*Analytica Chimica Acta*, 112 (1979) 233—243
*Computer Techniques and Optimization*


# CALCULATION OF THE INFORMATION CONTENT OF RETRIEVAL PROCEDURES APPLIED TO MASS SPECTRAL DATA BASES

GEERT VAN MARLEN*

*Department of Analytical Chemistry, Delft University of Technology, Jaffalaan 9, Delft (The Netherlands)*

AUKE DIJKSTRA and HENK A. VAN'T KLOOSTER

*Department of Analytical Chemistry, State University of Utrecht, Croesestraat 77A, Utrecht, (The Netherlands)*

(Received 17th April 1979)

SUMMARY

A procedure has been developed for estimating the information content of retrieval systems with binary-coded mass spectra, as well as mass spectra coded by other methods, from the statistical properties of a reference file. For a reference file, binary-coded with a threshold of 1% of the intensity of the base peak, this results typically in an estimated information content of about 50 bits for 200 selected *m/z* values. It is shown that, because of errors occurring in the binary-coded spectra, the actual information content is only about 12 bits. This explains the poor performance observed for retrieval systems with binary-coded mass spectra.

In recent years, information theory has been applied in different fields of analytical chemistry. The information content has been introduced as an optimization criterion in thin-layer chromatography [1], gas chromatography [2, 3], infrared [4, 5] and mass spectrometry [6—9]. It has been shown that a mass spectrum even with binary-coded intensities still provides a large amount of information [6]. The information content is diminished considerably by correlations between spectral features [9]. Similar observations have been made for binary-coded infrared spectra [4, 5].

The influence of errors and matching criteria on the retrieval of binary-coded mass spectra has been discussed [9]. It was concluded that the performance of the retrieval primarily depends on the extent of errors occurring in the coded spectra and is hardly affected by the matching criterion used.

In this paper, a method is described for calculating the information content for binary-coded mass spectra as well as spectra coded by other means. In addition, a new algorithm for feature selection is presented. Finally, an approach is outlined for prediction of the performance of a forward search system with binary-coded spectra, when data bases which differ with respect to the number and the nature of the compounds involved are considered.

## EXPERIMENTAL

The six mass spectral reference files used in this study are listed in Table 1. The spectra were binary-coded for a mass range of $m/z$ 1—300, by using an intensity threshold of $T$% of the intensity of the base peak, with $T$ varying from 0.1 to 20%. For some reference files, a 2-bit code was generated with three intensity levels, viz. 1, 5 and 20%, thus specifying peak intensities as: no peak, small peak, medium size and large peak, respectively.

For 1650 chemical compounds, duplicates were extracted from reference file F to enable an investigation of the effects of errors.

For the development and testing of the algorithms, a PDP11/45 minicomputer was used whereas the final computer programs were run on the IBM 370/158 computer system of the Delft University of Technology. All programs were written in FORTRAN IV.

## INFORMATION CONTENT

*Discontinuous distribution*

Retrieval from a reference file containing $N$ spectra each with $n$ coded features (in this case $m/z$ values) yields an amount of information called the "information content" of the retrieval procedure [9]. The information content for feature $j$, $I_H(j)$ in bits, without taking experimental and coding errors into account, is given by Shannon's equation [10]

$$I_H(j) = -\sum_{i=1}^{m} p_j(i) \log_2 p_j(i) \quad (1)$$

where $m$ represents the number of discrete values $i$ for feature $j$ with corresponding probabilities $p_j(i)$.

For binary-coded intensities, only two values for feature $j$ are distinguished: either below or above a given threshold, coded as "0" or "1", respectively.

TABLE 1

Mass spectral reference files used

| Ref. | Code-name | Number of spectra | Origin |
|---|---|---|---|
| A | ms9628 | 9628 | MSDC file, release 1971[a] |
| B | msalkane | 195 | Alkane Spectra from A |
| C | ms15796 | 15796 | MSDC file, release 1973[a] |
| D | ms02408 | 2408 | Hydrocarbon Spectra from C |
| E | ms11346 | 11346 | EPA/NIH file, release 1975[b] |
| F | ms22349 | 22349 | Mix of C and E |

[a]Mass Spectrometry Data Centre, Aldermaston, Gt. Britain.
[b]EPA/NIH Mass Spectral Data Base, 1975 edn., National Technical Information Service, Dept. of Commerce, 5285 Port Road, Springfield, Virginia 22151, U.S.A.

Since $\Sigma_{i=1}^{m} p_j(i) = 1$, it is obvious that in this case $p_j(0) = 1 - p_j(1)$. Equation (1) then reduces to

$$I_H(j) = -p_j \log_2 p_j - (1 - p_j) \log_2 (1 - p_j) \quad (2)$$

with $p_j = p_j(1)$. As an illustration for a large reference file, the probabilities and information contents for a number of $m/z$ values, calculated with eqn. (2), are compiled in Table 2.

If the spectral features are considered to be independent, the total information content can be calculated with

$$I_H(1, 2 \ldots n) = \sum_{j=1}^{n} I_H(j) = -\sum_{j=1}^{n} \sum_{i=1}^{m} p_j(i) \log_2 p_j(i) \quad (3)$$

If there is a dependence, eqn. (3) has to be replaced by

$$I_H(1, 2 \ldots n) = -\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \ldots \sum_{i_n=1}^{m} p(i_1, i_2 \ldots i_n) \log_2 p(i_1, i_2 \ldots i_n) \quad (4)$$

The total number of probabilities to be estimated amounts to $m^n$. For a small number of features $n$ and a relatively large number of spectra, eqn. (4) can be used to calculate the information content. In order to predict the information content of retrieval procedures for very large files from the statistical properties of small files, reliable estimates of $p$ must be available. However, in practice the number of spectral features is large (for mass spectra a few hundred $m/z$ values) and therefore it will be impossible to obtain an adequate estimate of

TABLE 2

Influence of the probability of occurrence $p$ and the mismatch probability $p_d$ on the information content $I_H$(Shannon, eqn. 2) and $I'_H$(Shannon, eqn. 14), respectively, for some binary-coded $m/z$ values from reference file F (in bits)

| $m/z$ | Threshold level 1% | | | | Threshold level 2% | | | |
|---|---|---|---|---|---|---|---|---|
| | $p$ | $I_H$ | $p_d$ | $I'_H$ | $p$ | $I_H$ | $p_d$ | $I'_H$ |
| 27 | 0.44 | 0.99 | 0.24 | 0.20 | 0.40 | 0.97 | 0.22 | 0.22 |
| 28 | 0.41 | 0.98 | 0.26 | 0.16 | 0.37 | 0.95 | 0.25 | 0.16 |
| 29 | 0.42 | 0.98 | 0.19 | 0.28 | 0.38 | 0.96 | 0.18 | 0.28 |
| 42 | 0.52 | 1.0 | 0.12 | 0.47 | 0.45 | 0.99 | 0.12 | 0.46 |
| 44 | 0.47 | 1.0 | 0.17 | 0.34 | 0.39 | 0.96 | 0.16 | 0.34 |
| 51 | 0.52 | 1.0 | 0.14 | 0.41 | 0.44 | 0.99 | 0.10 | 0.50 |
| 53 | 0.52 | 1.0 | 0.15 | 0.39 | 0.42 | 0.98 | 0.12 | 0.45 |
| 55 | 0.58 | 0.98 | 0.11 | 0.48 | 0.51 | 1.0 | 0.10 | 0.53 |
| 56 | 0.42 | 0.98 | 0.10 | 0.12 | 0.36 | 0.94 | 0.10 | 0.49 |
| 57 | 0.51 | 1.0 | 0.12 | 0.47 | 0.43 | 0.98 | 0.10 | 0.52 |
| 65 | 0.43 | 0.98 | 0.11 | 0.49 | 0.35 | 0.93 | 0.09 | 0.51 |
| 69 | 0.50 | 1.0 | 0.09 | 0.56 | 0.44 | 0.99 | 0.08 | 0.58 |
| 77 | 0.56 | 0.99 | 0.09 | 0.55 | 0.49 | 1.0 | 0.09 | 0.17 |
| 75 | 0.43 | 0.98 | 0.10 | 0.52 | 0.35 | 0.93 | 0.08 | 0.53 |
| 91 | 0.44 | 0.99 | 0.10 | 0.52 | 0.38 | 0.96 | 0.07 | 0.58 |

the $p$ values even with the sizes of the spectral data bases which are presently available. Application of eqn. (4) then leads to a maximum information content of $\log_2 N$ bits.

*Continuous distributions and correlations*

When the discontinuous distributions of the probabilities $p_j$ in eqn. (1) can be approximated by a continuous normal distribution, the sum in eqn. (1) is replaced by the integral

$$I_G(j) = -\int_{-\infty}^{+\infty} p_j(x) \log_2 p_j(x)\, dx \tag{5}$$

with $p_j(x)$ as the Gaussian distribution function for feature $j$ with value $x$ measured in histogram units and $I_G(j)$ the information content for the integral form. After integration eqn. (5) becomes

$$I_G(j) = \tfrac{1}{2} \log_2 2\pi e \sigma_j^2 \tag{6}$$

where $\sigma_j^2$ is the variance of the normally distributed feature $j$. Thus the information content is a logarithmic function of this variance. The $n$-dimensional equivalent becomes

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2 (2\pi e)^n |\mathrm{COV}| \tag{7}$$

when $I_G$ is the total information content for $n$ features and |COV| the determinant of the covariance matrix, filled with the variances $\sigma_i^2$ and the covariances $\sigma_{ij}$ [9]. Combination of eqns. (6) and (7) finally results in

$$I_G(1, 2 \ldots n) = \sum_{j=1}^{n} I_G(j) + \tfrac{1}{2} \log_2 |\mathrm{COR}| \tag{8}$$

with |COR| as the correlation determinant defined as

$$\mathrm{COR} = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \cdots & \cdots & \cdots & 1 \end{vmatrix}$$

The correlation coefficient $\rho_{ij}$ between the features $i$ and $j$ can be calculated with the estimates of the variances $\sigma_i^2$ and $\sigma_j^2$ and the covariance $\sigma_{ij}$ [9] from the equation $\rho_{ij} = \sigma_{ij}/(\sigma_i^2\, \sigma_j^2)^{\frac{1}{2}}$.

The second term in eqn. (8) is a correction of the information content caused by the interdependence of the features involved.

*Binary-coded spectra and feature selection*

Although binary-coded features can hardly be considered normally distributed, one can make an estimate of variances, covariances and correlation coefficients. The following equations can be derived:

$$\sigma_j^2 = N\,p_j\,(1-p_j)/(N-1) \tag{9}$$

$$\sigma_{ij} = (p_{ij} - p_i p_j)/[p_i p_j (1-p_i)(1-p_j)]^{\frac{1}{2}} \tag{10}$$

with $p_i$ and $p_j$ being the probabilities of coding feature $i$ or $j$, respectively, present ("1") and $p_{ij}$ the probability of coding both features present. Application of eqns. (2), (6) and (9) yields the information contents $I_H$ and $I_G$ as a function of $p_j$ for one binary feature $j$. The results are presented in Fig. 1. The approximation of $I_H$ by $I_G$ for values of $p$ between 0.15 and 0.85 results in a maximum error of 0.05 bit. Instead of making a correction in eqn. (8) for the differences between $I_H$ and $I_G$ [4, 9], the total information content is calculated from

$$I_G(1, 2 \ldots n) = \sum_{j=1}^{n} I_H(j) + \tfrac{1}{2} \log_2 |\mathrm{COR}| \tag{11}$$

The effect of the correction for correlated features (the second term in eqn. 11 is best illustrated in Table 3. In this Table the information contents calculated with eqns. (3), (4) and (11) are given for a set of $m/z$ values with low correlation derived from reference file A [9]. From these numbers, it appears that eqn. (11) gives a reasonable estimate of the total information content.

However, as shown in Table 4 for an artificial set of two highly correlated features, calculation of the information content with eqn. (11) will sometimes result in an overcorrection for the correlations between the features. In addition, the cumulative effect of the correlations will make it impossible to calculate the total information content for all the features concerned. To avoid this problem, features that tend to show the effect of overcorrection must be deleted during the calculation. This is achieved by applying a feature selection

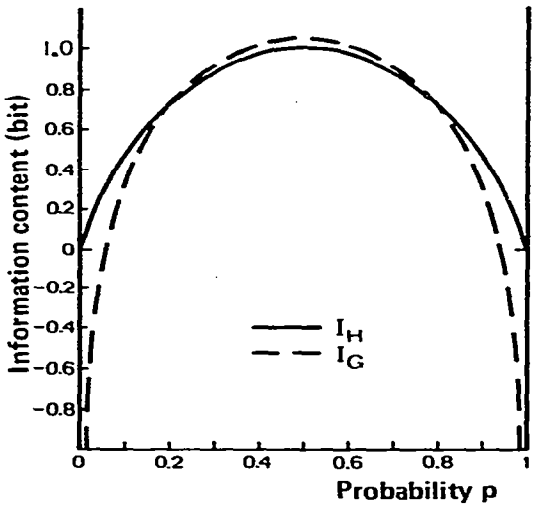


Fig. 1. The information contents $I_H$ (eqn. 2) and $I_G$ (eqns. 6, 9) as a function of the probability $p$ for a single binary-coded feature.

TABLE 3

Information contents for a set of $n$ poorly correlated $m/z$ values derived from reference file A (in bits)

| $n$ | $m/z$ value | $\Sigma I_H(j)$[a] | $I_H(1, 2 \ldots n)$[b] | $I_G(1, 2 \ldots n)$[c] |
|---|---|---|---|---|
| 1 | 77 | 1.0 | 1.0 | 1.0 |
| 2 | 69 | 2.0 | 2.0 | 2.0 |
| 3 | 27 | 3.0 | 3.0 | 3.0 |
| 4 | 50 | 4.0 | 3.8 | 3.8 |
| 5 | 45 | 4.9 | 4.6 | 4.7 |
| 6 | 40 | 5.9 | 5.4 | 5.5 |
| 7 | 57 | 6.9 | 6.2 | 6.2 |
| 8 | 75 | 7.8 | 6.8 | 7.0 |
| 9 | 81 | 8.7 | 7.4 | 7.7 |
| 10 | 44 | 9.7 | 8.1 | 8.4 |

[a]Sum of individual information contents, eqn. (3).
[b]Shannon information content, eqn. (4).
[c]Information content corrected for correlation, eqn. (11).

TABLE 4

Influence of a high correlation coefficient $\rho$ on the information content for a few sets of two artificial features 1 and 2 (in bits)

| $I_H(1)$[a] | $I_H(2)$[a] | $\rho_{12}$ | $\Sigma I_H(j)$[b] | $I_G(1, 2)$[c] |
|---|---|---|---|---|
| 1.0 | 1.0 | 1.0 | 2.0 | $-\infty$[d] |
| 1.0 | 1.0 | 0.9 | 2.0 | 0.8[e] |
| 1.0 | 1.0 | 0.8 | 2.0 | 1.3 |
| 0.9 | 0.8 | 1.0 | 1.7 | $-\infty$[d] |
| 0.9 | 0.8 | 0.9 | 1.7 | 0.5[e] |
| 0.9 | 0.8 | 0.8 | 1.7 | 1.0 |

[a]Shannon information content, eqn. (2).
[b]Sum of individual information contents, eqn. (3).
[c]Information content corrected for the correlation between 1 and 2, eqn. (11).
[d]Information content not defined (zero logarithm).
[e] Information content overcorrected for correlations.

algorithm, resulting in an information content $I_G(1, 2 \ldots l)$ for $l$ selected features with $l$ less or equal to $n$. A detailed description of the algorithm is given in the Appendix. These calculations have been carried out as a function of the intensity threshold $T$ for a number of binary-coded mass spectral reference files. The information content $I_G$ and the number of features selected are presented in Table 5.

*Extension to non-binary-coded spectra*

Since the method of calculating the information content uses only linear

TABLE 5

Information content $I_G(1, 2 \ldots l)$ and the number of selected $m/z$ values $l$ for the binary-coded reference files at 6 different threshold levels from eqns. (10) and (11) (in bits)

| Threshold level | 0.1% | | 1% | | 2% | | 5% | | 10% | | 20% | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference file | $I_G$ | $l$ | $I_G$ | $l$ | $I_G$ | $l$ | $I_G$ | $l$ | $I_G$ | $l$ | $I_G$ | $l$ |
| msalkane (B) | — | | 11 | 34 | — | | — | | — | | — | |
| ms9628 (A) | — | | 42 | 140 | — | | — | | — | | — | |
| ms02408 (D) | 29 | 95 | 22 | 80 | 19 | 77 | 16 | 77 | 12 | 76 | — | |
| ms15796 (C) | 47 | 157 | 46 | 156 | 41 | 159 | 32 | 159 | 24 | 164 | 17 | 171 |
| ms11346 (E) | 55 | 194 | 55 | 196 | 51 | 198 | 40 | 197 | 29 | 188 | 20 | 197 |
| ms22349 (F) | — | | 52 | 174 | 47 | 176 | — | | — | | — | |

correlations between the features, another estimate of these correlations may be obtained by applying the 2-dimensional expressions of eqns. (4) and (11) for the features 1 and 2:

$$I_H(1, 2) = -\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} p(i_1, i_2) \log_2 p(i_1, i_2) \tag{12}$$

$$I_G(1, 2) = I_H(1) + I_H(2) + \tfrac{1}{2} \log_2 (1 - \rho_{12}^2) \tag{13}$$

$I_H(1, 2)$ can easily be calculated with a large set of spectra, since only four $p$ values have to be estimated for binary-coded features. Equating $I_H(1, 2)$ and $I_G(1, 2)$ then leads to a new estimate for the correlation coefficient $\rho_{12}$; this may then be used in eqn. (11), thus circumventing overcorrection. The same feature selection procedure mentioned above has been applied to obtain the information content $I_G$ for the binary-coded reference files. Table 6 gives a brief example of the results.

This alternative manner of quantifying correlations between the features can also be used to assess the information content for non-binary-coded features, provided that the number of spectra is large enough to estimate all $p$ values used in eqn. (12). For some reference files, the features have been converted to a 2-bit code; the information contents and the numbers of features selected are given in Table 7.

TABLE 6

Information content $I_G$ (eqns. 11 and 13) and the number of selected $m/z$ values, $l$, for the binary-coded reference files (threshold level 1%) (in bits)

| Reference file | $I_G(1, 2 \ldots l)$ | $l$ | Reference file | $I_G(1, 2 \ldots l)$ | $l$ |
|---|---|---|---|---|---|
| msalkane (B) | 13 | 48 | ms15796 (C) | 60 | 257 |
| ms02408 (D) | 23 | 93 | ms11346 (E) | 78 | 282 |
| ms9628 (A) | 49 | 214 | ms22349 (F) | 68 | 276 |

TABLE 7

Information content $I_G$ and the number of selected $m/z$ values $l$ for 3 reference files, converted into a 2-bit code (eqns. 11, 12, 13) (in bits)

| Reference file | $I_G(1, 2 \ldots l)$ | $l$ |
|---|---|---|
| ms02408 (D) | 40 | 103 |
| ms15796 (C) | 110 | 288 |
| ms11346 (E) | 137 | 284 |

*Influence of errors*

For actual retrieval of spectra, the information content must be corrected for the uncertainty remaining after coding the spectra. The uncertainty in the coded spectra is caused by deviations in experimental, recording and coding conditions. Hence, for binary coded features, eqn. (2) must be replaced by:

$$I'_H = -p \log_2 p - (1-p) \log_2(1-p) + p[p_{1/0} \log_2 p_{1/0} + (1-p_{1/0}) \log_2(1-p_{1/0})] + (1-p)[p_{0/1} \log_2 p_{0/1} + (1-p_{0/1}) \log_2(1-p_{0/1})] \quad (14)$$

with $p_{1/0}$ and $p_{0/1}$ being the probabilities of coding a "0" or a "1", when the spectrum of the reference compound contains a "1" or a "0", respectively, for that feature. From the probability of finding a mismatch between two spectra of the same compound (the "mismatch probability" $p_d$) the "error probabilities" $p_{1/0}$ and $p_{0/1}$ can be calculated from [11]

$$p_{1/0} = p_d/2p \text{ and } p_{0/1} = p_d/2(1-p)$$

The mismatch probabilities are estimated from a set of 1650 pairs of spectra of the same compound, extracted from the reference file F. The large influence of the mismatch and error probabilities on the information content actually obtained in a retrieval process is shown for some $m/z$ values in Table 2. The mismatch probability ranges up to 0.26, with the maximum values in the low $m/z$ area.

With the assumption that the error probabilities for the different features are independent, the correction of the information content can also be done by replacing $I_H(j)$ by $I'_H(j)$ in eqn. (11), yielding $I'_G$. In Table 8 the information contents $I_G$ and $I'_G$, calculated for reference file F, binary-coded with two threshold levels, are given as an example.

Inspection of the results indicates that the information content drops dramatically from 52 to 11 bits and from 47 to 12 bits, respectively. This is much more than previously predicted [9]. Repeating the calculation for reference file A gave the same effect. This result explains the generally poor performance of a retrieval system based on binary-coded mass spectra [11].

## CONCLUSIONS

The model described for the calculation of the information content of

TABLE 8

Information contents $I_G$ (eqns. 10, 11) and $I'_G$ (bits) (eqns. 10, 11, 14) as a function of the number of selected $m/z$ values $l$ for reference file F, binary-coded at two threshold levels

| $l$ | Threshold level 1% | | Threshold level 2% | |
|---|---|---|---|---|
| | $I_G(1, 2 \ldots l)$ | $I'_G(1, 2 \ldots l)$ | $I_G(1, 2 \ldots l)$ | $I'_G(1, 2 \ldots l)$ |
| 10 | 8.7 | 4.3 | 8.4 | 4.4 |
| 20 | 14.9 | 6.8 | 14.5 | 7.1 |
| 30 | 20.3 | 8.5 | 19.5 | 8.9 |
| 40 | 25.0 | 9.7 | 23.7 | 10.3 |
| 50 | 29.1 | 10.5 | 27.4 | 11.3 |
| 60 | 32.7 | 11.0 | 30.8 | 12.0 |
| 70 | 36.1 | 11.2 | 33.8 | 12.5 |
| 80 | 39.1 | 11.3 | 36.5 | 12.7 |
| 90 | 41.8 | — | 38.9 | 12.8 |
| — | | | | |
| — | | | | |
| Max.[a] | 51.7 | 11.3 | 47.4 | 12.8 |

[a]The maximum number of selected $m/z$ values is 174, 77, 176 and 89, respectively.

retrieval procedures involves estimates of variances and correlation coefficients for a limited number of spectra in order to predict implicitly the probabilities for all possible codes. For features with extremely high correlations, the first model would lead to pessimistic estimates of the resulting information content. Use of a feature selection procedure is inevitable.

The second method for the calculation of the information content includes a different measure of the correlations (eqn. 13) which results in a contribution of almost all coded features to the information content. For this reason, the second method is expected to yield more reliable estimates of the information content. This method is also feasible for spectra coded by non-binary techniques, although it should be refined in order to deal with the influence of experimental and coding errors.

In the model discussed, the reference files are considered as select samples of complete populations of mass spectra. Consequently, in the design or evaluation of retrieval methods the reference file to be used for the calculation of the information content should be as representative of a certain population of compounds as possible. For example, the results for alkane spectra (reference file D), even without a correction of the information content for errors in the spectra, show clearly that binary coding does not yield enough information and is not recommended for retrieval purposes.

For retrieval from reference files containing spectra of a wider variety of compounds (such as C, E and F), binary coding is more promising. With a threshold of 1% of the intensity of the base peak, an information content of about 50 bits for 200 selected $m/z$ values is typical. This value decreases rapidly above threshold values of 2%, mainly because of the diminishing number of peaks per spectrum.

However, most of the information is lost because of the errors that occur in the binary-coded spectra. The performance of a retrieval system based on the binary-coded reference file A confirms this conclusion [11]. The prediction that a threshold of 2% would yield a better retrieval performance than 1% is confirmed by the preliminary results obtained with reference file F.

The authors gratefully acknowledge valuable discussions with Louis van Norel, Peter Cley and Jan Van den Hende.

## APPENDIX: FEATURE SELECTION ALGORITHM

To calculate the information content $I_G$ for a near-optimal subset of $l$ out of $n$ features, eqn. (11) is converted to

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2 |S| \tag{a1}$$

with the elements of determinant $S$ defined as $s_{ij} = 2I_H(i) \cdot 2I_H(j) \cdot \rho_{ij}$. This equation introduces a scaling which avoids computational problems such as arithmetical under- and over-flow. The diagonal elements of $S$ range in value from 1 to 4 and represent the exponential form of the information content for every feature ($I_G(i) = \frac{1}{2} \log_2 s_{ii}$). The non-diagonal elements vary from $-4$ to $+4$, since the correlation coefficient $\rho$ ranges from $-1$ to $+1$, and represent the "covariances" between the features. The information content is calculated after the covariance elements of $S$ have been zeroed by using the Gaussian elimination method [12] with

$$I_G(1, 2 \ldots n) = \tfrac{1}{2} \log_2 \prod_{i=1}^{n} s'_{ii} = \sum_{i=1}^{n} \tfrac{1}{2} \log_2 s'_{ii} \tag{a2}$$

where $s'_{ii}$ is the "variance" element after correction for the correlations between the features.

Because of the effect of overcorrection for correlations between the features, only $l$ features contribute to the calculated information content $I_G$. These features are selected together with their contribution to $I_G$ by the following procedure:

1. select the feature $i$ with the highest value of $s_{ii}$;
2. correct the variances of the remaining features $k$ for the correlation between $i$ and $k$ with

$$s'_{kk} = s_{kk} - s^2_{ik}/s_{ii} \tag{a3}$$

3. find the next feature $n$ with the highest corrected value of $s'_{nn}$;
4. correct the covariances of the remaining features $k$ for the correlation between feature $k$ and all features $j$ previously selected, excluding feature $n$, with

$$s'_{nk} = s_{nk} - s_{jk} \cdot s_{jn}/s_{jj} \tag{a4}$$

(Note that for $j = n$, $s'_{nk}$ is zero in which case $s'_{nk}$ is not calculated.)

5. replace $i$ by $n$ and restart at step 2.

The iterative procedure is terminated when all features have been selected or when the highest corrected value of $s'_{nn}$ in step 3 becomes less than 1; all remaining features $k$ will then have $s'_{kk}$ less than 1 and $\frac{1}{2} \log_2 s'_{kk}$ becomes negative.

The procedure described has been chosen to avoid the correction of all covariances in $S$; the elements need to be corrected only for the correlations with the features already selected. In programming this procedure the indices $i$, $j$, $k$ and $n$ are replaced by elements of a pointer array, in which the sequence of the features is stored, and which enables an indirect reference to the rows and columns of $S$ without time-consuming rearrangements. The square matrix $S$ is symmetric and only the upper triangle of the matrix need be stored, giving a considerable reduction in storage requirements. However, if possible the linear storage of $S$ should be avoided as otherwise all indices in the eqns. (a3) and (a4) have to be computed with a special function. The calculation of the information content $I_G$ for the mass spectral reference files in this study, with 300 binary-coded $m/z$ values, requires the storage of 300 × 300 elements of the matrix $S$. The execution time depends linearly on the number of coded features multiplied by the number of features selected. On the IBM 370/158 computer system this amounts to approximately 40 s for 100 features selected and 300 $m/z$ values coded.

REFERENCES


1 D. L. Massart, J. Chromatogr., 79 (1973) 157.
2 P. F. Dupuis and A. Dijkstra, Anal. Chem., 47 (1975) 379.
3 A. Eskes, P. F. Dupuis, A. Dijkstra, H. Declerc and D. L. Massart, Anal. Chem., 47 (1975) 2168.
4 P. F. Dupuis and A. Dijkstra, Fresenius' Z. Anal. Chem., 290 (1978) 357.
5 P. F. Dupuis, A. Dijkstra and J. H. van der Maas, Fresenius' Z. Anal. Chem., 291 (1978) 27.
6 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
7 F. Erni, thesis no. 4296, Eidgen. Techn. Hochschule, Zürich (1972).
8 F. Erni and J. T. Clerc, Helv. Chim. Acta, 55 (1972) 489.
9 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
10 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill., 1949.
11 G. van Marlen, A. Dijkstra and H. A. van't Klooster, Anal. Chem., 51 (1979) 420.
12 R. W. Hamming, Introduction to Applied Numerical Analysis, McGraw-Hill, New York, 1971, Ch. 5.