Estimation of mutual information from limited experimental data
Houtsma, A.J.M.

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.390125

Published: 01/01/1983

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
the \( f_r \) ratios (the most sensitive ratio) lie approximately two-thirds of the way to the \( f_r \) ratio for constant \( B \). Using this piece of experimental data, and for a first-order approximation assuming that one can linearly interpolate between Eqs. (1) and (3), one arrives at Eq. (7). This approximation can be arrived at graphically by performing the same interpolation on the graph on page 80 of Ref. 8.

\[
k_{33} \approx 1.064 \left[ 1 - \frac{(f_r/f_o)^2}{2} \right]^{1/2},
\]

when \( k_{eff} > 0.75 \). The application of Eq. (7) yields \( k_{33}^{max} = 0.875 \) for the 2605SC alloy. The authors of Ref. 1 used Eq. (6) to compute their \( k_{33} \) for the 2605SC alloy. When modified by using Eq. (7), their maximum \( k_{33} \) becomes 0.924 instead of 0.965.

IV. CONCLUSION

It has been empirically shown that metallic-glass piezomagnetic ribbons of the 2605CO composition are in a constant \( H \) boundary condition due to the transverse magnetization. The 2605SC alloy is in a "2/3 constant \( B \) condition" when \( k_{eff} > 0.75 \). This is due to partial rotation of the magnetization of this alloy at very low fields because of the low anisotropy constant. An empirical relation between \( f_r, f_o \), and \( k_{eff} \) was derived for the 2605SC alloy. The maximum coupling coefficient of the 2605SC specimens tested was determined to be \( k_{33} \approx 0.875 \) (uncorrected for leakage flux).

The application of this empirical relation to the data of Ref. 1 yields \( k_{33} \approx 0.924 \), which is lower than the value of 0.965 reported there.

ACKNOWLEDGMENTS

The authors wish to thank Dr. A. E. Clark and Dr. L. T. Kabacoff for annealing the metallic-glass samples used in this study. Thanks are due to Clementina M. Ruggiero for assistance in the numerical solution of two equations, to Dr. R. W. Timme and Dr. R. Y. Ting for the critical review of this paper, and to L. C. Colquitt for technical assistance in collecting the data.


Estimation of mutual information from limited experimental data

Adrianus J. M. Houtsma
Institute for Perception Research, Den Dolech 2, Eindhoven, The Netherlands

(Received 25 April 1983; accepted for publication 26 July 1983)

To obtain an unbiased estimate of mutual information from an experimental confusion matrix, one needs a minimum number of trials of about five times the number of cells in the matrix. This study presents a computer-simulated approach to derive unbiased estimates of mutual information from samples of considerably fewer data.

PACS numbers: 43.66.Yw, 43.85.Ta, 43.60.Cg [JH]

This letter will discuss a problem recently encountered while performing an absolute identification experiment with a set of many stimuli which differed along three physical dimensions. The purpose of that experiment was to examine independence of perceptual correlates of the three dimensions by trying to find out whether or not information conveyed through a three-dimensional set of stimuli equals the sum of the amounts of information conveyed through three separate sets of stimuli that differ only along one dimension. The problem encountered was, how to obtain a reliable estimate of mutual information from identification data for a large set of alternative stimuli, while keeping the number of required experimental trials within the realm of reality.

As it has been several decades since information theory first found widespread use in psychophysics, it may be worth while reviewing some fundamental ideas. If an event \( X \) has \( k \) possible outcomes \( \{x_1, x_2, \ldots, x_k\} \) and the \( i \)th outcome occurs with probability \( p(x_i) \), then the average uncertainty or entropy, according to the Shannon–Wiener theory, is

\[
H(X) = - \sum_{i=1}^{k} p(x_i) \log_2 p(x_i).
\]

If successive events are observed through a noisy transmission channel, an observation \( Y \) results, also with \( k \) possible outcomes. An entropy measure similar to Eq. (1) can be defined for \( Y \) as well. A useful measure of how much informa-
tion is received by the observer through the transmission channel is the mutual information between $X$ and $Y$,

$$T(X;Y) = \sum_{i,j} p(x_i,y_j) \log_2 \left( \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \right),$$  

(2)

where $p(x_i,y_j)$ is the joint probability of the $i$th transmitted and the $j$th observed message. Entropy and mutual information are both expressed in bits. In practice they cannot be computed from Eqs. (1) and (2), however, because the probabilities $p(x_i), p(y_j)$, and $p(x_i,y_j)$ are not known a priori. They must be estimated from frequencies of occurrence in empirical data. The maximum likelihood estimate of $H(X)$ is

$$\hat{H}(X) = -\sum_{i=1}^k \left( \frac{n_i}{n} \right) \log_2 \left( \frac{n_i}{n} \right),$$  

(3)

where $n_i$ is the actual number of times the outcome $x_i$ occurred in a total of $n$ successive events. Similarly, there is a maximum likelihood estimate for $T(X;Y)$:

$$\hat{T}(X;Y) = \sum_{i,j} \left( \frac{n_{ij}}{n} \right) \log_2 \left( \frac{n_{ij}}{n} \right),$$  

(4)

where $n_{ij}$ is the frequency of the joint event $(x_i,y_j)$ in a sample of $n$ events, and $n = \sum_{i,j} n_{ij}$. The frequencies $n_i$, $n_j$, and $n_{ij}$ can all be derived from an empirical confusion matrix, which is the typical form in which data from absolute identification experiments are cast.

Neither $\hat{H}$ nor $\hat{T}$ are unbiased estimates of $H$ and $T$. It can be shown (Miller, 1954) that $\hat{H}$ is an underestimate and $\hat{T}$ is an overestimate. Since entropy increases when outcomes of events are more uniformly distributed, one would expect the estimated entropy $\hat{H}$, derived from a small data sample, always to be on the low side since such uniformity in data distribution can only be reached asymptotically. On the other hand, since mutual information $T$ is a measure of response consistency, i.e., whether or not the same observation is consistently made for a given input event, one expects a mutual information estimate $\hat{T}$ always to be on the high side. Especially when there is little basis for consistency, e.g., when the transmission channel is very noisy and observations are rather random, a small sample of observations may nevertheless look reasonably consistent since the observer did not have sufficient opportunity to be inconsistent. An extreme example, of course, is a sample of one single observation which is always consistent with itself, no matter whether it is a correct or an incorrect one.

Miller (1954) demonstrated a method for computing the bias in $\hat{T}$ for data samples in which the number of trials is at least five times the number of cells in the confusion matrix. The actual number of alternative stimuli in an absolute identification experiment does not therefore have to be very large before the required number of trials becomes unpractically large (e.g., 50 000 trials for 100 alternative stimuli).

In most of the older experiments on absolute identification of relatively large stimulus sets, investigators often collected many trials by running groups of subjects simultaneously and pooling their data. Pollack (1953) measured about 2.5 bits of mutual information with a set of 25 pure tones differing only in frequency by presenting the entire set five times to ten subjects. The total number of trials thus obtained was 1250, twice the number of cells in the confusion matrix. Hake and Garner (1951) measured 3.25 bits for a stimulus set of 50 different points on a line by presenting 200 trials to 16 subjects. The total number of trials obtained was 1.28 times the number of matrix cells. Klemmer and Frick (1953), who measured recognition of the position of a dot in a square, presented for 400 alternative dot positions all 400 possible stimuli once to 80 subjects, obtaining a number of trials only 0.2 times the number of matrix cells. The most remarkable case is perhaps the study by Pollack and Ficks (1954) who measured information transfer for an auditory stimulus which could take on five values in each of six dimensions, i.e., a set of 15 625 different stimuli. They presented an average of about 100 trials to 36 subjects, but did not pool the data. It is obvious that, even if the data were pooled, the amount would be far short of the number of trials required for obtaining an unbiased estimate of mutual information from an overall confusion matrix. Such a matrix would have more than 200 million cells, requiring by Miller's criterion at least one billion trials! Instead, they transformed their data into six $5 \times 5$ confusion matrices, one for each physical dimension, estimated mutual information in each matrix by means of the 100 trials, and added the results under the assumption of independence for a grand total of 7.2 bits of mutual information. Finally, the author recently performed an absolute identification experiment with a set of vibrotactile stimuli which could assume five different values in each of three dimensions. On each trial three responses were given, one corresponding to each physical dimension. A total of 5000 trials was obtained on one subject. Data were processed (1) by the method of Pollack and Ficks with results of 0.89, 0.88, and 1.37 bits of mutual information along the respective dimensions, and (2) by a direct estimate of $T$ from the overall $(125 \times 125)$ confusion matrix having a trial/cell ratio of 0.32, resulting in 3.94 bits of mutual information. All results have been summarized in Table I.

One sees that in the first three examples the number of trials taken falls progressively short of the minimum stipulated by Miller for obtaining an unbiased estimate of mutual

---

**Table I. Summary of stimulus set and observation sample sizes in selected absolute identification experiments.**

<table>
<thead>
<tr>
<th>Author</th>
<th>stim. in set</th>
<th>trials</th>
<th>trials/matrix cells</th>
<th>mut. inf. (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pollack (1953)</td>
<td>25</td>
<td>1 250</td>
<td>2.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Hake and Garner (1951)</td>
<td>50</td>
<td>3 200</td>
<td>1.28</td>
<td>3.25</td>
</tr>
<tr>
<td>Klemmer and Frick (1953)</td>
<td>400</td>
<td>32 000</td>
<td>0.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Pollack and Ficks (1954)</td>
<td>15 625</td>
<td>100</td>
<td>...</td>
<td>7.2</td>
</tr>
<tr>
<td>Houtsma (current rep.)</td>
<td>125</td>
<td>5 000</td>
<td>0.32</td>
<td>3.94</td>
</tr>
</tbody>
</table>
information. Pollack and Ficks explicitly assumed independence of stimulus-response relations between the various dimensions. If this were not the case, their simple addition scheme would not work, as shown in the following example of a two-dimensional stimulus $S_{ij}$ and response $R_{kl}$, where both dimensions of the stimulus (and response) have only two possible values. A hypothetical confusion matrix, showing the conditional probabilities $p(R_{kl}/S_{ij})$, is shown in Fig. 1. Since every stimulus has a unique response, mutual information equals two bits if stimuli $S_{ij}$ are presented with equal a priori probabilities. If, however, one computes from this matrix the two confusion matrices of each separate dimension, as shown in the same figure, one obtains uniform distributions of conditional response probabilities (0.5) with zero bits of mutual information along each of the two dimensions. Information is clearly not additive here. In fact, one can show that for this simple two-dimensional case:

(a) If stimulus-response combinations for the two dimensions are independent, mutual information is additive, i.e., the sum of the amounts of information conveyed through each dimension equals the total amount of information received.

(b) If the occurrence of a particular stimulus-response combination along one dimension makes a particular combination in the other dimension more likely, the total amount of information received is larger than the sum of the amounts of information measured along each dimension.

(c) If the occurrence of a particular stimulus-response combination along one dimension makes a particular combination in the other dimension less likely, total mutual information is less than the sum of the amounts measured in each dimension.

The author's data from the three-dimensional tactile experiment suggest that case (b) applied, and if the same were true for the Pollack and Ficks experiment, their result of 7.2 bits of total mutual information for a six-dimensional stimulus is an underestimate.

A practical approach to the problem of how to obtain an unbiased estimate of $T$ from a limited sample of absolute identification data is to simulate an identification experiment with varying numbers of trials and varying amounts of response noise (to simulate “good” and “bad” performance). On each trial, an integer $X$ was chosen with equal probability in the range $1 < x < 125$. A response $Y = X + R$ was generated as well, where $R$ is a uniformly distributed random integer in the range $-S < R < S$. Trials in which $Y$ came out larger than 125 or smaller than 1 were repeated to keep responses within the proper range. Figure 2 shows plots of $\hat{T}$, the maximum likelihood estimate of $T$, as a function of the data sample size $L$. Each of these dashed curves, corresponding to a particular value of $S$, represents about ten computed points (not visible because they all fall nearly exactly on the curves). On the same set of coordinates empirical values of $\hat{T}$ from the author's experiment are shown as a solid curve, derived from the first $L$ empirical data points of the total of 5000 trials. The dotted curve shows these same empirical values of $\hat{T}$ but bias-corrected with Miller's (1954) formula. One can easily see how much mutual information estimates are overcorrected if that formula is applied to data samples that are too small.

FIG. 2. Computer-simulated estimates of mutual information (dashed curves) for a set of 125 alternative stimuli. Curve parameters are amounts of response noise $S$ in the simulation model. Solid curve represents empirical results from an absolute identification experiment with 125 possible different stimuli in which 5000 trials were taken. Dotted curve shows the same empirical results after application of Miller's (1954) bias correction.
The curves shown in Fig. 2 clearly show that $\hat{T}$ decreases monotonically with the number of trials $L$ to an asymptotic value $T$. They clearly show how much mutual information is overestimated when the number of experimental trials taken is insufficiently large. Curves such as these provide at the same time a reasonably good unbiased estimate of $T$ from data samples considerably smaller than those required when Miller’s bias correction is to be used. One can fit an empirically determined function $T(L)$ to the nearest simulated function and read off the corresponding asymptotic value of $T$, although it must be said that for very small data samples this may be a difficult task too. Second, one sees that far fewer trials are needed to estimate a relatively high value of $T$ than are needed for a low value of $T$. That is because in the latter case all or nearly all cells of the confusion matrix are used, whereas in the case of large information transfer far fewer cells are used, yielding a much smaller “effective” matrix. It takes more trials to estimate a particular kind of distribution over a large number of possible outcomes by empirical means than it takes to estimate a distribution over only a few possible outcomes. Finally, the computed data points that determine each of the curves of Fig. 2 show, by repeated computation, extremely small variance. Although an analytic expression for the variance of $\hat{T}$ is not simple to obtain, Miller and Madow (1954) and Rogers and Green (1954) have computed approximate expressions for the first two moments of the entropy estimate $E[\hat{H}]$ and $E[\hat{H}^2]$. Their results show that, even if the number of trials is smaller than the number of possible input events ($n < k$), the variance of $\hat{H}$ is small compared to its mean. For the same reason the variance of $\hat{T}$ should be small, compared to its mean, even for relatively small trial samples, which is supported by our simulation results. Therefore, the main problem of using insufficiently large numbers of trials to estimate mutual information in an absolute identification paradigm is not the variance of the estimate, but its bias.

**ACKNOWLEDGMENTS**

The author is indebted to L. D. Briada, N. I. Durlach, I. Pollack, J. Roufs, and P. Zurek for their encouragement and helpful comments. Work was supported by the National Institutes of Health, Grant 2R01 NS 11680-05, and by the Karmazin Foundation, while the author was at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA.