Design of a computer program for off-line processing of gas-chromatographic data

DESIGN OF A COMPUTER PROGRAM FOR OFF-LINE PROCESSING OF GAS-CHROMATOGRAPHIC DATA

M. H. J. van RIJSWICK

DESIGN OF A COMPUTER PROGRAM FOR OFF-LINE PROCESSING OF GAS-CHROMATOGRAPHIC DATA

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE TECHNISCHE WETENSCHAPPEN AAN DE TECHNISCHE HOGESCHOOL EINDHOVEN, OP GEZAG VAN DE RECTOR MAGNIFICUS, PROF. DR. IR. G. VOSSERS, VOOR EEN COMMISSIE AANGEWEZEN DOOR HET COLLEGE VAN DEKANEN IN HET OPENBAAR TE VERDEDIGEN OP DINSDAG 3 DECEMBER 1974 TE 16.00 UUR

DOOR

MATHIAS HUBERTUS JOHANNES VAN RIJSWICK

GEBOREN TE TIENRAY

DIT PROEFSCHRIFT IS GOEDGEKEURD DOOR DE PROMOTOREN
PROF. DR. IR. A. I. M. KEULEMANS EN DR. IR. R. S. DEELDER

Aan mijn ouders

CONTENTS

1. DATA PROCESSING

The purpose of all analytical methods is to obtain information. A gas chromatograph may be seen as an information source that sends the information as an encoded signal. The data processor acting as the receiver, has to decode the signal and deliver the information in a form intelligible to the person (or thing) asking for it. Data processing thus consists of two parts:
— to extract the information from a signal that contains also irrelevant and interfering components;
— to bring the information in a useful form.

The subject of the present work is to investigate the automation of the data processing and to develop a suitable computer program for it. The primary aim is that this program can be applied to any chromatographic signal. In order to cope with the most exacting cases, three requirements are implied:
— low detection limits,
— optimum accuracy and precision,
— automatic processing.

Low detection limits are required because it is generally unknown, a priori, which peaks are relevant. Optimum accuracy and precision are desired in order that the quality of the results is not unnecessarily limited by the quality of the processing. Automatic processing is necessary to make the performance independent of the user's skill. However, a "blackbox" design offers additional advantages:
— it minimizes the working knowledge, making the program easy to use;
— the consistency of the results is improved by excluding external interference which is often irreproducible and arbitrary;
— the performance can be specified rigorously, as it does not depend on the user's skill.

The design also has a number of drawbacks:
— prior knowledge available to the user is also excluded;
— the processing will be too sophisticated and inefficient for many chromatograms.

It may appear contradictory to conceive a "general" program, if such a program is implicitly inefficient for most applications. The explanation is that the appropriate simplifications for a particular application can be readily made from a well-designed general program.

The function of data processing is to bridge the gap between the information as it is contained in the chromatographic signal and the form in which it is desired. This transformation is commonly effected in three steps, as illustrated in fig. 1.1.

In the *data-extraction* step the irrelevant components of the signal and the redundancy are eliminated. The information is concentrated in parameters that

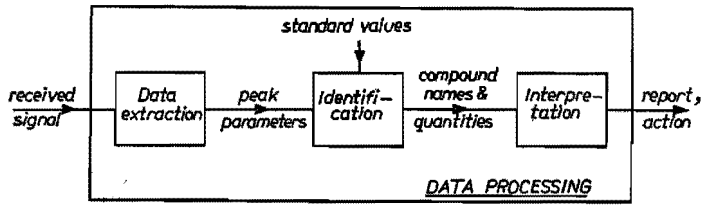


Fig. 1.1.

characterize the peaks.

In the *identification* step the peaks are identified from the retention value. Names and structures of the compounds can be found by comparison with tabulated values. Generalized structure-retention relations may aid in the elucidation of an unknown compound.

In the *interpretation* step the results of the identification are linked to operative consequences. The content of this step may vary from straightforward decision making to involved statistical analysis.

In the present work only the first two steps will be investigated. The interpretation step is closely related to particular applications and, therefore, less suited to a general discussion.

1.1. *Survey of literature*

Generally, data processing for chromatography will be similar to the data processing for other instrumental techniques giving a peak-like signal on a noisy background. Most of the applied methods will be identical. In detail, however, a number of distinct features leads to special requirements:

— Chromatographic analysis gives one peak for each compound. Being so little redundant, complete extraction of the information is essential.
— The shape of the peaks and the background is rather variable.
— Chromatography is mainly used for quantitative analysis, so that accurate background correction and precise calculation of the peak areas is required.

The cumulative effect of these details is that chromatographic-data processing requires a specially designed program. As the data processing is often the most time-consuming part of chromatographic analysis, it is obvious that its automation has already received great interest. Comprehensive reviews were recently given by Leathard[1-7]) and Ziegler[1-8]). We will mention here only those sources that have some relevance for the present work. Moreover, this section is confined to a discussion of the investigations that reported a program for integral processing.

Littlewood et al.[1-1,2,3]) described a program specially aiming at the separation of overlapping peaks by curve fitting. Detection methods able to find shoulder peaks were given. A number of processing parameters must be preset

by the user. Accuracy and precision were studied experimentally by running samples with known ratios of the compounds.

Westerberg [1–4]) discussed the peak-detection and resolution methods for a computer system that handles several concurrently operating on-line gas chromatographs. Limits for the detectability of overlapping peaks were derived. The errors for area calculation by triangulation and perpendicular drop were evaluated; it is concluded that these methods are too inaccurate so that curve fitting should be used. No details on the operation of the system were given.

Wijtvliet [1–5]) developed a computer program having as prime design goals: easy to use, failsafe and foolproof. Considerable effort was put in determining the peak-top location with utmost accuracy. The program however requires well separated peaks and a stable baseline. No detection limits were reported. Although the program may be run on standard values, presetting of some processing controls is required for obtaining optimum results.

Brouwer and Jansen [1–6]) described a program for automatic evaluation of complex spectra, using a combination of correlation detection and curve fitting. The method cannot be applied directly to chromatography because a known, invariable peak shape is assumed.

1.2. *Scope of this work*

Although not radically different, we believe that our approach is uncommon by a combination of two aspects:
— the explicit aim of optimum information extraction, and
— extended automation.
Parts of the problem have been investigated, either in the papers mentioned above or in papers concentrating on certain topics such as detection, accuracy and precision, curve fitting, etc.

The subject of the present work may be summarized as
— adaptation of existing techniques,
— development of new methods,
— integration of both in an operational program.

In chapter 2 the data extraction will be surveyed in greater detail in order to identify the elements of the problem. After reviewing the methods that have been applied, an outline of the selected techniques is given.

Chapter 3 gives a detailed account of the development of the algorithms. Special emphasize is given to the evaluation of the performance of the adopted methods, such as detection limits or potential accuracy and precision.

Chapter 4 is made up of three parts: it summarizes the performance specifications for readers who preferred to skip chapter 3; the application of the program to a "difficult" chromatogram is discussed; the possibilities for making the program more efficient and for simplifications are mentioned.

Chapter 5 is concerned with computer-aided identification. A matching crite-

rion is proposed for identification by "table matching", and its use is demonstrated by some examples. The use of structure-retention relations for identification is briefly discussed.

REFERENCES

[1-1]) A. B. Littlewood, T. C. Gibb and A. H. Anderson, in C. L. A. Harbourn (ed.), Gas chromatography 1968, Institute of Petroleum, London, 1969, p. 297.
[1-2]) A. H. Anderson, T. C. Gibb and A. B. Littlewood, Anal. Chem. **42**, 434, 1970.
[1-3]) A. H. Anderson, T. C. Gibb and A. B. Littlewood, in A. Zladkis (ed.), Advances in chromatography 1970, Miami, 1970, p. 75.
[1-4]) A. W. Westerberg, Anal. Chem. **41**, 1770, 1969.
[1-5]) J. J. M. Wijtvliet, Thesis Technological University Eindhoven, 1972.
[1-6]) G. Brouwer and J. A. J. Jansen, Anal. Chem. **45**, 2239, 1973.
[1-7]) D. A. Leathard, in H. Purnell (ed.), Advances in analytical chemistry and instrumentation, Vol. 11: New developments in gas-chromatography, Wiley, 1973, p. 29.
[1-8]) E. Ziegler, Computer in der instrumentellen Analytik, Akademische Verlagsgesellschaft, Frankfurt, 1973.

2. SURVEY OF DATA EXTRACTION

Data extraction consists of the separation of the peaks from the background and the concentration of the information in parameters that characterize the peaks. The purpose of this chapter is to examine the elements of this problem in more detail in order to conceive a framework for a suitable program.

2.1. *Peak models*

To be able to separate the peaks from the background, it is necessary to know their characteristics. Usually the signal is taken as a superposition of three components, viz. peaks, a deterministic baseline and random noise, as illustrated in fig. 2.1. The way how prior information about the characteristics of the components is used in the data extraction was extensively discussed by Kelly and Harris [2–1]). Two examples show how different prior information leads to different approaches:

— If the shape of the baseline is known and the contribution can be determined over the entire chromatogram, the peaks emerge as the residues after subtraction of this contribution. The peaks can then be characterized by the area under the curve, the location of the top and the centre of gravity, the width, etc. Thus almost nothing is assumed about the peak shape. Noise is considered as a source of uncertainty, causing that the true values of the peak parameters cannot be determined. This approach was among others followed by Wijtvliet [2–2]), who approximated the baseline by a horizontal line.

— Another approach is made if the peak shape is known and the baseline is less well defined. Now one can look for the presence of profiles in the signal that are congruent to the model profile. Brouwer and Jansen [2–3]) evaluated complex spectra in this way by taking a Gaussian peak model of fixed width and assuming that the baseline is a slowly varying function of time that can be eliminated by differentiation.

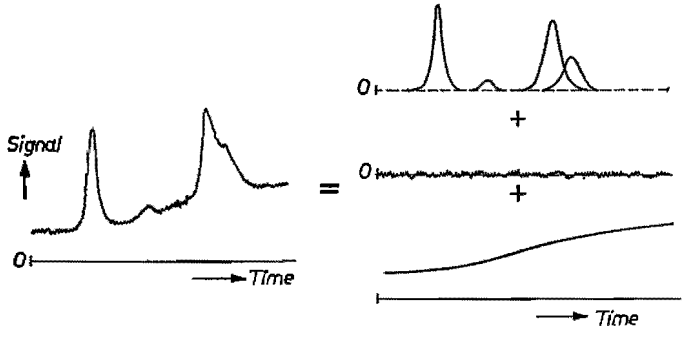


Fig. 2.1. A chromatogram is conceived of a superposition of peaks, random noise and a slowly drifting baseline.

The amount of information that can be obtained from a given signal depends to a large extent on the prior knowledge. Accurate peak parameters can only be obtained if an accurate background correction can be made. Overlapping peaks can only be dissected if an accurate peak model is available. In the following some relevant peak models are discussed.

A chromatographic peak is the residence-time distribution of the molecules of a certain compound. The distribution function might thus be derived from equations that describe the transport of the molecules through the column. Unfortunately, a realistic description of the transport leads to a complex set of differential equations that generally cannot be solved. There are various levels of simplification.

The simplest model [2–4]) yields that the peaks rapidly approximate a Gaussian distribution:

$$g(t, A, \mu, w) = \frac{A}{w\,(2\pi)^{1/2}} \exp\left[-\tfrac{1}{2}\left(\frac{t-\mu}{w}\right)^2\right]. \qquad (2.1)$$

This distribution is completely characterized by three parameters, viz. the area A under the peak, the location of the top μ and the parameter w as a measure for the width. Often, in chromatography, the width is taken at half height, so that for a Gaussian peak $w_{1/2} = w\,(8 \ln 2)^{1/2} \approx 2{\cdot}35\,w$. However, unless stated otherwise we will denote w as the peak width.

From an analytical point of view, the area and the location are the parameters that carry relevant information. The location, or retention time, is specific for the chemical identity of the compound. The area is related to the amount of the compound. With a linear detector response, the absolute quantity can be calculated if the sensitivity factor is known. The peak width is largely determined by the instrumental system and bears only minor specific information. Thus for a given analysis the areas and the retention times are free parameters from which generally nothing is known in advance, while the widths are approximately proportional to the retention times, the proportionality constant being largely determined by the chromatograph.

The Gaussian model gives only in few cases an accurate description of experimental peaks. For one thing, peaks are often asymmetric. An extension takes some instrumental factors into account. It is assumed [2–5]) that the chromatographic process yields a Gaussian shape, but mixing volumes in the injection port or in the detector modify this in a convolution with an exponential function, as illustrated in fig. 2.2. The resultant peaks will show a degree of tailing that depends on the time constant τ in the exponential function.

As most chromatographic peaks show some degree of tailing, this seems a plausible model. However, a purely instrumental contribution implies that the time constant is equal for all peaks. In practice some peaks show more tailing

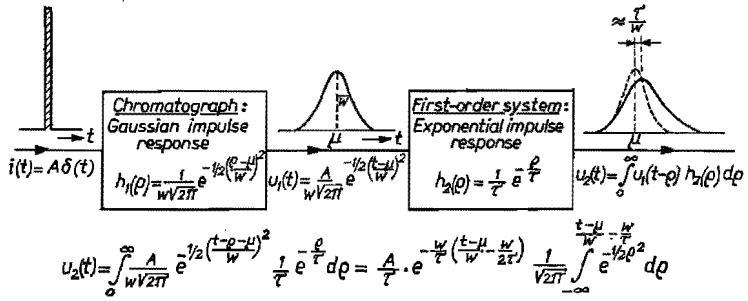


Fig. 2.2. Modification of a Gaussian peak by a first-order system.

than others. The reason is that tailing is mostly due to a competing non-linear retention mechanism, e.g. adsorption, which is dependent on the presence of certain specific groups.

Instead of deriving models which give a more realistic account of the physical processes in chromatography, one can also search pragmatically for functions that do simulate real peaks. Some parameters in such a function serve only for an adequate fit and have no pertinent physical meaning. The purpose of these "models" is that, if they are specific enough so that un-peaklike configurations are not exhibited, they can be applied for separating peaks from the background or for apportioning a composite peak. The models are often derived from the Gaussian function. A number of them were proposed by Fraser and Suzuki [2-6]).

A frequently mentioned model is the bi-Gaussian function, listed in table 2-I. This is a Gaussian function with different widths at the leading and at the trailing edge. Obviously this is not a correct physical model, but it is useful for simulating various asymmetrical shapes.

Another possibility is to take as a model the product of the Gaussian function $g(t)$ and a correction function $h(t)$: $f(t) = g(t)\, h(t)$. If $h(t)$ is developed [2-7]) in a series of Hermite polynomials and terms are grouped in decreasing order of magnitude, one obtains Edgeworth's series (with $\eta = (t - \mu)/w$):

$$f(\eta) = g(\eta) - \left(\frac{\gamma_1}{3!} g^{(3)}(\eta)\right) + \left(\frac{\gamma_2}{4!} g^{(4)}(\eta) + \frac{10\,\gamma_1^2}{6!} g^{(6)}(\eta)\right) + \ldots$$

$$= g(\eta)\left(1 + \frac{\gamma_1}{6}(\eta^3 - 3\eta) + \frac{\gamma_2}{24}(\eta^4 - 6\eta^2 + 3) + \right.$$

$$\left. + \frac{\gamma_1^2}{72}(\eta^6 - 15\,\eta^4 + 45\,\eta^2 - 15)\right) + \ldots . \qquad (2.2)$$

The additional parameters γ_1 and γ_2 are the coefficients of skewness and excess, as discussed below. This model, which was also proposed by Kelly and Harris [2-1]), has the advantage over the well-known [2-8]) Gram–Charlier A-

TABLE 2-I

FUNCTION	PARAMETERS	PARTIAL DERIVATIVES	MOMENTS
GAUSS $g(t)=\frac{A}{w\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{t-\mu}{w}\right)^2}$; $\eta=\frac{t-\mu}{w}$	A : area μ : top location w : width	$\frac{\partial g}{\partial A}=\frac{g(t)}{A}$ $\frac{\partial g}{\partial \mu}=\frac{g(t)}{w}\eta$ $\frac{\partial g}{\partial w}=\frac{g(t)}{w}(\eta^2-1)$	$m_0=A$ $m_1=\mu$ $m_2=w^2$ $m_3=0$ $m_4=3w^4$
EXPONENTIALLY CONVOLUTED GAUSS $f(t)=\int_0^t \frac{1}{\tau}e^{-\frac{t'}{\tau}}g(t-t')\,dt'$ $=\frac{A}{\tau}e^{-\rho(\eta-\frac{1}{2}\rho)}\int_{-\infty}^{\eta-\rho}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,dx$ $\rho=\frac{w}{\tau}$	A : area μ : top w : width } of unconvoluted Gauss τ : time constant	$\frac{\partial f}{\partial A}=\frac{f(t)}{A}$ $\frac{\partial f}{\partial \mu}=\frac{1}{\tau}\left(f(t)-g(t)\right)$ $\frac{\partial f}{\partial w}=\rho\left[\frac{1}{\tau}\left(f(t)-g(t)\right)-\eta\,\frac{g(t)}{w}\right]$ $\frac{\partial f}{\partial \tau}=\frac{f(t)}{\tau}(\rho\eta-1)-\frac{\rho^2}{\tau}\left(f(t)-g(t)\right)$	(Ref. 2-30): $m_0=A$ $m_1=\mu+\tau$ $m_2=w^2+\tau^2$ $m_3=\tau^3$ $m_4=3w^4+6\,w^2\tau^2+9\tau^4$
BI-GAUSS $f(t)=\frac{2A}{(w_1+w_2)\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{t-\mu}{w_i}\right)^2}$ $t<\mu: i=1$ $t\geq\mu: i=2$	A : area μ : top w_1 : front width w_2 : back width	$\frac{\partial f}{\partial A}=\frac{f(t)}{A}$ $\frac{\partial f}{\partial \mu}=\frac{f(t)}{w_i}\left(\frac{t-\mu}{w_i}\right)$ $\begin{cases} t<\mu: i=1 \\ t\geq\mu: i=2 \end{cases}$ $\frac{\partial f}{\partial w_i}=f(t)\left[\left(\frac{t-\mu}{w_i}\right)^2\frac{1}{w_i}-\frac{1}{w_1+w_2}\right]$ $\begin{cases} t<\mu: i=1 \\ t>\mu: i=2 \end{cases}$ $\frac{-f(t)}{(w_1+w_2)}$ $\begin{cases} t\geq\mu: i=1 \\ t\leq\mu: i=2 \end{cases}$	$m_0=A$ $m_1=\mu-\frac{2(w_1-w_2)}{\sqrt{2\pi}}$ $m_2=\left(\frac{w_1+w_2}{2}\right)^2+\left(\frac{3\pi-8}{4\pi}\right)(w_1-w_2)^2$ m_3, m_4 } See ref. 2-15
EDGEWORTH SERIES $f(t)=g(t)-\frac{m_3}{6}g^{(3)}(t)+\frac{m_4}{24}g^{(4)}(t)+\frac{m_3^2}{72}g^{(6)}(t)$ $=g(t)\left[1+\frac{\gamma_1}{6}(\eta^3-3\eta)+\frac{\gamma_2}{24}(\eta^4-6\eta^2+3)+\right.$ $\left.+\frac{\gamma_1^2}{72}(\eta^6-15\eta^4+45\eta^2-15)\right]$	A, μ, w } as in Gauss γ_1 skew γ_2 excess	$\frac{\partial f}{\partial A}=\frac{f(t)}{A}$ $\frac{\partial f}{\partial \mu}=\frac{f(t)}{w}\eta-\frac{g(t)}{w}\left[\frac{\gamma_1}{6}(3\eta^2-3)+\frac{\gamma_2}{24}(4\eta^3-12\eta)+\right.$ $\left.+\frac{\gamma_1^2}{72}(6\eta^5-60\eta^3+90\eta)\right]$ $\frac{\partial f}{\partial w}=\eta\frac{\partial f}{\partial \mu}-\frac{f(t)}{w}$ $\frac{\partial f}{\partial \gamma_1}=g(t)\left[\frac{1}{6}(\eta^3-3\eta)+\frac{\gamma_1}{36}(\eta^6-15\eta^4+45\eta^2-15)\right]$ $\frac{\partial f}{\partial \gamma_2}=g(t)\left[\frac{1}{24}(\eta^4-6\eta^2+3)\right]$	$m_0=A$ $m_1=\mu$ $m_2=w^2$ $m_3=w^3\gamma_1$ $m_4=w^4(\gamma_2+3)$

series that the order of magnitude of the terms is steadily decreasing so that less terms are required. In fact, the model (2.2) is already so flexible that for certain values of the additional parameters γ_1 and γ_2 an un-peaklike two-topped shape results, as shown by Kroon [2–9]). This implies that this function cannot be used for separating overlapping peaks unless the parameter values are bounded to some ranges.

A function which is so flexible that virtually any shape can be approximated, e.g. a series of orthogonal polynomials, cannot be considered as a peak model in the sense that it generalizes chromatographic phenomena. Such a function might be used for describing an unknown peak because the parameters in the function do characterize the peak uniquely. However, it is more convenient to characterize the peak by a set of quantities that can be calculated directly, the "moments" [2–10]). Let the observed signal be $f(t)$. The area under the peak is

$$A = \int_{-\infty}^{\infty} f(t)\,\mathrm{d}t.$$

In principle the integration should be carried out over the entire time axis, but in practice integration is restricted to the interval in which $f(t)$ differs noticeably from zero. The peak location is indicated by the centre of gravity:

$$\mu = \frac{1}{A}\int_{-\infty}^{\infty} t f(t)\,\mathrm{d}t. \tag{2.3}$$

A characterization of the shape which is independent of the area and the location is given by the "central moments" or "moments around the mean". The nth central moment is defined as

$$m_n = \frac{1}{A}\int_{-\infty}^{\infty} (t-\mu)^n f(t)\,\mathrm{d}t. \tag{2.4}$$

The area and the gravity centre are often called the zeroth and first moments. The second central moment is known as the variance and its square root as the standard deviation. It has several advantages to characterize the shape by dimensionless numbers, such as

$$\begin{aligned}
&\text{— the plate number } N = \frac{\mu^2}{m_2},\\
&\text{— the skew: } \gamma_1 = \frac{m_3}{m_2^{3/2}},\\
&\text{— the excess: } \gamma_2 = \frac{m_4}{m_2^{\,2}} - 3.
\end{aligned} \tag{2.5}$$

The interpretation of the higher moments in analytical information, e.g. physicochemical quantities such as diffusion coefficients, is still intricate [2–8]).

It is possible to calculate the moments of the peak models mentioned above by integration of eq. (2.4). This is summarized in table 2-I. For the higher moments of the Gaussian function the following relation can be derived:

$$m_n = (n-1)\, w^2\, m_{n-2}.$$

The parameters γ_1 and γ_2 in the Edgeworth series turn out to be identical to the skew and excess in (2.5).

Summarizing, in this section we discussed various ways to characterize the peaks. An unknown peak may be characterized by its moments or by a series of orthogonal functions, e.g. Hermite polynomials. However, often a peak model must be assumed, e.g. to dissect overlapping peaks or to distinguish peaks from the background. Suitable peak models are compiled in table 2-I.

2.2. *Background models*

Background is used here as a collective term for all components of the signal that carry no analytical information. These include both deterministic and random components. We shall first discuss their origin.

The signal of a chromatograph originates from the measurement of a certain physical property of the column effluent by the detector. A steady contribution comes from the carrier gas, including impurities, and the bleed of the stationary phase. With instrumental instabilities, e.g. due to flow controllers or thermostat, or with programmed changes of conditions, this contribution will be fluctuating or slowly drifting.

A second contribution may come from the injected substances. In trace analysis the column is often overloaded with solvent. The solvent peak is characterized by a steep front and an extended tail on which the trace peaks are superimposed (cf. fig. 2.3*a*). Usually the solvent peak is irrelevant and, therefore, considered as part of the background. In natural samples, as a rule of thumb, the number of compounds above a certain concentration is inversely proportional to the ratio of this concentration over the concentration of the largest peak. If the level at which the chromatogram becomes crowded with peaks is above the level of other noise contributions, the conglomerate of overlapping peaks forms a "hilly" background, called "compound noise" (cf. fig. 2.3*b*).

A third contribution is of electric origin. The noise from detector, amplifier, power supplies, etc., is mainly high-frequency noise. Spikes and steps may also be present. Noise from a flame-ionization detector is known to vary with the signal amplitude.

A fourth contribution is due to the recording. In digital sampling the signal

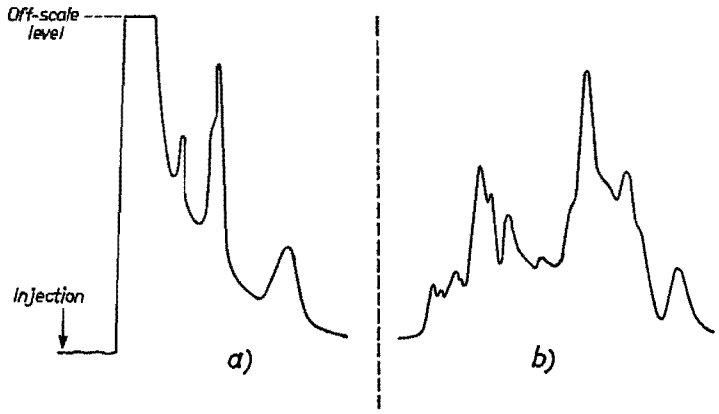


Fig. 2.3. Background contributions from the injected sample: (*a*) solvent peak, (*b*) compound noise.

will be distorted in two ways [2-11]). First, sampling means that the signal is only known at distinct times. Shannon's criterion [2-12]) states that the sampling rate must be higher than twice the highest frequency contained in the signal, otherwise the higher frequencies are "folded" over the lower, leading to distortions. Second, digitization means that the signal is rounded to the nearest discrete value, introducing an error of about half the least-significant digit unit.

A given background can be divided into random noise and a non-random baseline, as illustrated in fig. 2.4*a*. The baseline is usually approximated by a

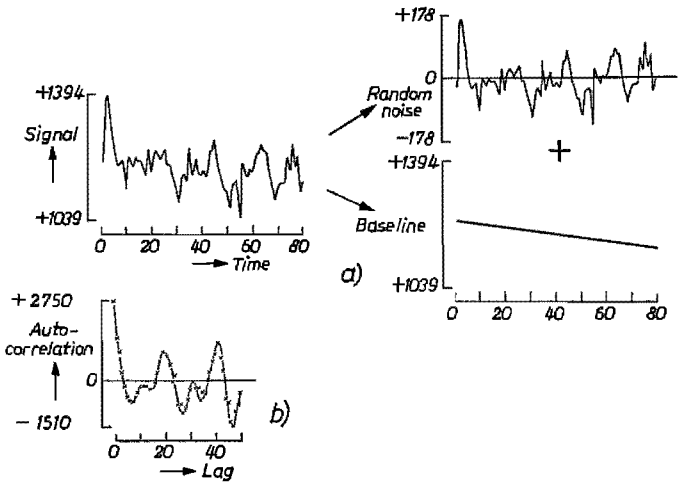


Fig. 2.4. Separation of a background trace in a polynomial baseline and random noise; (*a*) a real trace of background is approximated by a first-degree polynomial and random noise, (*b*) autocorrelation of the random noise.

low-degree polynomial, either a single function over the entire chromatogram or piecemeal functions if the background is discontinuous or wavy.

Random noise is described by its statistical characteristics, see e.g. Bendat and Piersol [2-12]). Since deterministic components are comprised in the baseline, the noise has a zero mean. Suppose a random noise signal $n(t)$ is sampled after intervals Δ: $N_k = n(k\,\Delta)$. The autocorrelation function $R(\tau)$ shows the average dependence of the noise amplitudes at a time lag τ:

$$R(d\,\Delta) = R_d = \langle n(t)\, n(t + d\,\Delta)\rangle \approx \frac{1}{m} \sum_{k=1}^{m} N_k\, N_{k+d}. \qquad (2.6)$$

The approximation is the more precise the larger the number of samples, m. Figure 2.4*b* shows the autocorrelation of the random noise in 2.4*a*. By definition, R_d has a maximum for $d = 0$; this value

$$R_0 = \frac{1}{m} \sum_{k=1}^{m} N_k^2$$

is called the variance or power of the noise. Its square root is called the (mean) amplitude. The magnitude of R_d, relative to R_0, indicates how strongly samples over the interval $d\,\Delta$ are related: if R_d differs appreciably from zero, this means that if the value at a certain moment is known, the value $d\,\Delta$ later is to some extent predictable. Unless the noise has some periodic component, e.g. caused by a thermostat cycle, the autocorrelation drops down to zero from the maximum R_0. A special type of noise is "white noise", for which $R_d = 0$ for all $d > 0$. The assumption that noise of successive samples is uncorrelated is a simple and very convenient noise model.

So far, an explicit noise model has not been applied in chromatography, except by Kelly and Harris [2-1]) who used the power-density spectrum. This is the frequency-domain equivalent of the autocorrelation function. These authors also discussed the validity of the assumption of stationary noise, which was implicitly made above.

2.3. *Peak detection*

With no prior knowledge about the locations or sizes of the relevant peaks, the amount of extracted information increases with the number of detected peaks. Hence the aim to detect as many peaks as possible, including trace peaks and overlapping peaks. Pushing this too far is likely to yield spurious peaks due to noise or to assign single peaks erroneously as composite. Clearly, the more the available knowledge about the peak shape is used in the detection, the better genuine peaks can be sorted out. This knowledge includes that peaks are more or less Gaussian, having a width varying approximately linearly with the location.

Existing detection methods in chromatography do not exploit much of the knowledge about the peak shape and the background. Conceiving a peak as "a signal that goes up and comes down", commonly the first or second derivative of the signal is compared with a threshold. A peak is detected if the threshold is exceeded.

Instead of deciding on the presence or absence of a peak on the momentary values of the signal or its derivatives, it appears more sensible to scan the signal for profiles that are congruent to the model profile. Let $g(\tau)$ denote the reversed standardized — i.e. unit area and located in the origin — peak model. It is known from communication theory [2–13]) that optimum peak detection in a signal with superposed uncorrelated noise is achieved by what is known as matched filtering, i.e. convolution of the signal $y(t)$ and $g(\tau)$:

$$z(t) = \int_{-\infty}^{\infty} y(t-\tau)g(\tau)\mathrm{d}\tau.$$

Matched filtering is a very selective means of distinguishing peaks from other components in the signal: a sort of resonance occurs where the local signal profile matches the shape of the peak model. Maxima in the filter output $z(t)$ are likely peak locations.

Detection is essentially a decision on the presence or absence of a peak. Two types of errors may be committed:

— a "miss" by deciding that a peak is not present when it is;

— "false alarm" by deciding that a peak is present when it is not.

Clearly, these errors are mutually antagonistic: if one wants to avoid a miss, the decision in favour of the presence will be made at the slightest indication, incurring many false alarms. A rule for decisions will therefore balance the "costs" associated with each type of error. Several such rules exist, differing in the adopted criterion (cf. ref. 2–13). For our purpose there is no clear appreciation of the costs. Since the aim was to detect as many peaks as possible, it appears sensible to minimize the probability of a "miss" for a fixed (small) probability of "false alarm".

Two types of detection limits should be distinguished. The first type is the limit at which a peak disappears in the background. The second limit specifies when overlapping peaks come so close that the composite peak cannot be distinguished from a single peak.

The detection limit due to background noise has not received much attention in chromatographic-data processing. Commonly an inflated threshold serves to suppress all minor peaks. In sec. 3.2.2 we will optimize the matched-filter detection, using a threshold based on the noise amplitude of the actually processed signal. The pertinent minimum detectable amount is derived.

To distinguish between single peaks and composite peaks is, to a large extent,

arbitrary because true single peaks do not exist. The point at which one starts to consider a peak as composite depends on the peak concept. Usually, the existence of two maxima is taken as the evidence of a composite peak. It has also been proposed to take the existence of more than two inflection points on the composite curve as evidence of overlap [2–14]). This criterion is more sensitive and can also detect shoulder peaks. By making more-detailed assumptions about the peak shape, the detection limit for composite peaks can be pushed further down [2–15,16]). However, this is rather speculative as the accuracy of the peak model cannot be checked due to the composite nature of all real peaks and the specificity of each shape.

2.4. *Baseline correction*

Baseline correction directly affects the peaks and is therefore of paramount importance to the quality of the peak parameters. Surprisingly, many investigations on accuracy and precision of parameter estimation [2–1,17,18,19]) did not report the method of baseline correction.

Commonly the baseline correction is made as follows: having determined the peak positions, the boundaries of each peak are located. The segments outside the peak regions, labelled b in fig. 2.5, are used to fit a baseline. This may be a continuous function over the whole chromatogram (global baseline, fig. 2.5 above) or a piecemeal correction to each peak or peak group (local baseline).

A global baseline is based on more data points and therefore less sensitive to erroneous location of the peak boundaries. However, it is difficult to fit a global baseline to a wavy or discontinuous background.

A local baseline is better suited to changing conditions but it hinges on the correct location of the peak boundaries. A simple constant or linear function will give a sufficiently accurate approximation. Commonly [2–20,21,22]) the minima before and after the peak are taken as boundaries and these are connected by a straight line, as shown in fig. 2.6*a*. A cwm between overlapping

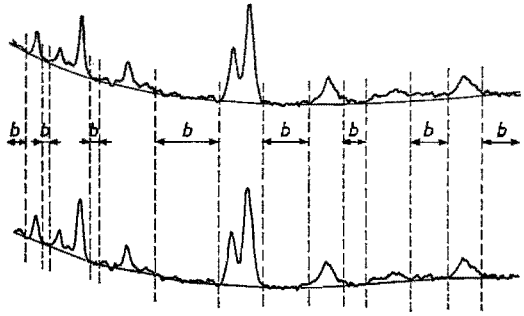


Fig. 2.5. Baseline approximation by fitting a function to the background segments, labeled b. A global baseline is a single function over the whole chromatogram (above). A local baseline is a piecemeal approximation to the segments bracketing each peak group (below).

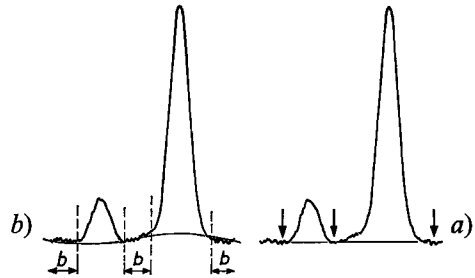


Fig. 2.6. Two methods of local baseline correction; (*a*) connection of valleys in a chromatogram, (*b*) least-squares fitting of a polynomial to the background segments.

peaks must be skipped [2-22]). Baan [2-23]) fitted a polynomial to the baseline segments bracketing a peak group, raising the degree of the polynomial until a satisfactory fit is obtained (cf. fig. 2.6*b*). The peak boundaries were determined by a threshold for the second derivative.

Considering these approaches it appears that the local-baseline approximation is able to cope with a greater variety of chromatograms and therefore better suited to our aims. The fitting of a polynomial will give a more accurate approximation as it is based on more data points and able to follow a curvature in the background. The crucial problem is the location of the peak boundaries. The minima before and after the peak are incorrect on a sloping baseline. A threshold for the signal or its derivatives also yields incorrect boundaries. We will elaborate an improved method for locating the boundaries.

2.5. *Parameter estimation*

Determinate and random errors from all stages of a chromatographic analysis, including sampling and injection, separation, recording and signal processing, are accumulated on the peak-parameter estimates. The errors in sampling and separation were comprehensively investigated by Rijks [2-24]). The errors in recording were discussed by Kelly and Horlick [2-11]). In studying the errors in signal processing we will ignore the errors from previous stages, although it should be kept in mind that however clever the processing, the inherent errors cannot be reduced and thus may eventually be the quality-determining factor.

A distinction should be made between systematic error and random error, or, as it is usually called, between accuracy and precision. Accuracy is a measure how close a result comes to the true value, neglecting the spread due to random errors. Precision is a measure how exactly the result is determined and thus characterizes the spread. In practice accuracy and precision are often mixed up with reproducibility, as a way to determine the spread is to repeat the experiment. This is not readily possible for studying the errors in signal processing as repeating the analysis will also change the errors in previous stages. However, suppose that a particular analysis can be repeated several times under virtually

constant conditions, so that the true values for each chromatogram are identical. The results will spread due to the non-reproducible, stochastic components in the signal. The standard deviation of the distribution of the results, σ, is thus determined by the noise. Conversely, if the statistical characteristics of the noise are known, it is possible to estimate the standard deviation without repeating the analysis:

Let some parameter p be a function of n data $Y_1, Y_2, \ldots, Y_n$: $p = g(Y_1, \ldots, Y_n)$. We will assume that the random noise superposed on the data is "white", i.e. uncorrelated, having mean amplitude $\sigma_y = \sqrt{R_0}$ (cf. (2.6)). The standard deviation σ_p of the parameter p, follows from the standard deviation of the data according to the so-called error-propagation expression:

$$\sigma_p^2 = \sigma_y^2 \sum_{k=1}^{n} \left(\frac{\partial p}{\partial Y_k}\right)^2. \tag{2.7}$$

Therefore, if the amplitude of the noise is known, the standard deviation of the parameters can be calculated without repeating the analysis.

Another source of error is associated with the estimation procedure. For example, if the peak top is located at the highest data point over the peak, the random error in the top location will be about one quarter of the sample interval, even in the absence of noise.

Often the random error and the systematic error have an antagonistic character. The estimation procedure can be designed to achieve a compromise. Consider fig. 2.7 of a Gaussian peak with superposed white noise. If the area of the area of the peak is determined by numerical integration of the sampled signal Y_i, i.e.

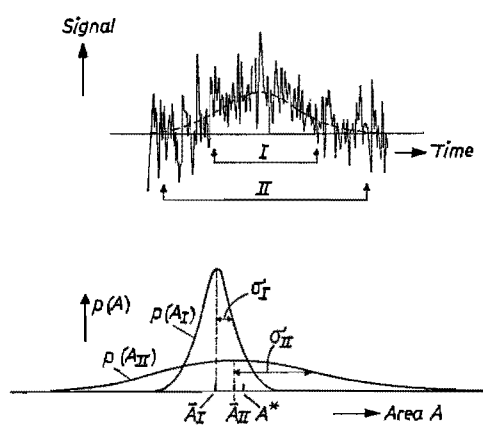


Fig. 2.7. Effect of the integration limits on the accuracy and precision of the estimated peak area A. On extending the integration limits (I → II), the mean of the probability distribution $p(A)$ shifts closer to the true value A^* ($\bar{A}_I \to \bar{A}_{II}$), but the standard deviation increases ($\sigma_I \to \sigma_{II}$).

$$A = \sum_{i=a}^{b} Y_i \Delta,$$

and the integration interval (a, b) and sample interval Δ are fixed, the lost area outside the boundaries causes a systematic error. The accuracy increases on extending the integration limits, but the precision decreases, as it follows from (2.7) that

$$\sigma_A^2 = \sigma_y^2 (b - a) \Delta^2.$$

As remote samples do not contribute substantially to the area, it is obvious that after some distance the gain in accuracy is overridden by the loss in precision. The integration limits may be adjusted to obtain a compromise.

Generally, let the systematic error ("bias") $\mu_p - \mu_p^0$ and the standard deviation $\sigma_p(K)$ of some parameter p be a function of a variable K (in the above example the integration limits may be at a distance K from the top). The mean squared error is

$$\sigma_p^2(K) + (\mu_p - \mu_p^0)^2.$$

The error is minimized if

$$\sigma_p \frac{\partial \sigma_p}{\partial K} + (\mu_p - \mu_p^0) \frac{\partial \mu_p}{\partial K} = 0.$$

Usually μ_p^0 is unknown, so that this is not a practical condition. In a region where both the bias and the standard deviation are monotonic functions, it seems sensible to choose K so that

$$\frac{\partial \sigma_p}{\partial K} - \left| \frac{\partial \mu_p}{\partial K} \right| = 0. \tag{2.8}$$

In the case of area integration this condition implies that the integration limits are extended until the value of the integral changes less than that of the standard deviation.

Accuracy and precision have been mainly studied experimentally from computer-generated peaks and random noise [2–17,18,19]). We will apply eq. (2.7) to obtain analytical expressions for the errors. These have the advantage over experimental relationships that the role of involved factors is better understood and, most important, that they can be used in eq. (2.8) to calculate a balance between accuracy and precision. The opposing trend in accuracy and precision has been recognized previously [2–17]), but no explicit condition for a trade-off has been reported.

Two methods of peak characterization were distinguished in sec. 2.1.

First a given peak can be characterized by its moments, without assuming anything about the shape.

The second method is to determine the parameters in a function that simulates more or less real peaks. This technique is usually called "curve fitting". The aim is to determine the parameters $\mathbf{p} = (p_1, \ldots, p_m)$ in a function $f(t, \mathbf{p})$ so that it fits best to the set of data $Y_1, \ldots, Y_n$ over the peak (Y_k is the baseline-corrected signal value at $t = t_k$). It can be shown that with white noise on the data it is sensible to minimize the sum S of the squared discrepancies between the function values and the data:

$$\min_{\mathbf{p}} S = \sum_{k=1}^{n} [f(t_k, \mathbf{p}) - Y_k]^2; \tag{2.9}$$

S is thus a function of the parameters $\mathbf{p}$. A necessary condition for the minimum is that the partial derivatives of S to the parameters are equal to zero:

$$\frac{\partial S}{\partial p_i} = 2 \sum_{k=1}^{n} [f(t_k, \mathbf{p}) - Y_k] \frac{\partial f(t_k, p)}{\partial p_i} = 0, \qquad i = 1, \ldots, n. \tag{2.10}$$

These are called the normal equations. The stated condition is necessary but not sufficient, because it also holds for a maximum in S and does not distinguish between local minima and the global minimum. Some problems encountered in curve fitting are:

— The choice of a suitable peak model that is both general and specific. This subject was already broached in sec. 2.1.

— To find a procedure which solves condition (2.9) or (2.10). As peak models are non-linear, the optimum parameter values must be approached iteratively, starting from some initial estimates. A good algorithm for optimization should require few amounts of computation and storage, and assure that the optimum values are attained.

— According to eq. (2.9) the optimum parameter values minimize the sum of squares. Whether this is physically a sensible condition depends on the correctness of the assumed peak model and noise model. Often additional relationships are known between the parameter values (e.g. peak width increases with retention time) or the values are restricted to some feasible regions (e.g. peak areas are positive). The formulation and the way to account for the constraints is described in chapter 3.

— Suppose that the optimal parameters $\hat{\mathbf{p}}$ have been found that yield a global minimum and satisfy all constraints. If the peak model is correct, the residues

$$E_k = Y_k - f(t_k, \hat{\mathbf{p}}), \qquad k = 1, \ldots, n,$$

should be entirely due to the noise on the data. Hence the residues should show almost the same characteristics as the random noise before and after the peaks. If this is not the case, the model is likely to be incorrect. A problem is how it can be deduced from the shape of the residues in which way the model should be modified to improve the fit.

— If the optimal parameters give a satisfactory fit, it is interesting to know how significant the values are. General expressions for the standard deviation of the parameters are known. We will apply these general expressions to some particular peak models to obtain analytical forms for the errors in the parameters. These errors are compared with the errors from the moment calculation.

Since the article by Fraser and Suzuki [2-25]) curve fitting has predominantly been applied for the calculation of the parameters of overlapping peaks [2-14,23,26,27]). The fitting function is then the sum of several displaced peaks with different areas, widths, etc. The standard deviation of the parameters has never been studied in chromatographic applications. Therefore several wrong ideas have pervaded:

— Littlewood [2-26]) thought about replacing chromatography as far as possible with mathematics, by using shorter columns. The goal was to "reduce the column to the point of virtual extinction". We will show that the standard deviation of the peak parameters increases rapidly with decreasing resolution.

— Chesler and Cram [2-18]) alleged that their complex peak model was excellently suited as a model for dissecting overlapping curves. However, this general model will give excessively large covariances of the parameters, which implies that the precision is very low.

— Attempts have been made to push the detection limits down, by moments analysis [2-15]) or slope analysis [2-16]), because it was believed that this was the limiting factor in the dissection of overlapping peaks. Apart from being unpractical, these attempts are also rather meaningless, as the parameters of closely spaced peaks cannot be determined precisely.

2.6. *Filtering*

Filtering is understood here as a certain operation on the signal. The aim of filtering may be to attenuate the random noise ("smoothing") or to obtain some derivative of the signal (differentiation). Since Savitzky and Golay [2-28]) gave a clear statement of the filtering problem, the polynomial filters proposed by them have gained an unassailed monopoly in the processing of analytical data. While these are convenient general-purpose filters, other filters are known which cause less distortion for the signal at hand or require less amounts of computation. We shall briefly discuss some types of filters.

As the signal is available in sampled and digitized form, our main interest lies

in the properties and design of digital filters. However, the result of the filtering is often more easily understood from the analogue form, as our peak models are also continuous.

Let Y_i, $i = 1, \ldots, n$, be the equidistantly sampled signal. The operation of a symmetric linear filter can be written as

$$Y_i^* = \sum_{k=-m}^{m} F_k Y_{i-k}, \qquad m \geqslant 0. \tag{2.11}$$

In this expression, F_k is the weighting factor for the contribution of Y_{i-k} to the filtered signal at point i, Y_i^*. Expression (2.11) is the discrete form of a convolution of the (unsampled) signal $y(t)$ and a filter function $f(\tau)$:

$$y^*(t) = \int_{-\infty}^{\infty} y(t-\tau) f(\tau)\, \mathrm{d}\tau. \tag{2.12}$$

If the sampling interval Δ is small, it is usually valid to assume that, if $Y_i = y(i\Delta)$ and $F_k = \Delta f(k\Delta)$ then $Y_i^* \approx y^*(i\Delta)$. This is very convenient because (2.12) is more easily evaluated when $y(t)$ is a peak model. On the other hand, (2.11) leads to a very simple result when Y_i is purely white noise. Symmetrical filtering according to (2.11) or (2.12) can only be performed off-line or on a delayed signal.

The filter (2.11) is a non-recursive filter because it operates only on the input signal. If the filter also acts on its own output, it is said to be recursive. A linear recursive filter is specified by

$$Y_i^* = \sum_{j=1}^{p} G_j Y_{i-j}^* + \sum_{k=-m}^{m} F_k Y_{i-k}, \qquad m, p \geqslant 0. \tag{2.13}$$

Often a filter can be put in both forms, e.g. the moving average:

$$Y_i^* = \frac{1}{2m+1} \sum_{k=-m}^{m} Y_{i-k} = Y_{i-1}^* + \frac{1}{2m+1} [Y_{i+m} - Y_{i-m-1}]. \tag{2.14}$$

The non-recursive form requires the summation of $2m + 1$ terms, whereas the recursive form requires only the summation of 3 and gives therefore great computational savings with a high m. In some cases the recursive filter achieves results for which a simple filter would require a very large or potentially infinite number of operations. Consider the discrete form of the analogue exponential filter:

$$Y_i^* = \sum_{k=0}^{i-1} \frac{1}{\tau} \exp(-k/\tau)\, Y_{i-k} = \frac{1}{\tau} Y_i + \exp(-1/\tau)\, Y_{i-1}^*. \qquad (2.15)$$

Here, the non-recursive form requires i multiplications and additions against just 2 for the recursive form. For large τ,

$$\exp(-1/\tau) \approx 1 - \frac{1}{\tau},$$

yielding

$$Y_i^* = Y_{i-1}^* + \frac{1}{\tau}(Y_i - Y_{i-1}^*). \qquad (2.16)$$

This equation states that the old filtered value is updated by a fraction of the difference between itself and the new sample. The time constant τ controls the degree of smoothing, e.g. if $\tau = 1$ there is no smoothing. Expression (2.16) can be generalized by allowing τ to be a function of i. For example, if $\tau = i$ we have an expression for the "current mean":

$$Y_i^* = Y_{i-1}^* + \frac{1}{i}(Y_i - Y_{i-1}^*) = \frac{1}{i} \sum_{k=1}^{i} Y_k. \qquad (2.17)$$

Having given some forms of linear filters, we will now discuss the effect of these filters on the deterministic components and the random component in the signal. For a linear filter these effects are independent. Among the transformations of the signal that a linear filter can perform, we are mainly interested in smoothing and differentiation.

Smoothing is effected by replacing the sampled value at a point by a weighted mean of the neighbouring samples. To leave a constant signal unaffected, the weights of a smoothing filter must satisfy the relation

$$\sum_{k=-m}^{m} F_k = 1 \qquad \text{or} \qquad \int_{-\infty}^{\infty} f(\tau)\, d\tau = 1. \qquad (2.18)$$

The first row in fig. 2.8 illustrates various smoothing filters; a way to calculate their effect on the noise is discussed below.

For obtaining the derivative of a continuous signal $y(t)$, we consider the following: let $f(\tau)$ be a function satisfying (2.18); if instead of convoluting $y(t)$ with $f(\tau)$ it is convoluted with the first derivative $f'(\tau)$, integration by parts shows:

$$y^*(t) = \int_{-\infty}^{\infty} f'(\tau)\, y(t-\tau)\, d\tau = -\int_{-\infty}^{\infty} f(\tau)\, y'(t-\tau)\, d\tau. \qquad (2.19)$$

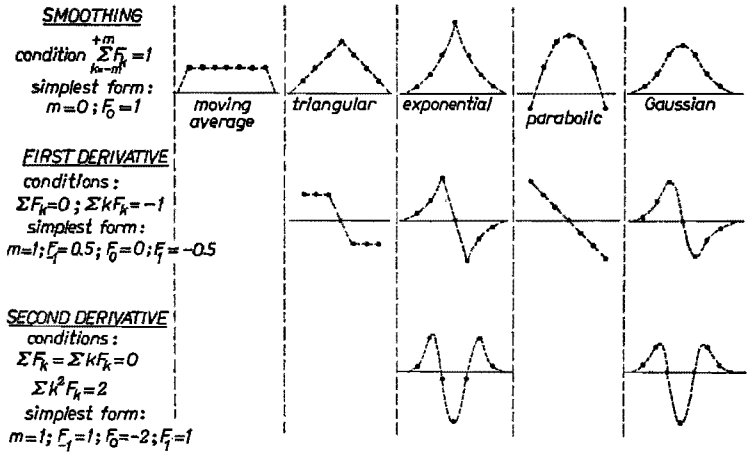


Fig. 2.8. Profiles of digital filters for smoothing and differentiation. Filters in one column are derivatives of the smoothing filter.

Thus the result of convoluting $y(t)$ and the derivative of $f(\tau)$ is identical with the smoothing of $y'(t)$ by the filter-response function $f(\tau)$. Or, in order to obtain the smoothed derivative of the signal $y(t)$, we merely have to convolute it with the derivative of a smoothing filter function. A number of digital filters for calculation of smoothed derivatives is illustrated in fig. 2.8, middle row. The requirement that a linear function is differentiated correctly implies that the filter weights must satisfy the conditions

$$\sum_{k=-m}^{m} F_k = 0; \qquad \sum_{k=-m}^{m} k F_k = 1. \tag{2.20}$$

The same reasoning can be made for the higher derivatives: a quasi n-fold differentiation of the signal $y(t)$ is performed by convoluting $y(t)$ and the nth derivative of $f(\tau)$. The digital form must satisfy $n+1$ conditions:

$$\sum_{k=-m}^{m} k^i F_k = 0, \qquad i = 0, 1, \ldots, n-1$$

and

$$\sum_{k=-m}^{m} k^n F_k = n!$$

Some filters for calculation of the second derivative are illustrated in fig. 2.8, last row.

The effect of a linear filter on random noise can be seen from the autocorrelation function, as defined in eq. (2.6). Let N_i, $i = 1, \ldots, m$, be the sampled

noise signal, characterized by the autocorrelation function

$$R_d = \frac{1}{m} \sum_{i=1}^{m} N_i N_{i+d}.$$

The filtered signal

$$N_i^* = \sum_{k=-a}^{a} F_k N_{i-k}$$

is characterized by the autocorrelation function

$$R_d^* = \frac{1}{m} \sum_{i=1}^{m} N_i^* N_{i+d}^* = \sum_{i=-2a}^{2a} R_{d+i} \sum_{k=\max(-a,i-a)}^{\min(a,a+i)} F_k F_{k-i}$$

$$= \sum_{i=-2a}^{2a} R_{d+i} \psi_{-i}. \qquad (2.21)$$

This result shows that the autocorrelation function is transformed by a similar linear "filtering" operation. If the original noise is white, i.e. $R_{d\neq 0} = 0$, eq. (2.21) reduces to

$$R_d^* = R_0 \sum_{k=\max(-a,-d-a)}^{\min(a,a-d)} F_k F_{k+d}. \qquad (2.22)$$

This result is important in two aspects. First it allows the mean amplitude of the noise after filtering to be calculated:

$$R_0^* = R_0 \sum_{k=-a}^{a} F_k^2. \qquad (2.23)$$

The noise attenuation of the filter is thus determined by the sum of the squared filter weights. It can be shown that for a fixed number of weights, the moving average has the greatest noise attenuation. Secondly it shows that a filtered white-noise signal will be correlated. The reverse, i.e. filtering in such a way that after filtering the noise is white, is called "whitening".

2.7. *Outline of the program*

Having made a survey of the elements of the data extraction, it must be considered how to organise these elements in a smoothly running program. For this purpose it may be inspiring to look at the way how a chromatogram is processed by a skilled analyst. This way is found to be a curious mixture of training, experience, a priori information and research. It appears that the processing does not proceed according to a fixed scheme, but by adapting a basic strategy to the information obtained in the course of the processing. Several stages of detailing can be distinguished. An initial scan reveals a quali-

tative impression of peak density, peak shape, signal-to-noise ratio, baseline trend, etc. These features serve to typify the chromatogram. Having gained an overall impression, attention is given to segments in order to recognize characteristic configurations, like solvent peak, overlapping peaks, shoulders, etc. Finally the individual peaks are scrutinized and qualitative impressions are quantified by interpolation, approximation and measurements.

The process of perception and mental processing is quite complex and is only roughly described as a cyclic sequence of sensing, hypothesis casting, search for supporting evidence, hypothesis modification and decision making. Generally, the processing is very powerful in the qualitative aspects, because it is able to cope with a wide variety of chromatograms and insufficient prior information is compensated by oriented research, hypothesis testing and decision making. Each chromatogram receives a matched treatment. On the other hand, the quantitative aspects are rather poor. The precision is limited owing to the inability or reluctance to do large amounts of measurements and calculations. For example, rapid but imprecise geometrical constructions such as tangents or perpendiculars are often preferred to numerical integration. Another drawback is that many arbitrary decisions are made, so that the processing and the results are irreproducible.

Imitation of this approach in a computer program would result in a very complex program: a few alternatives in each decision rapidly lead to a combinatorial explosion. In order to keep the program manageable — in size, in time and in mind — the number of alternatives must be restrained. This means that one program structure useful for achieving adaptivity, viz. pathway selection (fig. 2.9*a*), should only be considered for incompatible processing modes. A requirement for dynamic setting of processing controls (fig. 2.9*b*) is that the relation between the characteristic of the signal, e.g. S/N ratio, and the appropriate control setting, e.g. threshold value, is well defined. Iterative approximation (fig. 2.9*c*) is the method to be used if the relation between signal characteristics and optimum controls is not well defined, but some criterion for judging the quality of the processing is available. The optimum is approached by repeated adjustment of the controls. It is now important that the iteration will converge. Algorithms of this type are treated in detail by Tsypkin [2–29]). These three structures can be applied for small processing steps or wide ranging operations. By inserting one into another a powerful adaptive program structure can be obtained.

The usual way to design a program for solving a certain task is to make a top–down decomposition. This means that the task is divided into a number of sub-tasks, which are again decomposed in simpler sub-tasks, etc. The object of this decomposition is to arrive either at basic operations that can be programmed straightforwardly or at standard procedures for which ready-made algorithms are available. The decomposition is rather functional than time-

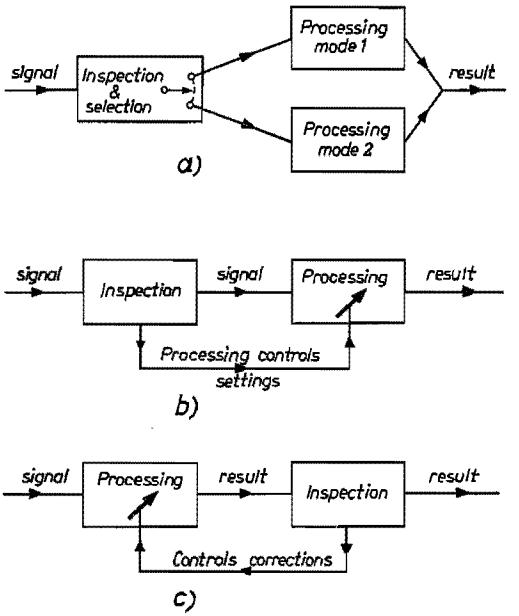


Fig. 2.9. Adaptive programming structures: (*a*) pathway selection, (*b*) dynamic setting of processor controls, (*c*) iterative approximation.

sequentional, that is, attempts are made to arrive at functionally simple sub-tasks. To arrive at an efficient and flexible program it may be worthwhile to structure the sub-tasks in a different way. Accordingly, the original top–down decomposition is complemented by a bottom–up assemblage which is not necessarily isomorphic.

In our program the processing is performed in three stages, viz. inspection, detection and estimation:

— By inspection some of the lacking information about the signal is obtained. The mean noise amplitude should be known for setting a threshold level in the peak detection. The peak width should be known for matched-filter detection.
— Peak detection locates the positions of the peaks, including trace peaks and overlapping peaks. Spikes, which interfere in the peak detection, are filtered out first. Peak boundaries are located so that peak regions can be separated from baseline segments.
— Estimation includes the baseline correction and the peak-parameters estimation. The latter can be done by calculation of the moments or by curve fitting.

Figure 2.10 shows a flow chart of the program. The modules, indicated by blocks in fig. 2.10, will be designed and described in the next chapter.

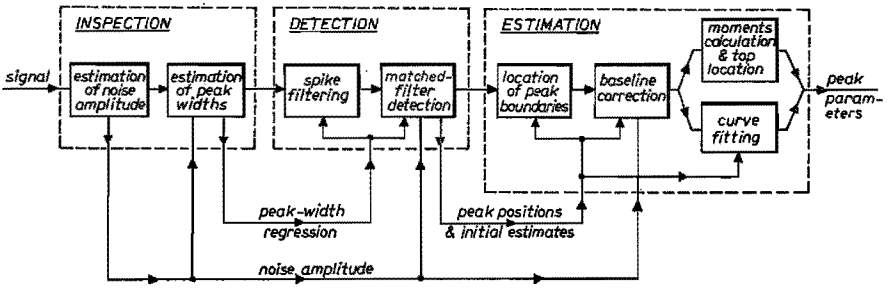


Fig. 2.10. Flow chart of the program for processing of a chromatographic signal.

REFERENCES

2-1) P. C. Kelley and W. E. Harris, Anal. Chem. **43**, 1170, 1971.
2-2) J. J. M. Wijtvliet, Thesis Technological University Eindhoven, 1972.
2-3) G. Brouwer and J. A. J. Jansen, Anal. Chem. **45**, 2239, 1973.
2-4) A. I. M. Keulemans, Gas chromatography, Rheinhold, 1959, p. 120.
2-5) G. McWilliams and H. C. Bolton, Anal. Chem. **41**, 1755, 1969.
2-6) R. D. B. Fraser and E. Suzuki, Anal. Chem. **41**, 37, 1969.
2-7) H. Cramér, Mathematical methods of statistics, Princeton, 1946, p. 227.
2-8) O. Grubner, in J. C. Giddings and R. A. Keller (eds), Advances in chromatography, Dekker, New York, 1968, vol. 6, p. 173.
2-9) D. J. Kroon, Thesis University of Amsterdam, 1962, p. 119.
2-10) E. Grushka, N. M. Myers and J. G. Giddings, Anal. Chem. **41**, 889, 1969.
2-11) P. C. Kelly and G. Horlick, Anal. Chem. **45**, 518, 1973.
2-12) J. S. Bendat and A. G. Piersol, Measurement and analysis of random data, Wiley, 1966, p. 19.
2-13) A. D. Whalen, Detection of signals in noise, Academic Press, 1971, p. 126.
2-14) A. W. Westerberg, Anal. Chem. **41**, 1770, 1969.
2-15) E. Grushka, N. M. Myers and J. C. Giddings, Anal. Chem. **42**, 21, 1970.
2-16) E. Grushka and G. C. Monacelli, Anal. Chem. **44**, 484, 1972.
2-17) S. N. Chesler and S. P. Cram, Anal. Chem. **43**, 1922, 1971.
2-18) S. N. Chesler and S. P. Cram, Anal. Chem. **45**, 1354, 1973.
2-19) M. Goedert and G. Guiochon, J. chromatog. Sci. **11**, 326, 1973.
2-20) N. Guichard and G. Sicard, Chromatographia **5**, 83, 1972.
2-21) G. Schomburg and E. Ziegler, Chromatographia **5**, 96, 1972.
2-22) R. A. Landowne, R. W. Morosani, R. A. Herrmann, R. M. King and H. G. Schmuss, Anal. Chem. **44**, 1961, 1972.
2-23) A. Baan, Graduation Report, Technological University Eindhoven, 1971.
2-24) J. A. Rijks, Thesis Technological University Eindhoven, 1973.
2-25) R. D. B. Fraser and E. Suzuki, Anal. Chem. **38**, 1770, 1966.
2-26) A. B. Littlewood, T. C. Gibb and A. H. Anderson, in C. L. A. Harbourn (ed.), Gas chromatography 1968, Institute of Petroleum, London, 1969, p. 297.
2-27) S. M. Roberts, D. H. Wilkinson and L. R. Walker, Anal. Chem. **42**, 886, 1970.
2-28) A. Savitzky and M. J. E. Golay, Anal. Chem. **36**, 1627, 1964.
2-29) Ya. Z. Tsypkin, Adaptation and learning in automatic systems, Academic Press, 1971, p. 17.
2-30) E. Grushka, Anal. Chem. **44**, 1733, 1972.

3. ALGORITHMS

In this chapter the algorithms of the program, as outlined in sec. 2.7, are described. The fact that no coded algorithms are presented does not imply that coding is straightforward, even in a so-called high-level language. However, the difficulties of programming are not specific for the present work.

As our prime objective is to obtain optimum information extraction, this aspect will be stressed in the present treatment. To get efficient algorithms it was often necessary to structure them in a different way. Thus the actually implemented algorithms may have a different form.

3.1. Inspection

As illustrated in fig. 2.10, the inspection stage consists of the random-noise estimation and the initial-peak detection.

3.1.1. *Random-noise estimation*

Random noise can be described by its autocorrelation function, as defined by eq. (2.6). Its mean amplitude must be known in order to set a threshold in the peak detection. The mean noise amplitude, denoted as σ_y, is also used to calculate the standard deviations of the parameter estimates.

The random noise amplitude can be derived from the scattering of the sampled data around a local polynomial fit. Let $p_n(t)$ be an nth degree polynomial fitted to m data (t_i, Y_i), $i = 1, \ldots, m$. If $p_n(t)$ accounts sufficiently for the non-random components in the signal, the mean squared error provides an estimate for the power of the noise:

$$\sigma_y^2 \approx \frac{1}{m-n-1} \sum_{i=1}^{m} [Y_i - p_n(t_i)]^2.$$

The simplest case is to fit a zeroth-degree polynomial to two successive samples Y_{i-1} and Y_i:

$$p_0 = \tfrac{1}{2}(Y_{i-1} + Y_i).$$

Summing the mean squared differences over all data points yields the expression for the random-noise amplitude:

$$\sigma_y^2 = \frac{0{\cdot}5}{m-1} \sum_{i=2}^{m} (Y_i - Y_{i-1})^2. \tag{3.1}$$

A constant baseline is eliminated in this way. If the signal is purely random

noise, evaluation of (3.1) and substitution of (2.6) yields

$$\sigma_y^2 = R_0 - R_1.$$

This shows that for white noise the amplitude is accurately estimated.

As actually the squared first differences are summarized in (3.1), the result will be biased owing to large contributions of points on the peaks. These contributions can be eliminated iteratively, assuming that random noise is almost normally distributed. The differences $Y_i - Y_{i-1}$ will also be distributed normally, so that it may be assumed that differences which exceed the range between $-3\sigma_y/2$ and $+3\sigma_y/2$ are points on a peak. Hence the following algorithm (k is the iteration number):

(i) ($k = 0$): calculate from eq. (3.1) the mean amplitude σ_y^0;

(ii) recalculate, by discarding the differences outside the range $(-3\sigma_y^k/2, +3\sigma_y^k/2)$, yielding σ_y^{k+1};

(iii) if σ_y^{k+1} differs significantly from σ_y^k (more than 10%), k is raised by 1 and (ii) is repeated.

The procedure was tested with computer-generated uncorrelated noise. As expected, accurate estimates were obtained. For correlated noise, obtained by filtering the uncorrelated noise with a moving average filter, the bias was in close agreement with the predicted value. Superposition of peaks of varying width and size showed that the iterative procedure successfully discarded these contributions.

3.1.2. *Initial peak location*

Peak detection by matched filtering requires the knowledge of the peak shape and the peak widths. As all peaks have a more or less Gaussian shape, it is obvious to use this model. The peak width in chromatography increases with the location. Commonly a linear relationship is accepted as a reasonable approximation. In the following procedure the coefficients in the linear relation are estimated from the widths of the major peaks by regression. For a Gaussian peak the width is defined as the distance between the top and the inflection points. We will generalize this definition to other peak shapes. This implies that the width at the leading edge may be different from the width at the trailing edge.

The top of a peak is located where the first derivative (f.d.) changes its sign from positive to negative. The maximum of the f.d. before the top and the minimum after it are the inflection points. Let $Y_i^{(1)}$ denote the f.d. at the ith data point. The f.d. can be calculated efficiently with a linear differentiating filter:

$$Y_i^{(1)} = \frac{1}{\sum k^2} \sum_{k=-m}^{m} k\, Y_{i+k}. \tag{3.2}$$

This expression can also be put in a recursive form which is easier to calculate for large m:

$$Y_i^* = Y_{i-1}^* + \frac{1}{2m+1}(Y_{i+m} - Y_{i-m-1}),$$

$$Y_i^{(1)} = Y_{i-1}^{(1)} - \frac{1}{\sum k^2}[(2m+1)\,Y_i^* - m\,Y_{i-m-1} - (m+1)\,Y_{i+m}].$$

The filter width m controls the noise attenuation. According to eq. (2.23) the mean amplitude of the filtered noise is

$$\sigma_y^* = \sigma_y\left(\frac{1}{\sum k^2}\right)^{1/2} = \sigma_y\left(\frac{3}{2m^3+3m^2+1}\right)^{1/2}. \tag{3.3}$$

If the signal contains a linear baseline and random noise, the f.d. will be within the dual threshold $\pm 5\sigma_y^*$. The start of a peak is detected where the positive threshold is first exceeded; the end of the peak is where the f.d. returns from below the negative threshold. In between, the top and the inflection points can be found.

It may appear from eq. (3.3) that the noise amplitude can be attenuated at will by increasing the value of m. However, it can also be demonstrated that the extrema in the f.d. shift outwards on increasing m. This is undesirable for three reasons. First the apparent widths would be too large. Second, fusing of partially resolved peaks may occur. Third, the magnitude of the f.d. at the extrema decreases, which reduces the minimum detectable amount.

It can be shown that a value $m = w_p$ (peak width) gives a negligible broadening, while maximizing the signal-to-noise ratio in the f.d. At first, $m = w_p$ seems not a very practical condition because the peak width is unknown. However, by using an adaptive mechanism both the peak-width regression and the filter-width optimization can be done concurrently. Suppose a few peaks have been detected using a provisional value of m. They are plotted as points in a location–width (μ–w_p) graph, cf. fig. 3.1. One could now take a value for m corresponding to the linear regression to these points. However, if some of the peaks happen to be very broad, the too large value of m would broaden future peaks, inducing a still larger value of m, etc. It is necessary to take a safety margin. Assuming a Student T-distribution, the spread about the regression line is used to calculate a confidence interval (dashed curves) [3–1]. The lower curve is used to calculate the appropriate value of m and the point where this value can be updated. Upon detection of an additional peak, the regression line and the confidence limits are recalculated.

Outliers, due to composite peaks or spikes, bias the regression line and the confidence limits, thus blocking the filter updating mechanism. A test for out-

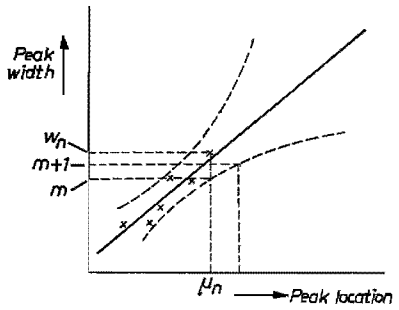


Fig. 3.1. Linear regression of the peak width and the peak location (drawn line). Crosses indicate peaks detected ($\mu_i - w_i$). The appropriate filter width m and the updating point ($m \rightarrow m + 1$) are calculated from the lower confidence limit (dashed curves).

liers compares the sums of squares from the regression including and excluding each peak. If significant differences are found, the peak is temporarily dropped (until the regression is repeated for a newly added peak).

Simulated chromatograms with both uncorrelated and correlated noise and peaks of various sizes, showed that the widths were estimated quite accurately. Spikes and broad composite peaks were effectively removed. Only in the case that all peaks were very broad, i.e. widths order of magnitude larger than the initial value of m, and the S/N ratio was low, did the procedure fail to detect any peaks, as the high threshold was never surpassed.

3.2. Detection

The detection stage is made up of two parts, viz. spike filtering and peak detection.

3.2.1. *Spike filtering*

A spike superposed on the baseline is easily distinguished from genuine peaks because its width is much smaller than expected from the linear regression. A spike superposed on a peak, however, distorts the profile of the output of the matched filter. It requires a complex detection logic if this type of distortion must be taken into account. It is easier to remove the spikes in advance by non-linear filtering. This means that the outlying value is replaced by a value interpolated from neighbouring samples. The procedure presented here is designed with special care not to discard sharp peak tops or steep peak sides.

Consider a segment of five successive samples: Y_{i-2}, Y_{i-1}, Y_i, Y_{i+1}, Y_{i+2}. To find out whether the central sample Y_i is a spike, a parabola,

$$f(i + k) = a + b\,k + c\,k^2,$$

is fitted to the neighbouring samples. The coefficients of the parabola are found as

$$a = \tfrac{1}{6}\,(-Y_{i-2} + 4Y_{i-1} + 4Y_{i+1} - Y_{i+2}),$$
$$b = \tfrac{1}{10}\,(-2Y_{i-2} - Y_{i-1} + Y_{i+1} + 2Y_{i+2}),$$
$$c = \tfrac{1}{6}\,(+Y_{i-2} - Y_{i-1} - Y_{i+1} + Y_{i+2}).$$

The discrepancy at the central point is denoted as $\Delta_i = Y_i - a$. For a signal of purely white noise, Δ_i is a statistical quantity with zero mean and standard deviation $\sigma_\Delta = 1{\cdot}4\,\sigma_y$. Two ratios serve as test quantities.

The ratio $R_1 = \Delta_i/\sigma_y$ compares the discrepancy with the mean noise amplitude. If the central sample is a spike, R_1 will be very large. For white noise on a smoothly varying baseline, R_1 will be almost normally distributed with zero mean and s.d. = 1·4, so that a spike threshold may be set at $|R_1| > 5$. However, R_1 may also be very large where the parabola gives a bad fit, as on sharp peak tops and steep peak sides. These cases are generally characterized by a strong curvature in the signal.

The second test quantity $R_2 = \Delta_i/c_i$ compares the discrepancy with the local curvature. A small value of R_2 indicates a strong curvature. A large value of R_2 is, in itself, not indicative of a spike, as c_i has zero mean for pure noise. However, if both R_1 and R_2 are large, a spike will be present. A threshold for R_2 is derived by considering a number of cases.

— The top of a sharp peak can be simulated by the case $Y_{i-2} = Y_{i+2} = 0$; $Y_{i-1} = Y_{i+1} = p$; $Y_i = 4p$ (p is an arbitrary value). This yields $\Delta_i = \frac{8}{3}p$, and $c = \frac{1}{3}p$. The condition for a spike is: $R_2 > 8$.

— To avoid that Y_i is taken as a spike if Y_{i+1} is a spike, the case $Y_{i-2} = Y_{i-1} = Y_i = Y_{i+2} = 0$; $Y_{i+1} = p$, yields $R_2 > 4$.

— A damped oscillation may be simulated as $Y_{i-2} = Y_{i-1} = Y_{i+2} = 0$; $Y_i = p$; $Y_{i+1} = fp$. This yields $R_2 = 4 - 6/f$. It is sensible to reject Y_i if $-1 < f < 0$, or, $10 < R_2 < \infty$.

Combining these results, a spike is detected if $|R_1| > 5$ and $R_2 > 10$.

A test with Gaussian peaks and superposed white noise showed that peaks having the width $w_p > 0{\cdot}4$ (sample interval units) were not affected.

3.2.2. *Peak detection by matched filtering*

The matched-filter detection was briefly outlined in sec. 2.3. Straightforward application is not possible because two other contributions must be taken into account, viz. random noise, with known amplitude, and an unknown deterministic baseline. In the following sections these factors are investigated.

3.2.2.1. Elimination of the baseline

The following method for elimination of the baseline is an elaboration of the detection by correlation used by Brouwer and Jansen [3-2]) for the processing of complex spectra. It is assumed that the baseline in the neighbourhood of a peak is approximately a linear function. Let $y'(t)$ denote the first derivative

of the signal. In the differentiated signal the assumed peak shape $g(\tau)$ is modified to $g'(\tau)$. Hence the modified matched filter becomes:

$$z^*(t) = \int_{-\infty}^{\infty} y'(t-\tau)\, g'(\tau)\, \mathrm{d}\tau. \tag{3.4}$$

Integration by parts in either direction proves the identity

$$z^*(t) = -\int_{-\infty}^{\infty} y''(t-\tau)\, g(\tau)\, \mathrm{d}\tau, \tag{3.5}$$

$$= -\int_{-\infty}^{\infty} y(t-\tau)\, g''(\tau)\, \mathrm{d}\tau. \tag{3.6}$$

Expression (3.5) shows that the effect of the modified matched filter is, in fact, a smoothing of the second derivative of the signal. A linear baseline is therefore eliminated.

Expression (3.6) shows that the two operations in (3.4), viz. differentiation of $y(t)$ and convolution, can be conveniently combined into a single operation, viz. convolution with the second derivative of the peak model.

So far no particular model was assumed. As all peaks are more or less Gaussian, it is obvious to use

$$g''(t) = \frac{1}{w_f{}^3\,(2\pi)^{1/2}} \exp\left[-\tfrac{1}{2}\left(\frac{\tau}{w_f}\right)^2\right]\left[\left(\frac{\tau}{w_f}\right)^2 - 1\right]. \tag{3.7}$$

We call w_f the width of the matched filter. For a sampled signal the operation (3.6) is replaced by a summation:

$$Z_k{}^* = \sum_{i=-m}^{m} G_i\, Y_{k-i}. \tag{3.8a}$$

The weighting factors are calculated from (3.7) (Δ is sampling interval):

$$G_i = \Delta\, g''(i\,\Delta). \tag{3.8b}$$

The value of m was set to $m = 4w_f/\Delta$; beyond this value the weights are almost zero.

3.2.2.2. Optimization of the filter width

Optimum detection is attained for the filter width that maximizes the signal-to-noise ratio after filtering, defined as

$$\mathrm{S/N}^* = \frac{\text{height of the filtered peak}}{\text{mean noise amplitude after filtering}} \tag{3.9}$$

(a marking "*" denotes an attribute of the filtered signal).

The mean noise amplitude before filtering, σ_y, was estimated in (3.1). Assuming the random noise to be uncorrelated, the attenuated amplitude after filtering is found, using (2.23):

$$\sigma_y^* = \sigma_y \left(\sum_{i=-m}^{m} G_i^2 \right)^{1/2} . \tag{3.10}$$

The sum of the squared filter weights may be approximated, using eqs (3.7) and (3.8b):

$$\sum_{i=-m}^{m} G_i^2 = \Delta \int_{-\infty}^{\infty} [g''(\tau)]^2 \, d\tau = \frac{\Delta}{8 \, w_f^5 \sqrt{\pi}} . \tag{3.11}$$

Let the signal contain a Gaussian peak, eq. (2.1), having area A and width w_p, i.e. $y(t) = g(t, A, \mu, w_p)$. Inserting this and (3.7) in (3.6) and evaluating the integral shows that the filtered signal is indeed the second derivative of this peak, but its width is increased to

$$w_p^* = (w_p^2 + w_f^2)^{1/2}.$$

Let $K = w_f/w_p$. The filtered peak is thus apparently broadened by a factor $(1 + K^2)^{1/2}$. The height of the filtered peak is equal to the magnitude of the second derivative at $t = \mu$:

$$\frac{A}{w_p^{*3} \, (2\pi)^{1/2}} .$$

Substitution of this, (3.10) and (3.11) in (3.9) yields (using $K = w_f/w_p$):

$$\mathrm{S/N^*} = \frac{A}{\sigma_y \, w_p \, (2\pi)^{1/2}} \left(\frac{8 \sqrt{\pi}}{3} \frac{w_p}{\Delta} \right)^{1/2} \left(\frac{K^5}{(1 + K^2)^3} \right)^{1/2} . \tag{3.12}$$

The first term in (3.12) is the signal-to-noise ratio before filtering. The second term contains the ratio $f = w_p/\Delta$ which may be seen as a dimensionless sampling density (number of samples over the peak width). The third term can be optimized. As a function of K this term reaches the optimum at $K = 5$, as graphed in fig. 3.2. This means that for optimum peak detection the width of the matched filter should be proportional to the width of the peaks in the signal, viz. $w_f = w_p \sqrt{5}$. Figure 3.2 shows that this is not a sharp optimum, so that a smaller value of K, which is advantageous from the point of resolution and amount of computation, does not entail a great loss, e.g. with $K = 1{\cdot}4$, S/N* is only 10% less.

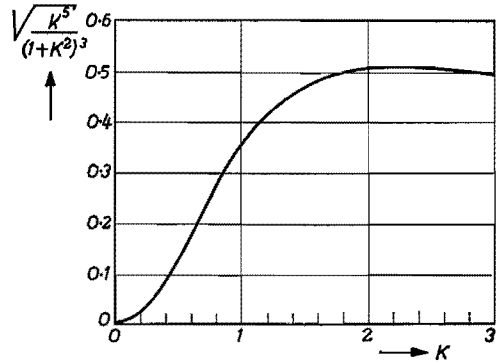


Fig. 3.2. Signal-to-noise ratio in the output of the modified matched filter, as a function of the filter width (cf. eq. (3.12)).

3.2.2.3. Threshold level and detection limit

From the filter output Z_k^* it must be decided whether a peak is present. With a peak in the signal, the second derivative becomes negative, so that a negative threshold level must be exceeded. Without assuming any particular distribution for the filtered noise, the Bienaymé–Chebychev inequality states that for a signal of purely random noise with variance σ_y^{*2},

$$p\,[|Z_p^*| > a\,\sigma_y^*] < \frac{1}{a^2}. \tag{3.13}$$

Hence, the probability of detecting a spurious peak due to noise using a threshold $5\sigma_y^*$ is less than 0·04. In fact this probability will be considerably smaller because the noise is almost normally distributed. For the threshold $5\sigma_y^*$ and $K = 1{\cdot}4$, in order to be detectable, a peak must have a minimum signal-to-noise ratio before filtering:

$$\mathrm{S/N_{min}} = \frac{5}{\sqrt{f}}. \tag{3.14}$$

The minimum peak area required for detection, which is directly proportional to the minimum detectable amount, is

$$A_{\mathrm{min}} = 12\,\sigma_y\,(w_p\,\Delta)^{1/2}. \tag{3.15}$$

By comparison, the commonly accepted limit with no filtering is S/N > 5, yielding $A_{\mathrm{min}} = 12\,\sigma_y\,w_p$. The optimum limit if the differentiation of the signal were not required is

$$S/N > \frac{3}{\sqrt{f}} \quad \text{or} \quad A_{min} = 8\,\sigma_y\,(w_p\,\Delta)^{1/2}.$$

These limits are illustrated in fig. 4.1.

3.2.2.4. Fusing limits

The fusing limit is defined as the minimum resolution between two peaks at which the two can be detected separately. Resolution is expressed as the distance in location over the mean peak width:

$$Rs = \frac{|\mu_2 - \mu_1|}{w_p}. \tag{3.16}$$

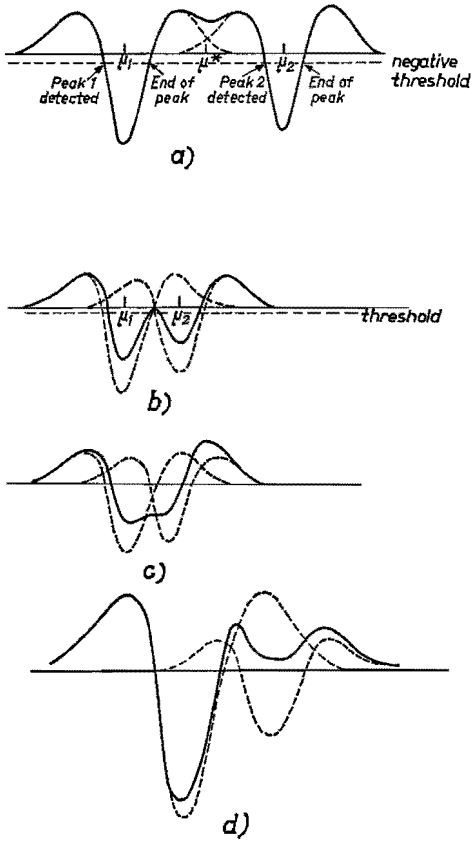


Fig. 3.3. Second derivatives of two overlapping Gaussian peaks (dashed curves) and of the composite curve (drawn curve).
(*a*) Two almost separated peaks. Detection by minima in the second derivative yields a spurious peak at μ^*. (*b*) Fusing limit for detection by multiple pairs of inflection points (zeros in the second derivative). (*c*) Fusing limit for the detection by minima in the second derivative: upon closer spacing the minima will coincide. (*d*) Two overlapping peaks giving one positive minimum in the second derivative of the composite curve.

The output of the matched filter can be seen as the smoothed second derivative of the signal. For two overlapping Gaussian peaks the filter output is the sum of the, broadened, second derivatives of the peaks, cf. fig. 3.3*a*. If peaks are detected where the filter output exceeds a certain negative threshold, the limit of separability is attained when the end of the first peak and the start of the second peak coincide, cf. fig. 3.3*b*. If the threshold is close to zero, i.e. large S/N, this limiting case is described by two conditions, as derived by Westerberg [3–3]:

$$\frac{\partial^2 y(t)}{\partial t^2} = \frac{\partial^3 y(t)}{\partial t^3} = 0. \tag{3.17}$$

This fusing limit depends on the ratio of the areas of the two peaks. In fig. 4.2, curve b, the limit for two Gaussian peaks of equal width is plotted.

Clearly, other methods of detection have different fusing limits. Often the existence of two tops in the signal is taken as the evidence of a composite peak. The fusing limit for this criterion, evaluated from the conditions [3–3])

$$\frac{\partial y(t)}{\partial t} = \frac{\partial^2 y(t)}{\partial t^2} = 0,$$

is plotted in fig. 4.2, curve a. A more sensitive detection method counts the minima in the second derivative [3–4]), or, equivalently, the zeros in the third derivative [3–5]). If the two minima coincide, cf. fig. 3.3*c*, the conditions apply

$$\frac{\partial^3 y(t)}{\partial t^3} = \frac{\partial^4 y(t)}{\partial t^4} = 0.$$

We solved these conditions numerically; the limit is plotted in fig. 4.2, curve c. A complication is that not all minima correspond to real peaks, e.g. the minimum marked μ^* in fig. 3.3*a*. Morrey [3–5]) proposed as an extra condition that both minima must be on the negative side. Taking this into account, the fusing limit follows the left branch of curve c (until area ratio 2·3) and then the right branche of b. Figure 3.3*d* illustrates a case of a discarded minimum from a real peak. We propose another condition, by which most of the positive minima from real peaks are retained:

(i) a single positive minimum between two negative minima is discarded;
(ii) if two positive minima are found between two negative minima, the smaller minimum is discarded.

These fusing limits were evaluated for Gaussian peaks with no noise. It is known, however, that the second derivative is very prone to noise. Smoothing is required to prevent noise from causing accidental minima, so that a single peak would be erroneously assigned as composite. The smoothing by the

matched filter depends on the filter width, becoming optimum with $K = \sqrt{5}$. A side effect is the broadening of peaks by a factor $(1 + K^2)^{1/2}$, which results in a loss of resolution in the filtered signal:

$$\mathrm{Rs}^* = \frac{\mathrm{Rs}}{(1 + K^2)^{1/2}}. \tag{3.18}$$

This expression shows that the loss is small if $K < 1$. In the algorithm described in the following section, the value $K = 0{\cdot}5$ is used, which gives the fusing limit (fig. 4.2, drawn curve) shifted by a factor 1·1 with respect to the noise-free limit.

3.2.2.5. Implementation

In the detection procedure the mean noise amplitude estimated in sec. 3.1 and peak-width regression estimated in sec. 3.1.2 are used. As the width of the matched filter should be taken proportional to the peak width, we had the choice between two alternatives: either, to adapt the filter weights continuously, or, to put the data on a logarithmic scale on which all peaks have equal width. The amounts of computation are roughly equal. However, because the linear relationship is only approximately valid, it is not necessary to have a strict proportionality. Updating the filter weights intermittently, after 10% changes of the width, gave no appreciable loss in detection, but reduced the amount of computation drastically. The filter weights are calculated from eqs (3.8b) and (3.7). A small correction of the calculated weights is usually necessary to ensure that the conditions for the proper differentiation of a second-degree polynomial are satisfied, i.e.

$$\sum_{i=-m}^{m} G_i = \sum_{i=-m}^{m} i\,G_i = 0; \qquad \sum_{i=-m}^{m} i^2\,G_i = 2. \tag{3.19}$$

The smallest set of weights which satisfies these conditions is (+1, −2, +1). These weights are approximately obtained if $w_f = 0{\cdot}6$, which is therefore a minimum filter width.

To combine a sensitive detection of trace peaks and a sensitive detection of overlapping peaks, a two-step filtering was designed:

(i) Using a matched filter with $K = 1{\cdot}5$, the start of a peak, t_s, is detected where the filter output first exceeds the threshold $-5\sigma_y{}^*$: the end t_e where the output returns within the threshold. The width of the detected peak is provisionally estimated as

$$w_p = \tfrac{1}{2}\,(t_e - t_s)/(1 + K^2)^{1/2};$$

the location is estimated from the minimum between t_s and t_e; the area

is estimated from the amplitude h at the minimum:

$$A = \tfrac{1}{8}\, h\, (t_s - t_e)^3\, (2\pi)^{1/2}.$$

(ii) The filtering is repeated over the regions of the large peaks (S/N > 20) with $K = 0{\cdot}5$; appropriate minima are taken as the locations of peaks.

The estimates of the location, width and area are needed as initial values in the iterative curve fitting, in case of overlapping peaks.

3.3. Estimation

The estimation includes three operations, viz. location of the peak boundaries, baseline correction and peak-parameter estimation.

3.3.1. *Location of peak boundaries*

The importance of accurate peak boundries was discussed in sec. 2.4.

The location is based on the assumption that the slope of the background in the neighbourhood of a peak is constant. The boundaries are defined as the point on the leading edge and the point on the trailing edge, at the smallest distance, where the signal slopes are equal. This definition implies that the boundaries are correct on a linear background.

The inaccuracy of the boundaries on a curved background could be reduced by assuming a parabolic shape, but a conflict between accuracy and precision results. We justify the linearity assumption by three arguments:

— the inaccuracy on the curved background will be small, because the background is a slowly varying function of time (compared with the peaks);
— if the background segments bracketing a peak are small compared with the base width of the peak, a curved baseline cannot be fitted because it is likely to yield large interpolation errors under the peak;
— if the bracketing segments are large compared with the base width, the effect of inaccurate boundaries on the fitted baseline is negligible.

The procedure for locating the boundaries is illustrated in fig. 3.4.

(i) starting from the peak top, the minimum and the maximum in the first derivative are located;
(ii) alternately at each side, the first derivative at the next point is calculated, using eq. (3.2), until the f.d. at the trailing edge becomes larger than the f.d. at the leading edge;
(iii) it is checked whether a different pair at a closer distance can be found.

If peaks overlap, the derived boundaries will be tangential to the coombs between the peaks. The boundaries of the peak group are determined by the above procedure, starting from the maximum in the f.d. on the first peak and the minimum in f.d. on the last peak.

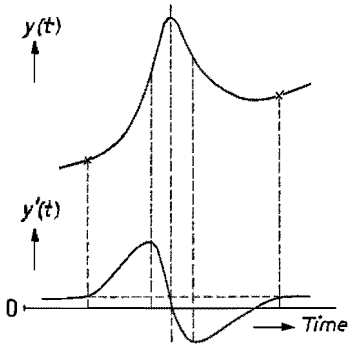


Fig. 3.4. Location of the peak boundaries as the first points on either side of the top where the slopes are equal.

3.3.2. *Baseline correction*

The peak regions are staked out by the determined boundaries. The remaining background segments are used to interpolate the baseline correction. To each peak or group of overlapping peaks a local baseline is fitted.

The least-squares fitting of a polynomial is done iteratively, starting with a zeroth-degree function and increasing the degree until a satisfactory fit is obtained. In this way the baseline function has the lowest degree required to follow the trend in the segments. Higher-degree terms add progressively to the uncertainty in the correction interpolated under the peak.

It is very attractive to use a set of orthogonal polynomials [3–6]. Let $p_k(t)$ be a polynomial of degree k in the independent variable t; two members of the set of orthogonal polynomials over the data (t_i, Y_i), $i = 1, \ldots, m$, satisfy the condition

$$\sum_{i=1}^{m} p_k(t_i)\, p_l(t_i) = 0 \qquad \text{if} \qquad k \neq l. \tag{3.20}$$

The function

$$f_n(t) = a_0 p_0(t) + a_1 p_1(t) + \ldots + a_n p_n(t)$$

is then a polynomial of degree n. Least-squares fitting of this function to the data leads to the set of $n + 1$ normal equations:

$$\sum_{j=0}^{n} a_j \sum_{i=1}^{m} p_j(t_i)\, p_k(t_i) = \sum_{i=1}^{m} Y_i\, p_k(t_i), \qquad k = 0, \ldots, n. \tag{3.21}$$

Using the property (3.17), these equations reduce to separate equations:

$$a_k = \frac{\sum_{i=1}^{m} Y_i \, p_k(t_i)}{\sum_{i=1}^{m} p_k^2(t_i)} . \tag{3.22}$$

The generation of the orthogonal polynomials is described by Forsythe [3–6].

An obvious advantage of orthogonal polynomials is that no involved matrix equations need to be solved. For our purpose it is very convenient that the coefficient of the current highest-degree term can be added or withdrawn without changing the coefficients of the lower-degree terms. This facilities the iterative raising of the degree of the polynomial.

The improvement of the fit can be studied from the sum of squared residuals. Let S_n denote the sum after fitting the nth-degree function:

$$S_n = \sum_{i=1}^{m} [Y_i - f_n(t_i)]^2. \tag{3.23}$$

Substitution of (3.20) and (3.22) yields:

$$S_n = \sum_{i=1}^{m} Y_i^2 - \sum_{k=0}^{n} a_k \sum_{i=1}^{m} p_k^2(t_i).$$

We assume the noise to be uncorrelated, so that the F-test for an additional term can be used. The ratio

$$F_n = (m - n - 1)\,(S_{n-1} - S_n)/S_n \tag{3.24}$$

follows an F distribution with 1 and $m - n - 1$ degree(s) of freedom. The ratio is a measure of how much the additional term has improved the fit. The significance of the calculated ratio can be seen from tables or from approximating functions [3–7].

3.3.3. *Peak-parameter estimation*

Single peaks are most conveniently characterized by its moments. Traditionally, the coordinate of the maximum is taken as the retention time. Overlapping peaks must be dissected by curve fitting.

3.3.3.1. Peak-top location

The coordinate of the largest sample provides an initial estimate of the top. The associated random error is about one quarter of the sampling interval for a noise-free signal. In the presence of noise the error may be several intervals.

Attempts to improve upon this estimate are based on a reconstruction of the top from the sampled data. Wijtvliet [3-8]) and Goedert and Guiochon [3-9]) estimated the location from the maximum of a parabola fitted to samples around the top. Two errors are associated with this estimate:
— a random error from noise on the samples;
— a systematic error due to the incorrectness of the parabolic model.

Assuming uncorrelated noise, the random error can be evaluated by applying the error-propagation expression (2.7). Let the parabola

$$f(p+i) = a\,i^2 + b\,i + c$$

be fitted to $2m+1$ samples $Y_{p-m}, \ldots, Y_{p+m}$; Y_p is the centre of the fit. The maximum of the parabola is at

$$\text{top} = \left(p - \frac{b}{2a}\right)\Delta.$$

The standard deviation is derived by variance analysis on the coefficients:

$$\sigma_{\text{top}} = \frac{\Delta\sigma_y}{|2a|}\left[\frac{3}{2m^3+3m^2+m} + \left(\frac{b}{2a}\right)^2 \frac{180}{8m^5+20m^4+10m^3-5m^2-3m}\right]^{1/2}. \tag{3.25}$$

If the fit is properly positioned, $|b/2a| < \frac{1}{2}$; the second term between square brackets becomes negligible for $m > 2$. The coefficient a can be estimated as the curvature at the top of a Gaussian peak with area A and width w_p:

$$2a = \frac{-A}{w_p^3\,(2\pi)^{1/2}}\,\Delta^2.$$

Introducing the signal-to-noise ratio

$$\text{S/N} = \frac{A}{\sigma_y\, w_p\,(2\pi)^{1/2}}$$

and the sampling density $f = w_p/\Delta$, eq. (3.25) simplifies to

$$\frac{\sigma_{\text{top}}}{w_p} = \frac{f}{\text{S/N}}\left(\frac{3}{2m^3+3m^2+m}\right)^{1/2}. \tag{3.26}$$

The random error decreases as the value of m increases.

The systematic error, on the contrary, increases as m is increased, because only the top part of a peak has approximately a parabolic shape. Wijtvliet [3–8]) showed that the bias between the maximum of a Gaussian peak and the maximum of the fitted parabola depends on the value of m, the width of the peak and the off-centredness, i.e. the distance between the centre of the fit and the maximum of the peak.

The optimum value of m that balances the two errors will be a complex function of the peak shape, the signal-to-noise ratio and the sampling density. This optimum can be approached however:

— The systematic erorr will be related to the closeness of the fit, measured by the standard deviation of the fit:

$$s = \left(\frac{1}{2m-2} \sum_{i=-m}^{m} [Y_{p+i} - f(p+i)]^2 \right)^{1/2}. \tag{3.27}$$

— If the region used for fitting is approximately parabolic, s provides an accurate estimate for the random-noise amplitude, independent of m. The systematic error will also be small and independent of m.

— Increasing m so far that the signal shape becomes non-parabolic, the value of s will increase due to lack of fit. The systematic error will also increase.

The effect of substituting σ_y by s in eq. (3.25) is that the first term, $s/2a$, reflecting the systematic error, will first be constant as m is increased and increases if non-parabolic parts of the peak are included. The competition with the decreasing square-root term results in a minimum in the calculated value of σ_{top}. The value of m at the minimum is taken. Although this will not exactly be the optimum value, it satisfies two requirements:

— the derived value of m will increase with decreasing S/N;

— on asymmetrical peaks the lack of fit will result in a narrow fitting region, so that the bias is small.

The procedure for locating the peak top with iterative adjustment of the fitting region is:

(i) p is set equal to the index of the largest sample; starting value of m = maximum $(2, \frac{1}{2}f)$;

(ii) a parabola is fitted to $2m + 1$ samples $Y_{p-m}, \ldots, Y_{p+m}$; the maximum of the parabola is located: top $= (p - b/2a)\,\Delta$; σ_{top} is calculated from (3.25), using $\sigma_y = s$ (eq. (3.27));

(iii) if $|b/2a| > \frac{1}{2} + \sigma_{top}$, the fitting region is shifted: $p = p \pm 1$, in order to centre the fit properly, and (ii) is repeated;

(iv) if the standard deviation σ_{top} has decreased, m is increased by 1, and (ii) is repeated.

Wijtvliet [3–8]) elaborated a method for reducing the systematic error from off-centredness. Although reducing the error by an order of magnitude, typically

from 0·01 w_p to 0·001 w_p, it appears of limited analytical interest because real peaks are always composite so that the top is shifted [3–10]).

Figure 4.7 shows the resultant error after adjustment of the fitting region, for a Gaussian peak, as a function of the signal-to-noise ratio, at some sampling densities. The error was determined with simulated Gaussian peaks and various sequences of superposed white noise. The off-centredness was varied randomly between $-\frac{1}{2}$ and $+\frac{1}{2}$. The curves represent average values.

3.3.3.2. Moments calculation

The definition of the moments was given in sec. 2.1. The moments of a sampled peak are calculated by numerical integration. Let the peak boundaries be at $t = a\,\Delta$ and $t = b\,\Delta$; Y_i is the baseline-corrected signal. The simplest procedure is to perform a summation:

— zeroth moment, or peak area $A = \sum_{i=a}^{b} Y_i\,\Delta$; (3.28)

— first moment, or centre of gravity $\mu = \frac{1}{A} \sum_{i=a}^{b} i\Delta\, Y_i\,\Delta$; (3.29)

— nth central moment $m_n = \frac{1}{A} \sum_{i=a}^{b} (i\Delta - \mu)^n\, Y_i\,\Delta$. (3.30)

The precision of the numerical integration is increased by differently weighting samples near to the boundaries [3–11]). Generally, a curved integrand is more precisely integrated by a parabolic form, e.g.

$$A = \Delta\left[\tfrac{3}{8}(Y_a + Y_b) + \tfrac{7}{6}(Y_{a+1} + Y_{b-1}) + \tfrac{23}{24}(Y_{a+2} + Y_{b-2}) + \sum_{i=a+3}^{b-3} Y_i\right]. \quad (3.31)$$

The systematic error and the random error in the moments are dependent on the signal-to-noise ratio, the sampling density and the integration limits [3–12,13]). For a given signal only the limits can be adjusted to minimize the errors.

The random error in the area is derived by applying the error-propagation expression (2.7):

$$\sigma_A = \sigma_y\,\Delta\,(b - a)^{1/2}. \quad (3.32)$$

If the integration limits are taken at the distance $\pm\, k\, w_p$ from the top,

$$b - a = 2\,k\,w_p/\Delta.$$

Substituting the signal-to-noise ratio S/N and the sampling density f, (3.32) can be rewritten as

$$\frac{\sigma_A}{A} = \frac{1}{\mathrm{S/N}} \left(\frac{k}{f\pi} \right)^{1/2}. \tag{3.33}$$

This error is plotted in fig. 3.5 as a function of k, for various S/N and f.

The systematic error in the area, Δ_A, is equal to the area outside the integration limits. For a Gaussian peak:

$$\frac{\Delta_A}{A} = 1 - \mathrm{erf}\left(\frac{k}{\sqrt{2}} \right); \tag{3.34}$$

erf (x) denotes the well-known error function. The systematic error is also plotted in fig. 3.5, drawn curve.

As a function of k, the systematic error and the random error have opposing trends. From the derived expressions (3.33) and (3.34) the value of k that minimizes the mean squared error may be calculated. Generally, however, the systematic error is unknown as it depends on the peak shape. A practical condition for optimizing the integration limits is to increase k until the value of the integral changes less than standard deviation, i.e. until

$$\frac{\partial \sigma_A}{\partial k} > \left| \frac{\partial A}{\partial k} \right|. \tag{3.35}$$

Applying this condition, cf. eq. (2.8), to the area integration of a Gaussian peak, the error curves in fig. 4.4 (drawn curves) are derived:

$$\varepsilon_A = (\sigma_A{}^2 + \Delta_A{}^2)^{1/2}.$$

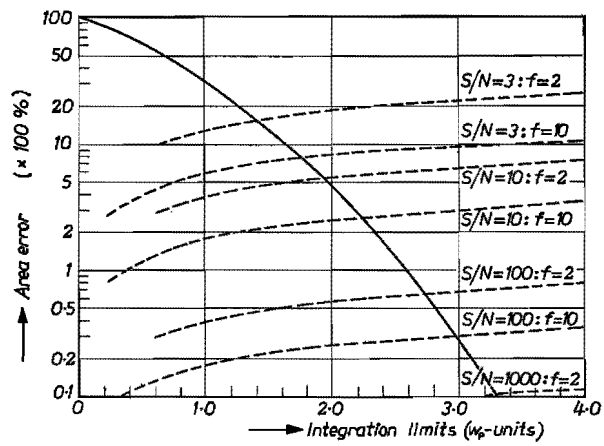


Fig. 3.5. Systematic error (drawn curve) and random errors (dashed curves) in the area of a Gaussian peak, as a function of the integration limits. The limits are taken symmetrically to the top. Signal-to-noise ratio S/N and sampling densities f as indicated.

A virtue of this procedure for iterative optimization of the integration limits is that it does not assume a certain peak shape; it only assumes an expression for the random error such as (3.32). The derived error curves are, approximately, also valid for non-Gaussian shapes. Only peaks with extended flat tails will have a larger systematic error contribution.

The random error in the centre of gravity is derived as

$$\sigma_\mu^2 = \sigma_y^2 \frac{\Delta^2}{A^2} \sum_{i=a}^{b} (i\,\Delta - \mu)^2 \approx \sigma_y^2 \frac{\Delta}{A^2} \tfrac{1}{3} [(b\,\Delta - \mu)^3 - (a\Delta - \mu)^3]. \quad (3.36)$$

Taking $(b\,\Delta - \mu) = (\mu - a\,\Delta) = k\,w_p$, and substituting S/N and f yields

$$\frac{\sigma_\mu}{w_p} = \frac{1}{\mathrm{S/N}} \left(\frac{k^3}{3 f \pi} \right)^{1/2}. \quad (3.37)$$

Studying the systematic error in the centre of gravity and the effect of the integration limits thereupon, it is, of course, not reasonable to assume symmetrical extension of the integration limits. Instead, one limit is fixed, $k = k_1$, and the other is varied:

$$\mu = \frac{1}{A} \int_{\mu - k_1 w_p}^{\mu + k w_p} t\, g(t)\, \mathrm{d}t \quad (3.38)$$

and

$$\frac{\partial \mu}{\partial k} = \frac{1}{A} (\mu + k\,w_p)\, g(\mu + k\,w_p). \quad (3.39)$$

The condition $\partial \sigma_\mu / \partial k = \left| \partial \mu / \partial k \right|$ was solved numerically (the "solution" $k = 0$ is of course omitted), for a Gaussian peak. The resulting error is plotted in fig. 4.5 (drawn curves). For Gaussian peaks the systematic error is zero. For non-Gaussian shapes the relationships may be assumed to be qualitatively valid.

The procedure for calculation of the moments is

(i) the peak area and the centre of gravity are provisionally calculated by numerical integration between the peak boundaries;

(ii) starting from the provisional centre, the integration limits are iteratively extended at each side until the increment of the sum is less than the change in the standard deviation;

(iii) samples near the integration limits are weighted as in (3.31).

3.3.3.3. Curve fitting

In this section, expressions for the errors in the estimated parameters are derived and an outline of the algorithm is given. The theory of curve fitting

is well known, cf. ref. 3-11. The following account serves to have the appropriate expressions available.

Let the signal be sampled at n points (t_k, Y_k). The function to be fitted is $f(t, \mathbf{p})$; $\mathbf{p} = (p_1, \ldots, p_m)$ is the vector of m optimizable parameters. The least-squares criterion requires minimization of

$$S(\mathbf{p}) = \sum_{k=1}^{n} [f(t_k, \mathbf{p}) - Y_k]^2. \tag{3.40}$$

If the function is non-linear in the parameters, the optimum parameter vector $\hat{\mathbf{p}}$, i.e. the vector that minimizes S, must be approached iteratively, starting from initial estimates $\mathbf{p}^0$. In the hth iteration a correction vector $\Delta\mathbf{p}^h$ is calculated:

$$\mathbf{p}^h = \mathbf{p}^{h-1} + \Delta\mathbf{p}^h. \tag{3.41}$$

Let ${}^hf_k$ represent $f(t_k, \mathbf{p}^h)$. The corrections are found by solving the set of normal equations

$$Z\Delta\mathbf{p}^h = \mathbf{b}, \tag{3.42a}$$

where

$$Z = \left[z_{ij} = \sum_{k=1}^{n} \frac{\partial {}^hf_k}{\partial p_i} \frac{\partial {}^hf_k}{\partial p_j} \right]; \quad \mathbf{b} = \left[b_i = \sum_{k=1}^{n} \frac{\partial {}^hf_k}{\partial p_i} (Y_k - {}^hf_k) \right];$$

$$i, j = 1, \ldots, m.$$

The solution can be found by matrix inversion:

$${}^h\Delta\mathbf{p} = Z^{-1}\,\mathbf{b}. \tag{3.42b}$$

If the noise over the sampled data is uncorrelated, having mean amplitude σ_y, the error-propagation expression (2.7) can be invoked to determine the standard deviation of the parameter corrections, and hence the standard deviations of the estimated parameters (σ_i denotes the standard deviation of p_i):

$$\sigma_i = \sigma_y \sqrt{z_{ii}^{-1}}. \tag{3.43}$$

The standard deviation of the parameter estimates is determined by the noise amplitude and by the diagonal elements in the inverse of the matrix of the normal equations. The diagonal elements become available in most algorithms for solving matrix equations. The noise amplitude can also be estimated from the sum of squares, taking the lost degrees of freedom into account:

$$\sigma_y \approx \left(\frac{1}{n-m} S(\hat{\mathbf{p}}) \right)^{1/2}. \tag{3.44}$$

Expression (3.43) is a well established part of the theory of parameter estimation. Yet, surprisingly, in chromatography no one has made use of it for estimating the errors.

3.3.3.3.1. *Errors for a single Gaussian peak*

Expression (3.43) is quite general, assuming only uncorrelated noise. In order to obtain more-tangible expressions for the errors, the fitting of a single Gaussian peak is considered, i.e. $f(t, \mathbf{p}) = g(t, A, \mu, w)$. Let the peak be sampled at equally spaced intervals Δ.

First, it is assumed that the position and width are known, e.g. from previous analyses. This yields a 1×1 matrix equation; the single element in the matrix can be approximated analytically:

$$z_{11} = \sum_{k=1}^{n} \left(\frac{\partial^h f_k}{\partial A} \right)^2 \approx \frac{1}{\Delta} \int_{-\infty}^{\infty} \left(\frac{\partial f(t, \mathbf{p})}{\partial A} \right)^2 \mathrm{d}t = \frac{1}{\Delta} \frac{1}{2w\sqrt{\pi}}. \qquad (3.45)$$

Hence,

$$z_{11}^{-1} = 2w\Delta\sqrt{\pi},$$

yielding

$$\sigma_A = \sigma_y (2w\Delta\sqrt{\pi})^{1/2}.$$

Substituting the signal-to-noise ratio S/N and the sampling density, this can be rewritten to

$$\frac{\sigma_A}{A} = \frac{1}{\mathrm{S/N}} \left(\frac{0{\cdot}56}{f} \right)^{1/2}. \qquad (3.46)$$

This expression may be compared with the expression for the random error in the zeroth moment (3.33). It may appear from (3.46) that for a given S/N the error can be made arbitrarily small by increasing the sampling density. However, actual physical signals are always bandwidth-limited so that beyond a certain sampling density successive noise contributions will not be uncorrelated and the error does not decrease further. From the correlation width of the noise, cf. sec. 2.2, this limiting density can be assessed.

Generally, the location of the peak and the width are also unknown. By approximating the matrix elements in a similar way as in (3.45), the matrix of the normal equations is obtained:

$$\begin{pmatrix} \dfrac{1}{2w\,\Delta\sqrt{\pi}} & 0 & \dfrac{-A}{4w^2\,\Delta\sqrt{\pi}} \\ 0 & \dfrac{A^2}{4w^3\,\Delta\sqrt{\pi}} & 0 \\ \dfrac{-A}{4w^2\,\Delta\sqrt{\pi}} & 0 & \dfrac{3\,A^2}{8w^3\,\Delta\sqrt{\pi}} \end{pmatrix}; \tag{3.47a}$$

inverse:

$$\begin{pmatrix} 3w\,\Delta\sqrt{\pi} & 0 & \dfrac{2w^2\,\Delta\sqrt{\pi}}{A} \\ 0 & \dfrac{4w^3\,\Delta\sqrt{\pi}}{A^2} & 0 \\ \dfrac{2w^2\,\Delta\sqrt{\pi}}{A} & 0 & \dfrac{4w^3\,\Delta\sqrt{\pi}}{A^2} \end{pmatrix}. \tag{3.47b}$$

This yields the expressions for the relative errors

$$\frac{\sigma_A}{A} = \frac{1}{\mathrm{S/N}}\left(\frac{0{\cdot}85}{f}\right)^{1/2}; \qquad \frac{\sigma_\mu}{w} = \frac{\sigma_w}{w} = \frac{1}{\mathrm{S/N}}\left(\frac{1{\cdot}13}{f}\right)^{1/2}. \tag{3.48}$$

These errors are plotted in figs 4.4 and 4.5 (dashed curves). Comparing these with the error for area-only fitting, eq. (3.46), it appears that the sensitivity to noise is enhanced. The cause is a covariance between the area and the width. The degree of interdependence is measured by the normalised off-diagonal term

$$\varrho_{ij} = \frac{z_{ij}^{-1}}{(z_{ii}^{-1}\, z_{ij}^{-1})^{1/2}}. \tag{3.49}$$

The magnitude of the normalized off-diagonal terms ranges between —1 and +1. If two parameters are strongly correlated $|\varrho_{ij}| \approx 1$, and neither parameter can be pointed down precisely. In such case knowledge of the value of one parameter considerably improves the precision of the other. As shown, knowledge of the peak width decreases the standard deviation of the area by a factor 0·81.

3.3.3.3.2. *Errors for overlapping Gaussian peaks*

For two overlapping peaks of equal width w, at positions μ_1 and μ_2 and areas A_1 and A_2, the standard deviations become a function of the resolution

$$\mathrm{Rs} = \frac{|\mu_1 - \mu_2|}{w}.$$

First, it is assumed that the positions and widths are known in advance. This is a realistic assumption in many routine analyses where the location and shape of the peaks are known and only the concentrations vary. The diagonal terms are as for single peaks:

$$z_{11} = z_{22} \approx \frac{1}{2w\Delta\sqrt{\pi}}.$$

The off-diagonal terms can be approximated:

$$z_{12} = z_{21} = \frac{1}{\Delta}\sum_{k=1}^{n}\frac{\partial f_k}{\partial A_1}\frac{\partial f_k}{\partial A_2}\Delta \approx \frac{1}{\Delta}\int_{-\infty}^{\infty}\frac{\partial f(t)}{\partial A_1}\frac{\partial f(t)}{\partial A_2}\mathrm{d}t = \frac{1}{2w\Delta\sqrt{\pi}}\exp(-\tfrac{1}{2}\mathrm{Rs}^2). \tag{3.50}$$

The effect of the resolution on the standard deviation is best expressed relative to the standard deviation at "infinite" resolution, i.e. two single peaks:

$$\alpha_A(\mathrm{Rs}) = \frac{\sigma_A(\mathrm{Rs})}{\sigma_A(\infty)} = \sqrt{\frac{z_{11}^{-1}(\mathrm{Rs})}{z_{11}^{-1}(\infty)}} = \sqrt{\frac{1}{1-\exp(-\frac{1}{2}\mathrm{Rs}^2)}}. \tag{3.51}$$

This "loss factor" gives the multiplication factor for the error in the area of single peaks.

Generally, both the locations, areas and widths of two overlapping peaks will be unknown. Simultaneous fitting of all parameters gives a 6×6 matrix. Although it might be possible, as before, to obtain analytical expressions for the matrix elements, it is virtually impossible to invert the involved 6×6 matrix in closed form. Therefore, for Rs increasing from 0·25 to 10 in steps of 0·25, the matrix elements were calculated numerically and the matrices were inverted. The calculated loss factors for area, location and width are plotted in fig. 4.7. These curves substantiate two important conclusions:

— The precision of parameter estimates of overlapping peaks decreases rapidly with decreasing resolution, e.g. if S/N = 20 and $f = 2$, the relative error in the area of a single peak is 3%; for a doublet with Rs = 1·5 the error increases to 20% (loss factor = 6). Hence, little significant information can be obtained from the composite curve of closely spaced peaks.

— The precision also decreases with increasing number of optimizable parameters, e.g. if Rs = 1 the area-only fitting is attributed with a loss factor 1·6 while the simultaneous fitting of all parameters yields a loss factor 20. Loosely speaking, the larger the number of unknowns over which the information contained in the experimental data must be distributed, the less significant the derived values.

These conclusions are useful for assessing the errors for non-Gaussian models in curve fitting. These models contain additional parameters to account for

asymmetry, excess, etc. The preceding results indicate that for single peaks the errors increase with the number of interdependent parameters. The skew and excess in Edgeworth series are nearly independent. The time constant on the exponentially convoluted Gaussian is strongly correlated with the location parameter, the correlation coefficient (3.49) becoming nearly 1 for small values of the time constant. For overlapping peaks, even independent parameters in a peak model will become correlated due to mutual interference of the peaks. As a result the errors will sharply rise when using more-complex peak models. For dissecting overlapping peaks it is therefore important to have an accurate peak model which has as few degrees of freedom as possible.

3.3.3.3.3. *Implementation*

Peak model: the 4 peak models compiled in table 2-I are available in the curve-fitting procedure. The preceding error discussion showed that the number of optimizable parameters should be restricted. Therefore, a Gaussian model is assumed initially. After convergence of the fit of Gaussian peaks, the residues are examined to see whether the introduction of additional peaks or an extension of the peak model can improve the fit significantly.

The test on the introduction of an additional peak is identical to matched-filter detection on the residues. Differentiation is not necessary because the baseline is already eliminated. The threshold is derived from the mean amplitude of the residues, eq. (3.44).

The test on extension of the peak model is based on the assumption that the extended peak model $f(t, \mathbf{p}, p_{m+1})$ is approximately linear in the additional parameter p_{m+1}. Let $\hat{\mathbf{p}}$ denote the optimum vector. Hence,

$$f_k(\hat{\mathbf{p}}, p_{m+1}) = f_k(\hat{\mathbf{p}}) + p_{m+1} \frac{\partial f_k}{\partial p_{m+1}}. \tag{3.52}$$

Let the residues after convergence be denoted as

$$E_k = Y_k - f_k(\hat{\mathbf{p}}) \tag{3.53}$$

Minimization for p_{m+1} of

$$\sum_{k=1}^{n} [f_k(\hat{\mathbf{p}}, p_{m+1}) - Y_k]^2 = \sum_{k=1}^{n} \left(p_{m+1} \frac{\partial f_k}{\partial p_{m+1}} - E_k \right)^2$$

yields

$$p_{m+1} = \frac{\sum_{k=1}^{n} E_k \, \partial f_k / \partial p_{m+1}}{\sum_{k=1}^{n} (\partial f_k / \partial p_{m+1})^2}. \tag{3.54}$$

The corresponding reduction of the sum of squares is

$$S(\hat{\mathbf{p}}) - S(\hat{\mathbf{p}}, p_{m+1}) = p_{m+1} \sum_{k=1}^{n} E_k \frac{\partial f_k}{\partial p_{m+1}}. \tag{3.55}$$

The significance of the reduction is measured by the F-test for an additional term (cf. (3.24))

$$F = (n - m - 1) \frac{S(\hat{\mathbf{p}}) - S(\hat{\mathbf{p}}, p_{m+1})}{S(\hat{\mathbf{p}}, p_{m+1})}. \tag{3.56}$$

The calculated F-values for each of the three extended models are compared with tabulated values. If the values are significant, the fitting function is extended using the model with the highest F-value. The curve-fitting procedure is then repeated until the convergence test is satisfied. The test on the introduction of a new peak is similar to the extension of the model, if p_{m+1} is considered as the area of the additional peak.

Initial estimates: Initial estimates are derived in the detection procedure, as described in sec. 3.2.2.5. The contiguity of the initial estimates is very important for fast convergence to the global minimum. Tests proved that the derived values are sufficiently close to give fast convergence. With simulated peaks no trapping into local minima was observed.

Algorithm: Marquardt's modification of Newton's algorithm [3-14]) was applied. This algorithm is safeguarded against divergence. If the calculated corrections lead to divergence, the diagonal elements in the matrix of the normal equations are increased, which brings about a shift of the corrections in the sense of the gradient. A drawback is that the matrix equations must be solved once again. However, the algorithm usually converged at once from the derived initial estimates. Choleski's algorithm [3-15]) is used to invert the matrix equations.

Constraints: To keep the parameter values within feasible ranges the updated values were tested after each iteration whether

— areas, widths and time constants are positive;
— the peak location is within the fitting region;
— the width of an individual peak does not deviate more than 20% from the mean peak width.

Although often trivial, these constraints are essential for fast convergence and

for avoiding unrealistic "solutions". To account for the inequality constraints, we developed the following iterative scheme:

(i) the unconstrained correction vector is calculated $\Delta\mathbf{p}^0 = Z^{-1} \cdot \mathbf{b}$;

(ii) suppose that k calculated corrections $p_1, \ldots, p_k$ exceed the maximum allowable corrections $\max_1, \ldots, \max_k$ respectively.

(iii) Lagrange's method of undertermined multipliers yields k additional equations: $\lambda_i (\Delta p_i - \max_i) = 0$. Solving these together with the normal equations leads to the modified equation:

$$\left(\begin{array}{c|c} Z & E \\ \hline E^T & 0 \end{array}\right)\begin{pmatrix} \Delta\mathbf{p}^1 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{max} \end{pmatrix}. \tag{3.57}$$

$E = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k)$ is an $m \times k$ matrix of the unit vectors of the constrained parameters, 0 is a $k \times k$ matrix of zeros and $\Delta\mathbf{p}^1$ is the constrained solution. It is postulated that the inverse matrix is of the form

$$\left(\begin{array}{c|c} Z^{-1} + F & G \\ \hline G^T & H \end{array}\right). \tag{3.58}$$

As the product of the two matrices is the identity matrix,

$$\begin{aligned} F &= Z^{-1} E H E^T Z^{-1}, \\ G &= -H E^T Z^{-1}, \\ H &= -(E^T Z^{-1} E)^{-1}. \end{aligned} \tag{3.59}$$

Substitution yields

$$\begin{aligned} \Delta\mathbf{p}^1 &= \Delta\mathbf{p}^0 - Z^{-1} E H (\mathbf{max} - E^T \Delta\mathbf{p}^0), \\ \boldsymbol{\lambda} &= H (\mathbf{max} - E^T \Delta\mathbf{p}^0). \end{aligned} \tag{3.60}$$

This is very convenient because only the matrix H must be calculated, which is the inverse of a $k \times k$ submatrix of Z^{-1}. Usually the number of constraints k is much lower than the number of parameters m, so that the calculation of H requires considerably less time than the solution of (3.57) by direct inversion.

(iv) It is checked whether in the constrained solution other parameters exceed their maximum allowable correction. If so, then (iii) is repeated including constraints for the parameters in question.

Convergence test: The curve-fitting algorithm has converged if each of the following criteria is satisfied:

— The sum of squares between two iterations changes less than 5%.

— The damping factor in Marquardt's algorithm is less than 2.

— For each parameter:

$$|\Delta p_i| < \varepsilon_a + \varepsilon_r |p_i| \qquad \text{or} \qquad |\Delta p_i| < \tfrac{1}{4}\sqrt{\frac{z_{ii}^{-1}\, S(\mathbf{p})}{n-m}}\,;$$

ε_a and ε_r are respectively an absolute tolerance and a relative tolerance. For the area only a relative tolerance was used, i.e. $\varepsilon_a = 0$, $\varepsilon_r = 0{\cdot}005$, and for the location and width an absolute tolerance was set, $\varepsilon_a = 0{\cdot}05$ (sample interval). The term under the radical sign is an estimate for the variance of the parameter p_i, cf. (3.43) and (3.44).

REFERENCES

3-1) N. R. Draper and H. Smith, Applied regression analysis, Wiley, New York, 1966, p. 21.
3-2) G. Brouwer and J. A. J. Jansen, Anal. Chem. **45**, 2239, 1973.
3-3) A. W. Westerberg, Anal. Chem. **41**, 1770, 1969.
3-4) A. B. Littlewood, T. C. Gibb and A. H. Anderson, in C. L. A. Harbourn (ed.), Gas chromatography 1968, Institute of Petroleum, London, 1969, p. 297.
3-5) J. R. Morrey, Anal. Chem. **40**, 905, 1968.
3-6) G. E. Forsythe, J. Soc. ind. appl. Math. **5**, 74, 1957.
3-7) M. Abramowitz and I. A. Segun(eds), Handbook of mathematical functions, Dover, New York, 1968, p. 946.
3-8) J. J. M. Wijtvliet, Thesis Technological University Eindhoven, 1972, Ch. 3.
3-9) M. Goedert and G. Guiochon, Chromatographia **6**, 39, 1973.
3-10) J. A. Rijks, Thesis Technological University Eindhoven, 1973, p. 47.
3-11) P. R. Bevington, Data reduction and error analysis for the physical sciences, McGraw-Hill, New York, 1969, pp. 200 and 270.
3-12) S. N. Chesler and S. P. Cram, Anal. Chem. **43**, 1922, 1971.
3-13) M. Goedert and G. Guiochon, J. chromatog. Sci. **11**, 326, 1973.
3-14) D. W. Marquardt, J. Soc. ind. appl. Math. **11**, 431, 1963.
3-15) R. Zurmühl, Matrizen, Springer, Berlin, 1964, p. 66.

4. RESULTS AND APPLICATIONS

The purpose of this chapter is threefold:

— To summarize the performance specifications of the program, especially to the potential users who may not be acquainted with or interested in the techniques discussed in chapter 3.

— To discuss the application of the program to one chromatogram selected as representative for a wide range of analyses.

— To discuss the tailoring of the program to processing with different priorities such as speed and low cost.

4.1. Performance specifications

The program was developed with three major aims:

— automatic processing,

— low detection limits,

— optimum accuracy and precision of the results.

We discuss below to what extent these aims have been attained. In order to make the discussion self-contained and comprehensible for non-experts in data processing, the performance data will be presented without reference to the underlying mechanisms developed in chapter 3.

We will try to compare our results with those reported in literature but it should be stated right away that such comparisons are quite limited. Owing to the lack of standard tests and criteria, most of the reported results stem from ad hoc tests. Extrapolation to our standards often needs some speculation. Moreover, some programs are constructed with quite different aims in view so that the results are hard to compare by any standards.

4.1.1. *Automatic processing*

In addition to the chromatogram, five data must be specified as input to the computer, viz. (1) the number of chromatograms for processing, (2) an upper bound for the number of samples in one chromatogram, (3) the time of the first sample relative to the injection time, (4) the sampling interval, (5) the format of the plot output. Specification of these data giving pertinent information about the data acquisition or the output format, requires no understanding of the program.

By comparison, Wijtvliet's program [4–1]) requires a similar number of data for standard processing, but to obtain the best results ten odd parameters must be preset. Littlewood's program [4–2,3,4]) requires the specification of plate number, threshold value for spurious peak filtering and filter widths.

4.1.2. *Detection limits*

The detection of peaks is limited for two reasons:

— Small peaks may get lost in random noise or submerged in a mass of small neighbouring peaks. This will be called the "detectability limit".
— Two peaks may come so close that the composite peak cannot be distinguished from a single peak, by the applied detection method. This will be referred to as the "fusing limit".

The detectability limit results from a balance between maximizing the detection of real peaks and minimizing that of spurious peaks such as noise spikes or baseline bumps. The limit is a function of the signal-to-noise ratio (S/N), which is the ratio of the peak height over the mean noise amplitude, and the peak width w (expressed in sample intervals): $S/N > 5/\sqrt{w}$. Since, for a fixed height, the area is proportional to the width, the minimum area required for detection, and hence the minimum detectable amount, increases proportionally to the square root of the width, as illustrated in fig. 4.1, drawn curve. This limit, attained by the program, is only slightly higher than the optimum limit (lower dashed curve). Littlewood's program [4–2]) detects peaks when the slope of the signal is greater than a certain threshold. Because the knowledge of the peak width is not exploited, the minimum area will increase linearly with the width (upper dashed curve). Wijtvliet's method of detection uses a threshold for the signal obtained in the course of the "average-below-average" baseline approximation. As the threshold in this way depends on the peak density and the peak heights, it is not possible to assess the detectability limit generally. In the most favourable case it will approach the upper dashed curve.

The fusing limit depends on the area ratio of the overlapping peaks. Figure

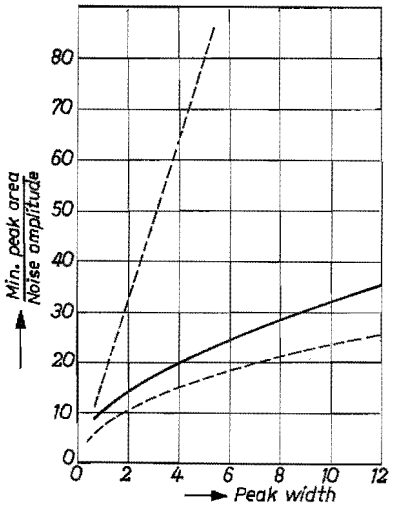


Fig. 4.1. Detection limits of Gaussian peaks in white noise. Drawn curve: limit attained by the program; lower dashed curve: limit for matched-filter detection; upper dashed curve: limit for slope detection with fixed filter width.

4.2, drawn curve, depicts the minimum resolution between two Gaussian peaks required for separate detection, in our program. Figure 4.3 illustrates some cases near the limit: it appears that the method is as discriminative as the expert's eye. The lower dashed curve in fig. 4.2 indicates the limit that could be attained, in principle, by our method if no noise were present. The discrepancy results from a loss in resolution due to filtering of the noise necessary to prevent the detection of spurious peaks. The limit of our method is almost the lowest meaningful limit, because actual pure peaks show nearly the same profile as the composites in fig. 4.3. It will also be discussed in the next section that the precision of parameter estimates of overlapping peaks decreases rapidly with resolution. By comparison, the upper dashed curve in fig. 4.2 indicates the minimum

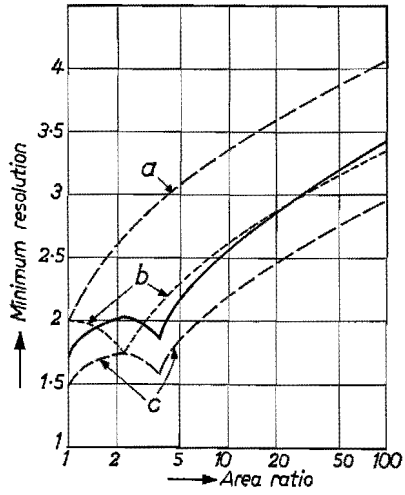


Fig. 4.2. Fusing limits for overlapping Gaussian peaks of equal width. Drawn curve: limit attained by the program; curve a: limit to two maxima on the composite peak (shoulder limit); curve b: limit to two pairs of inflection points; curve c: limit to two maxima in the second derivative.

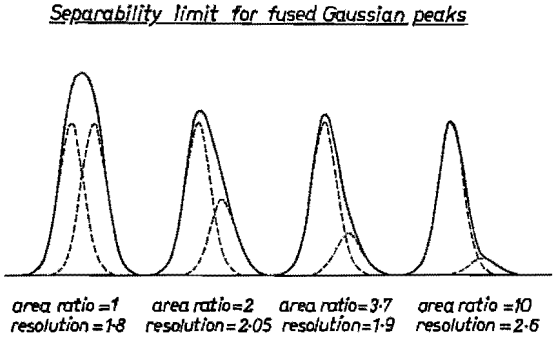


Fig. 4.3. Separability limit for fused peaks.

resolution for the existence of two tops on the composite peak. Curve b is the limit pertaining to Westerberg's method of counting pairs of inflection points [4–5]). The detection methods of Morrey [4–6]) and Brouwer and Jansen [4–7]) will attain the left branch of curve c below the area ratio 2·25, and curve b above this ratio. Littlewood's second method [4–2]) is able to attain the same limit as our method, but it detects additional spurious peaks between almost separated doublets.

4.1.3. *Accuracy and precision*

The systematic and random errors that will be discussed are associated with the *estimation* of peak parameters. Errors from preceding steps in the analysis, e.g. sampling, injection, separation, detection and recording, are ignored. However, no matter how clever the data processing, the inherent errors cannot be corrected or reduced generally. Rather, the estimation errors pile up on top of these errors. This should be reminded when interpreting the presented error curves. A low error figure usually means that the preceding steps will be the quality-determining factors.

The estimation errors can be attributed to two causes:

— random noise on the signal; the estimated parameter values are consequently also random quantities with associated standard deviations;
— inexactness or model errors in the calculation procedure.

4.1.3.1. Peak area

The area of a single peak is determined by numerical integration of the base-line-corrected signal. No particular peak shape is assumed. The result is also free from errors commonly associated with the determination of peak height, inflection points, halfwidth, etc. The precision depends on the signal-to-noise ratio, on the sampling density and on the integration limits. The former two are fixed for a given signal, but in our calculation procedure the limits are adjusted for a trade-off between accuracy and precision. The resultant error, i.e. the root of the sum of the squared random and systematic errors, for a Gaussian peak, is plotted in fig. 4.4 (drawn curves) as a function of the signal-to-noise ratio S/N, at various sampling densities f (expressed as the number of samples per peak width, or, equivalently, as the peak width expressed in sample intervals). The systematic part in the plotted error is one order of magnitude smaller than the random part. Because in the limits adjustment no peak shape is assumed, these curves may be considered as representative for all peaks, except those with a very extended tail. Similar curves were given previously by Chesler and Cram [4–8,9]) and by Goedert and Guiochon [4–10,11]). These authors did not mention a criterion or procedure for balancing accuracy and precision. The results in fig. 4.4 are comparable to those given in ref. 4-10, except that our results are not limited by a systematic error from fixed integration limits.

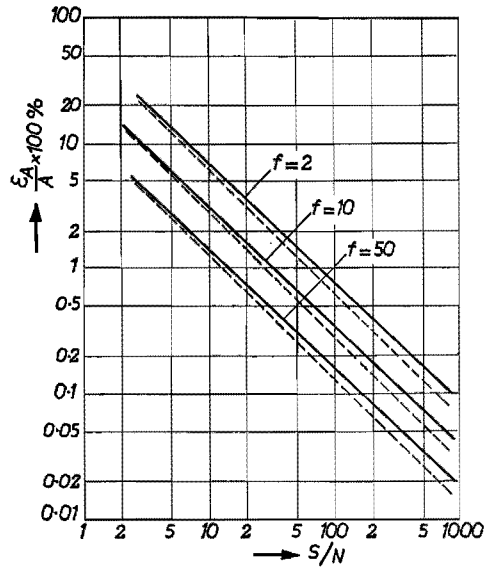


Fig. 4.4. Relative area error, as a function of the signal-to-noise ratio, at the sampling densities indicated, for a Gaussian peak. Drawn curves: resultant error with numerical integration; dashed curves: random error with curve fitting.

Although, in our program, curve fitting is not used for single peaks, the random errors with curve fitting are also plotted in fig. 4.4 (dashed curves). These errors are largely insensitive to the limits of the fitting region. If the peak model is correct then there is no systematic error. The slight difference between the error curves from integration and from curve fitting shows that the tedious computations for curve fitting do not pay off, the more so as curve fitting hinges on the correctness of the peak model.

Commonly, quantitative chromatographic analysis is associated with relative errors from 0·5 to 5%, in the experimental part. This means that only at very low S/N estimation errors will be a limiting factor.

4.1.3.2. Centre of gravity

The centre of gravity is calculated by numerical integration. The error is dependent on the same factors as in the area calculation. The integration limits are chosen so as to balance the systematic and random errors. The resulting error in the centre of gravity relative to the peak width is plotted in fig. 4.5, drawn curves.

The error from fitting of a Gaussian peak is plotted by the dashed curves. At high S/N the errors differ by a factor of 3. The reason is that with curve fitting the peak location is predominantly determined by samples on the peak body, whereas in the integration samples are weighted by their distance to the centre. Remote samples, where the ratio of signal amplitude to noise amplitude

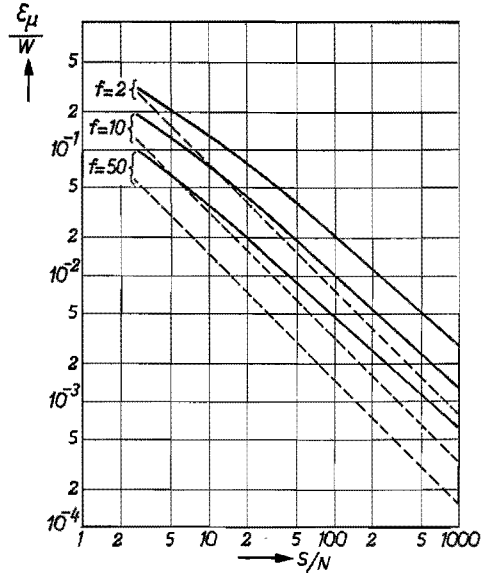


Fig. 4.5. Error in the centre of gravity relative to the peak width, as a function of the signal-to-noise ratio, at the sampling densities indicated, for a Gaussian peak. Drawn curves: resultant error with numerical integration; dashed curves: random error with curve fitting.

is low, therefore contribute more heavily. At low S/N, the integration limits contract so that these remote samples are omitted, resulting in a convergence of the errors from the two methods. For the calculation of the location parameter of a single peak we preferred the integration method because it is computationally simpler and does not require a suitable peak model. Moreover, the difference between the errors is insignificant in view of the possible peak shift for fused peaks.

4.1.3.3. Peak top

The top location is determined by the maximum of a parabola fitted to the top section of a peak, as described by Wijtvliet [4-1]) and Goedert and Guiochon [4-11]). Additional refinements included are the iterative centring and adjustment of the fitting section for improving accuracy and precision. The resulting error for Gaussian peaks and white noise is depicted in fig. 4.6.

Comparison of figs 4.5 and 4.6 shows that the errors are of the same order of magnitude, so that from this point of view there is no preference for one parameter over the other. For asymmetrical peak shapes the width of the fitting section is automatically reduced, to avoid large systematic errors. The error will however be higher than in fig. 4.6.

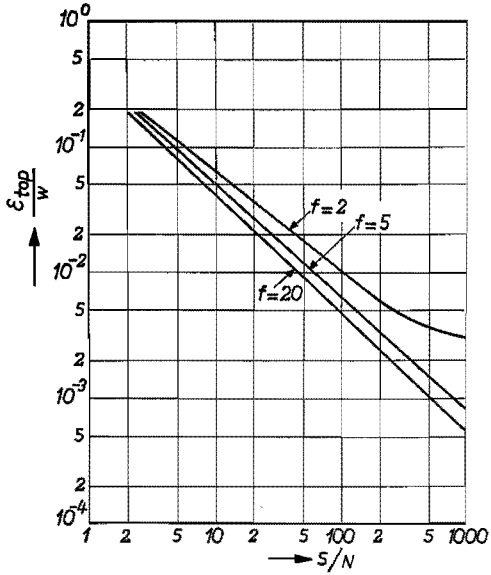


Fig. 4.6. Error in the peak-top location relative to the peak width, as a function of the signal-to-noise ratio, at the sampling densities indicated, for a Gaussian peak.

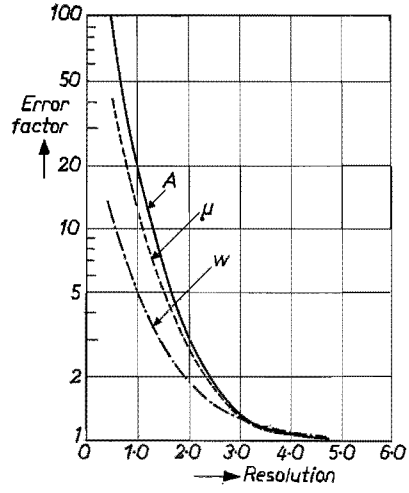


Fig. 4.7. Multiplication factors for the standard deviations of the parameters of overlapping peaks as a function of the resolution, derived for Gaussain peaks, of equal width (A = area, μ = location, w = width).

4.1.3.4. Multiple peaks

Overlapping peaks are dissected by curve fitting. The peak parameters are derived from the parameters in the model for fitting. For well-resolved peaks

(resolution > 5), the precision of the parameters is dependent on the signal-to-noise ratio and the sampling density. The errors are identical to the errors for curve fitting of single peaks (dashed curves in figs 4.4 and 4.5).

With decreasing resolution the errors are enhanced. The factors by which the errors for single peaks are multiplied at decreasing resolution are plotted in fig. 4.7. The dramatic increase of these factors at resolutions below 2 explains why it is meaningless to push the fusing limit further down: little significant information can be obtained from a composite curve of closely spaced peaks.

4.2. Application

Real chromatograms are not very expedient for demonstrating the correctness of the approach, because the true results are usually unknown. Real chromatograms are however very suitable to expose flaws in the design. They will also show the tolerances for deviations from the "model" form (i.e. uncorrelated noise, Gaussian peaks, smooth baseline) and the reliability of the results. We discuss the processing of one such deviating chromatogram in detail: this reveals the limitations of the present chromatogram more clearly than drawing examples from various chromatograms.

The chromatogram in fig. 4.8 incorporates a number of interesting features:
— correlated noise,
— solvent peak,
— closely spaced peaks riding on the solvent tail mask the background level,
— background signal does not return to its initial level,
— trace peaks.

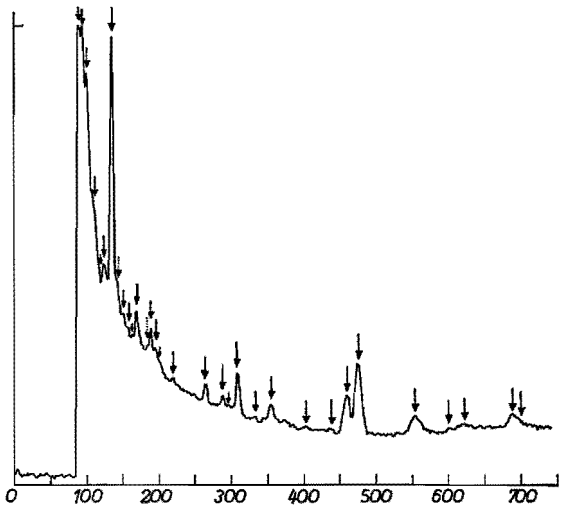


Fig. 4.8. Pesticide chromatogram. Bolds arrows indicate peaks detected. Broken-tail arrows indicate additional peaks detected during the curve fitting. Small arrows indicate neglected peaks.

4.2.1. *Experimental*

The chromatogram in fig. 4.8 is an analysis of pesticides in blood-serum extract, on a glass capillary column, with electron-capture detection. The chromatogram was provided by Franken, who described the chromatographic procedure extensively in ref. 4-12.

The detector signal, after amplification, was sampled, digitized and punched on paper tape. The data-acquisition system was assembled and described by Wijtvliet [4-1]). Sampling rate: 1 sample/s; in total 743 samples.

The chromatogram was processed on the Burroughs B6700 computer. The program is written in Algol-60. Figure 2.10 gives a scheme of the program's main procedures. The source length is about 20K words.

Processing time for this chromatogram was 60 seconds. A timetable is given in table 4-I. Although processing times on different computers are hard to compare, the processing time, excluding curve fit, is estimated to be 5–10 times longer than Wijtvliet's program on the same computer.

TABLE 4-I

Processing time for chromatogram 4·8 (seconds)

noise estimation	0·3
initial peak detection	1·8
spike filtering	0·2
peak detection	4·4
boundaries location	0·5
baseline correction	3·2
peak-top location (13 peaks)	1·5
moments calculation (13 peaks)	2·6
curve fitting	45·9

4.2.2. *Results*

Initial inspection: The noise amplitude was estimated 40 (arbitrary units). This agrees closely with the autocorrelation function of the first 80 samples, depicted in fig. 2.4*b*. Initial detection revealed 13 peaks; positions on time axis: 88, 124, 134, 169, 264, 288, 308, 457, 473, 442, 687. These are indeed the major peaks in the chromatogram. From these peaks the linear regression of the peak width against retention time was found: $w = 0{\cdot}5 + 0{\cdot}008\ t_R$. The linear correlation coefficient was high ($\approx 0{\cdot}9$) if the peak at 88 was omitted.

Peak detection: The 26 peaks detected are indicated in fig. 4.8 by bold arrows. Two additional peaks found in the curve-fitting procedure are marked by

broken-tail arrows. Positions 333, 402 and 436 appear to be spurious peaks. A serious omission is the peak at 295. Further omissions, surmised by visual inspection, are indicated by small arrows. The intermediate output showed that 109 peaks were provisionally detected from which 83 were subsequently rejected, being below the noise threshold. Most of the suggested omissions could be recovered with a slightly lower threshold, but at the expense of a larger number of spurious peaks.

Baseline correction: Figure 4.9 is a typical plot output. Below, the raw chromatogram is drawn together with the fitted baseline. The baseline-corrected signal is plotted on top on a 5-fold enlarged scale. The baseline is fitted piecemeal

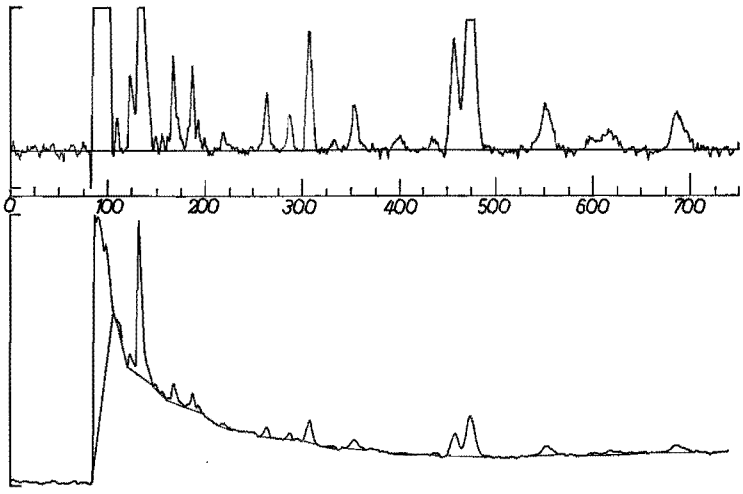


Fig. 4.9. Plot output of chromatogram 4.8. Below: chromatogram with fitted baseline; above: baseline-corrected chromatogram on 5-fold enlarged scale.

to each peak group. Under the peak conglomerates on the start of the solvent tail the fitted baseline has an unrealistic angularity. This is a consequence of the local baseline approximation: the traces of the background bracketing the peak groups are too small for indicating the curved trend, so that a linear approximation was used. Although a human operator might conjecture a smoother baseline by interpolation and extrapolation, it seems that in this part of the chromatogram any baseline is largely arbitrary.

On the whole the fitted baseline seems quite reasonable, although some minor details may be questioned, e.g. at 80, 200, 225, 315, 445.

Parameter estimation: The results are displayed in table 4-II. The column "specification" lists the type of the peak, i.e. whether a single peak or the model for curve fitting in case of overlap. A number of asterisks indicating the peak magnitude in a logarithmic measure facilitates the retrieval of the major peaks in an otherwise monotonous list of figures. To the major peak parameters an estimated standard deviation is added. Although these values only indicate the estimation errors, their specification was found useful to avoid that the com-

TABLE 4-II

NR.	START	END	SPECIFICATION	AREA	ERROR	TOP	S.D.	CENTER	S.D.	WIDTH	S.D.	PLATES	SKEW	EXCESS
1	83	106	****** GAUSS FIT	59816	27.4	87.22	0.39	87.22	0.39	1.69	0.26	2659	0.0	0.0
2	83	106	****** GAUSS FIT	70291	34.5	92.29	0.46	92.29	0.46	2.29	0.33	1931	0.0	0.0
3	83	106	****** GAUSS FIT	43920	33.1	98.64	0.81	98.64	0.81	2.29	0.60	1863	0.0	0.0
4	105	115	* SINGLE PEAK	1389	8.4	110.40	0.15	110.60	0.24	1.32	0.36	7007	-0.7	-3.3
5	120	148	*** GAUSS FIT	4798	12.5	124.23	0.81	124.23	0.81	2.22	0.86	3139	0.0	0.0
6	120	148	****** CONVOLUTION	45742	8.6	132.36	0.42	133.61	0.42	2.05	0.40	4126	0.4	0.6
7	120	148	*** GAUSS FIT	6393	45.3	140.46	1.39	140.46	1.39	2.22	1.13	4014	0.0	0.0
8	148	158	SINGLE PEAK	438	22.9	150.40	0.12	150.14	0.53	0.49	1.79	90039	-13.4	-124.9
9	150	163	SINGLE PEAK	954	19.0	156.73	0.09	156.73	0.31	1.14	0.29	14792	-1.2	-0.5
10	160	170	*** SINGLE PEAK	7229	21.1	168.25	0.05	169.29	0.14	3.19	0.14	2812	1.0	0.5
11	179	199	** GAUSS FIT	3664	4.8	187.77	0.07	187.77	0.07	1.49	0.08	15812	0.0	0.0
12	179	199	* GAUSS FIT	1394	7.4	193.84	0.12	193.84	0.12	1.51	0.13	16568	0.0	0.0
13	179	199	GAUSS FIT	817	19.8	183.51	0.27	183.51	0.27	1.29	0.25	20185	0.0	0.0
14	214	229	SINGLE PEAK	1339	11.3	219.01	0.08	221.02	0.48	3.17	0.30	4849	0.8	-0.8
15	255	272	** SINGLE PEAK	4370	3.7	263.93	0.05	263.68	0.19	2.91	0.20	8192	-0.8	0.5
16	279	294	* SINGLE PEAK	2309	6.6	287.19	0.10	287.03	0.27	2.36	0.27	14824	-0.5	-0.4
17	300	318	*** SINGLE PEAK	8394	1.9	307.92	0.09	308.07	0.08	2.33	0.08	17481	0.3	-0.3
18	329	338	SINGLE PEAK	590	20.0	333.38	0.18	333.40	0.60	2.02	0.52	27312	1.1	-0.6
19	345	366	** SINGLE PEAK	4244	4.2	353.97	0.04	354.74	0.27	3.71	0.25	9134	0.0	0.1
20	392	411	* SINGLE PEAK	1626	9.6	401.79	0.35	400.54	0.49	3.93	0.24	10373	-0.8	-0.7
21	430	444	SINGLE PEAK	969	15.1	437.08	0.60	436.93	0.65	2.66	0.46	27026	-0.5	-0.9
22	447	466	**** GAUSS FIT	12330	2.5	457.53	0.10	457.53	0.10	3.69	0.10	15389	0.0	0.0
23	447	466	***** GAUSS FIT	27525	1.2	473.70	0.06	473.70	0.06	4.51	0.06	11043	0.0	0.0
24	534	574	*** SINGLE PEAK	7719	2.9	552.07	0.16	552.17	0.26	5.81	0.11	9024	0.0	-0.2
25	587	636	* GAUSS FIT	1921	15.8	599.37	0.88	599.37	0.88	5.25	0.88	13044	0.0	0.0
26	587	636	* GAUSS FIT	3595	8.6	617.95	0.59	617.95	0.59	6.41	0.59	9281	0.0	0.0
27	672	706	** GAUSS FIT	5922	5.3	686.39	0.27	686.19	0.27	4.89	0.19	19885	0.0	0.0
28	672	706	* GAUSS FIT	1823	15.4	697.08	0.83	697.08	0.83	4.25	0.48	26958	0.0	0.0

puter output is attended by unwarranted significance. The plate numbers, under the heading "plates", show a large spread. This may be explained partially because the ratio of retention time to peak width is not strictly constant and also because the outliers come from the trace peaks whose widths are subject to large errors. The columns "skew" and "excess" list the quantities that carry additional information about the shape. For reasons of space the standard deviations are not listed. Generally these quantities are significant only for large peaks.

Six groups of overlapping peaks were found, viz. (86,92,99), (123,133), (188,194), (457,473), (599,617) and (687,698). The other peaks were supposed to be single. The baseline under the solvent peak is arbitrary so that a discussion of the first group is not meaningful. In any case, the large errors indicate that the results are unreliable. Figure 4.10*a* (below) shows the group (123,133) after fitting of two peaks to the baseline-corrected signal. For the larger peak the convolution model was found to give the best fit. However, the residuals (i.e. the discrepancies between the signal and the sum of the fitted peaks, plotted on top in fig. 4.10*a*) greatly exceed the random-noise amplitude. The shape of the residual signal leads to the introduction of an additional peak at 141. Although this improves the fit, cf. fig. 4.10*b*, the residuals are still too large

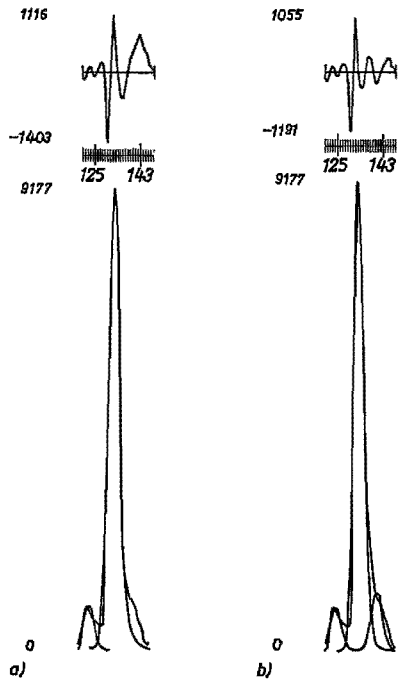


Fig. 4.10. Plot output of curve fit of overlapping peaks. Below: baseline-corrected signal and the fitted peaks. Above: residues (on the same scale); (*a*) fit of two initially detected peaks; (*b*) after introduction of an additional peak.

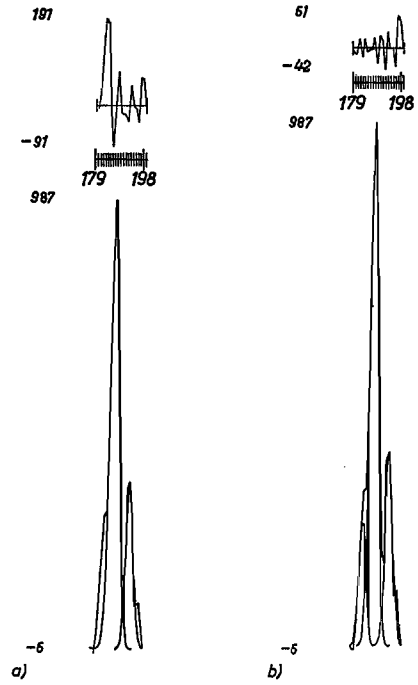


Fig. 4.11. See legend to fig. 4.10.

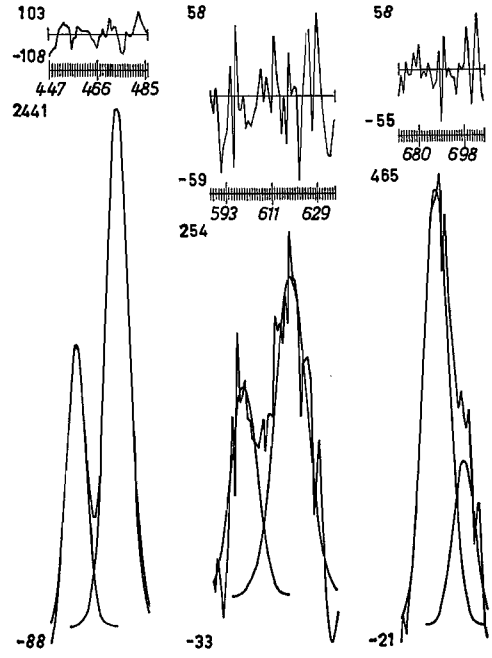


Fig. 4.12. Plot output of curve fit of overlapping peak groups in chromatogram 4.8.

owing to the questionable baseline. Figure 4.11 shows the same pattern for the peak group (188,194). Here, after introduction of an extra peak at 183, the mean of the residuals agrees with the noise amplitude. Figure 4.12 depicts the curve fits of the other peak groups. It is interesting to compare the results of the last two groups. In the groups (599,617) the signal-to-noise ratio S/N = 6 (for the largest peak); in the group (687,698) S/N = 12. As the errors are inversely proportional to S/N, the errors for the last group would be expected half of those in the former. However, in the first group the resolution Rs = 3·2 and in the last Rs = 2·3. Figure 4.7 indicates that the error factors for these resolutions differ by a factor 2 which explains why the errors in table 4-II are almost equal (cf. column "%error").

4.3. Tailoring to constraints

The sheer size of the program and the processing time will be prohibitive in many applications. The following modifications will be discussed:

— ways of speeding up the present program without affecting the quality of the results;

— simplification, taking some loss in detection and precision into account.

None of these modifications were actually tested, so the effects can only be roughly estimated.

4.3.1. *Reduction of processing time*

Generally, processing time depends on the number of samples, on the number and widths of the peaks and on the number of overlaps. Roughly speaking, the time for initial inspection and spike filtering is only proportional to the number of samples. The time for peak detection is proportional to the product of samples and mean peak width. The time for calculation of the moments and for peak-top location is proportional to the number of peaks and the mean peak width. The time for curve fitting is proportional to the number of samples in the fitting region, and also increases with the third power of the number of peaks in the peak group.

The program can be speeded up by a different arrangement of some processing stages and combination of similar operations, e.g. combination of initial peak detection and spike filtering, or combination of peak detection and boundary location. The gain will be about 10% of the processing time before curve fitting.

As most of the processing time is devoted to curve fitting, savings here will be more rewarding. The time for curve fitting depends on

— the number of iterations,

— the number of operations in each iteration.

The number of iterations depends on the quality of the initial estimates, the

optimization algorithm and the convergence criteria. There is no easy way to improve these. The number of operations depends on the number n of peaks to be fitted, on the number m of parameters in the peak model and on the number p of samples for fitting. Generally, p will be in the range 1·5 mn to 3 mn. Set-up of the point equations costs the evaluation of 2 mnp function values if the derivatives are calculated numerically or the equivalent of 1·5 mnp function evaluations for analytical derivatives. The conversion to the normal equations takes $\frac{1}{2}p\,(mn)^2$ multiplications of matrix elements. Inversion of a symmetrical $mn \times mn$ matrix with Choleski's algorithm takes about $\frac{1}{6}\,(mn)^3 + 2\,(mn)^2$ multiplications [4–13]. From these expressions it follows that the reduction of either m, n or p will drastically reduce the number of operations:

— n may be temporary reduced by fitting first the larger peaks, adding the smaller peaks after convergence;
— by starting with a simple model (Gaussian peak, fixed width), the number of parameters is small in the initial iterations;
— the function values and the partial derivatives are not evaluated at all p samples, but only at samples on the peak body;
— p may be varied during the iterations, starting with a very low value, e.g. 1·2 nm, and increasing gradually until in the final iteration all samples are taken into account;
— parameters may be optimized sequentially instead of simultaneously, as will be elaborated below.

Some preliminary tests showed that these strategies may reduce the time for curve fitting by a factor 2–4.

4.3.2. *Simplification*

Often the sophisticated data processing is unnecessary because the accuracy and precision of the experimental part are an order of magnitude worse. With some loss in precision, the program may be considerably reduced in length and required processing time.

Initial inspection: Iterative noise estimation can be speeded up by larger thresholds. Initial peak detection may be done without the intricate regression procedure for filter-width updating.

Peak detection: The matched filter becomes tedious for broad peaks because of the large number of iterations. An almost equally effective filter, depicted in fig. 4.13, requires in its recursive form only 4 multiplications, irrespective of the filter width. Using inflection-point pairs, shoulder peaks can still be detected.

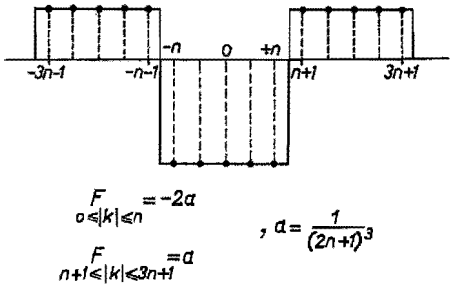


Fig. 4.13. Profile of a filter for calculating second derivatives.

Baseline correction: Instead of fitting a polynomial, the boundaries of the peak group can be connected by a straight line.

Parameter estimation: For single peaks the moments can be calculated by numerical integration between the peak boundaries. The top location is estimated by fitting a single parabola over the region (—1·2 w, +1·2 w). For all except the very smallest peaks the precision will be satisfactory.

Curve fitting is still the only accurate means of apportioning overlapping peaks. A bi-Gaussian peak model is able to cover a wide range of actual shapes. A simplified, fast procedure for curve fitting is based on the following considerations:

— If an unredundant number of samples is taken, i.e. $p = mn$, the point equations need not be converted to the normal equations.
— The samples for fitting are taken near the positions where the parameters are most sensitive (optimum in partial derivatives).
— Similar parameters are optimized simultaneously, while other parameters are kept constant. For example, the areas are optimized while positions and widths are kept constant, then the positions are optimized while areas and widths are constant, etc. Because each peak is supposed to interfere only with its direct neighbours, this results in a set of tri-diagonal equations that can be solved with 5 mn multiplications (Choleski: $\frac{1}{8}(mn)^3 + 2(mn)^2$).

These simplifications are estimated to reduce the program length to one third and the processing time to about one tenth, thus bringing it on a par with Wijtvliet's program while maintaining a number of important advantages:

— automatic, by requiring no presetting of processing parameters;
— detection of overlapping peaks and shoulder peaks;
— local baseline to each peak group;
— separation of overlapping peaks by curve fitting.

REFERENCES

4–1) J. J. M. Wijtvliet, Thesis Technological University Eindhoven, 1972.
4–2) A. B. Littlewood, T. C. Gibb and A. H. Anderson, in C. L. A. Harbourn (ed.), Gas chromatography 1968, Proceedings of the Seventh International Symposium, Institute of Petroleum, London, 1969, p. 297.
4–3) A. H. Anderson, T. C. Gibb and A. B. Littlewood, Anal. Chem. **42**, 434, 1970.
4–4) A. H. Anderson, T. C. Gibb and A. B. Littlewood, in A. Zladkis (ed.), Advances in chromatography 1970, Miami, 1970, p. 75.
4–5) A. W. Westerberg, Anal. Chem. **41**, 1770, 1969.
4–6) J. R. Morrey, Anal. Chem. **40**, 905, 1968.
4–7) G. Brouwer and J. A. J. Jansen, Anal. Chem. **45**, 2239, 1973.
4–8) S. N. Chesler and S. P. Cram, Anal. Chem. **43**, 1922, 1971.
4–9) S. N. Chesler and S. P. Cram, Anal. Chem. **44**, 2240, 1972.
4–10) M. Goedert and G. Guiochon, J. chromatog. Sci. **11**, 326, 1973.
4–11) M. Goedert and G. Guiochon, Chromatographia **6**, 39, 1973.
4–12) J. J. Franken, in S. G. Perry (ed.), Gas Chromatography 1972, Institute of Petroleum, London, 1973, p. 77.
4–13) R. Zurmühl, Matrizen, Springer, Berlin, 1964, p. 66.

5. IDENTIFICATION

5.1. Introduction

A general review of techniques for identification by chromatography, including ancillary techniques, is given by Leathard and Shurlock [5-1]). This chapter will be confined to computer-aided identification from chromatographic data only.

It is convenient to distinguish between primary and secondary identification. Primary identification means structural elucidation of an unknown compound, without relying on tabulated data. Secondary identification is based on recognition: the compound must have been identified before and analysed under the same conditions. As chromatographic analysis provides only one quantity specific for an eluted compound, which does not allow interesting inferences about molecular structure, it is mainly suitable for secondary identification.

Secondary identification is unreliable, without further evidence. Due to its limited precision, a retention value should not be considered as a point on a uni-dimensional scale but as the mean of a probability distribution extending over a range of values. Probability distributions are bound to overlap to some extent, so that identification is always uncertain. More fundamentally, identification is unreliable because data from only a small number of the myriad of chemical compounds will be available in the data collection. Circumstantial evidence, such as sample origin and pretreatment steps, are essential for reliable identification. For these reasons, we believe that computerized table searching should be based on a probabilistic matching criterion, and the circumstantial evidence should be reflected in the selection of certain classes of compounds and, possibly, different a priori ratings for each of these classes.

The reliability can be increased by analysing the sample on different stationary phases or at different temperatures. Coinciding retention data may be resolved, and structural information can be obtained. The matching criterion should therefore be able to combine the results of different analyses.

Attempts have been made to establish relations between molecular structure and retention values, so that chromatography can also be used for primary identification. Two aims can be distinguished:

— the identification of structural features from retention data of one compound on different stationary phases;
— the prediction of the unknown retention value of a compound from the known values of related compounds on one stationary phase.

The first method is based on a clustering of compounds with a common specific group if the retention values on different phases are plotted in a multi-dimensional space [5-2]). At present, this still requires a lot of inspired guesswork and, at best, contributes pieces of information. The prediction of retention values is

based on the assumption that the contributions of structural elements are additive [5–3]).

In the following sections some aspects of computer-aided identification will be discussed. A probabilistic matching criterion is presented in sec. 5.2. Section 5.3 discusses the structuring of the data collection, necessary for efficient searching. In sec. 5.4 some examples are given that demonstrate the application of the matching criterion. In sec. 5.5 a method for predicting retention values is investigated.

5.2. Matching criterion

Due to the finite precision of both the measured retention value and the retention values in the data collection ("library"), unique identification is impossible. It is the function of the matching criterion to weigh the evidence between possible candidates.

There is also the possibility that the true compound is not present in the library. This possibility may not be neglected even if there is perfect agreement between the measured value and some library value, although it becomes more likely if no suitable library value can be found. To assess this would require knowledge of the number of possible compounds and the distribution of the retention values. These are not available, so only the compounds in the file are considered.

Let R_m represent the measured retention value, and R_i, $i = 1, \ldots, N$ the N values in the library; they are both random quantities with variances σ_m^2 and σ_i^2 characteristic of the precision of measurement. A certain compound A in the library is the more likely the right compound, the closer R_a is to R_m. It seems reasonable to asume that if A is the right compound, then the discrepancy

$$\Delta_a = R_m - R_a$$

will be normally distributed with zero mean and variance equal to the sum of the variances of both retention values:

$$\sigma_\Delta^2 = \sigma_m^2 + \sigma_a^2.$$

In other words, the conditional probability density of a discrepancy Δ_a for a given substance A is

$$p(\Delta_a \mid \mathrm{A}) = \frac{1}{[(\sigma_m^2 + \sigma_a^2)\, 2\pi]^{1/2}} \exp\left(-\tfrac{1}{2} \frac{\Delta_a^2}{\sigma_\Delta^2}\right). \tag{5.1}$$

The conditional probability of A, given the discrepancy Δ_a, follows from Bayes' rule:

$$P(\mathrm{A} \mid \Delta_a) = \frac{p(\Delta_a \mid \mathrm{A})\, P(\mathrm{A})}{p(\Delta_a)}\,; \tag{5.2}$$

$P(\mathrm{A})$ is the a priori probability of compound A. The denominator $p(\Delta_a)$ is the unconditional probability density of the observed discrepancy, and this is assumed to be uniform and constant for all library compounds. If A and B are two possible candidates, the likelihood ratio L_{ab} characterizes numerically the support by the measurement R_m for A as against B:

$$L_{ab} = \frac{p(\Delta_a \mid \mathrm{A})\, P(\mathrm{A})}{p(\Delta_b \mid \mathrm{B})\, P(\mathrm{B})}. \tag{5.3}$$

The likelihood ratio expresses the relative merit of two compounds. An absolute merit is expressed by the two-sided excess probability for the observed discrepancy:

$$P(> |\Delta_a| \mid \mathrm{A}) = 2 \int_{|\Delta_a|}^{\infty} p(\Delta \mid \mathrm{A})\, \mathrm{d}\Delta = 1 - \mathrm{erf}\left(\frac{\Delta_a}{\sigma_\Delta \sqrt{2}}\right), \tag{5.4}$$

where $\mathrm{erf}\,(x)$ denotes the well-known error function.

So far only a single measurement was considered. The criterion can easily be extended for multiple measurements of the same compound under different conditions. Let there be n measurements, $R_{m,k}$, $k = 1, \ldots, n$ and hence n discrepancies for a candidate compound. The likelihood ratio for two compounds A and B contains the product of the conditional probability densities:

$$L_{ab} = \frac{P(\mathrm{A}) \prod_n p(\Delta_{a,k} \mid \mathrm{A})}{P(\mathrm{B}) \prod_n p(\Delta_{b,k} \mid \mathrm{B})}. \tag{5.5}$$

The n standardized discrepancies for a compound A

$$\Delta_{a,k}^{*} = \frac{R_{m,k} - R_{a,k}}{(\sigma_{m,k}^{2} + \sigma_{a,k}^{2})^{1/2}} \tag{5.6}$$

are assumed to be n independent normally distributed variables with zero mean and unit variance. Accordingly, the sum of the squared standardized discrepancies follows a χ^2-distribution with n degrees of freedom. Consistent with the definition in (5.4), the absolute merit of a compound A is expressed by the integral probability for a χ^2-distribution with n degrees of freedom of exceeding

$$\sum_{k=1}^{n} (\Delta_{a,k}^{*})^2.$$

Values for this probability are found in ref. 5-4.

5.3. File structure and search

Identification by searching in a library of retention data is well suited to be done by a computer. The method described below was developed for an EL-X8 computer, having a core store capacity of 48K words of 27 bits and 512K drum capacity. Retention indices are used for identification [5–5]).

In an unstructured library a certain value can only be found by searching sequentionally through the whole library. This is obviously not very efficient if the library is large. Three methods are used to improve the efficiency of the search:

— classification of the data, to limit the scope of the search;
— coding of the information, to reduce the size of the searched classes;
— sorting of the values, so that more-rapid searching methods can be used.

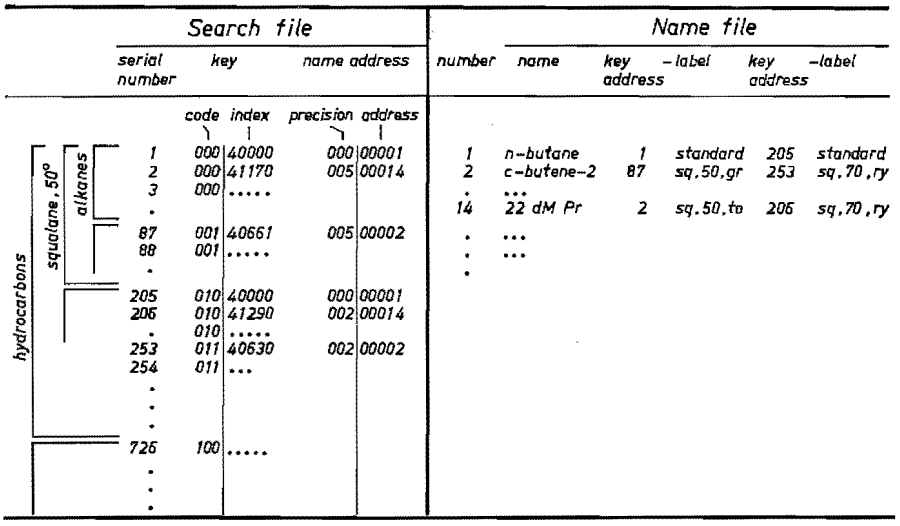
A classification is made at three levels. Primary classes are based on chemical nature, e.g. hydrocarbons, steroids, alcohols, etc. It is assumed that it is known from circumstantial evidence which class must be searched. A second classification is made according to the conditions of the analysis (stationary phase, temperature). This will also be known in advance. Further subclassification is made for related compounds, e.g. hydrocarbons are subdivided in aliphatic, aromatic, cyclic, unsaturated, etc. By selecting certain subclasses for searching and, possibly, giving them different a priori ratings, further circumstantial evidence can be introduced.

Coding is especially important for computer search, because of the limited capacity of the high-speed random-access store. Data transports between different stores usually take considerable time. Retention indices range between 100 and 4000, and the state-of-the-art precision is about 0·05 unit in the lower range (< 1000) and about 0·2 in the upper range. So 5 digits or 16 bits are sufficient. By putting the coded classifications (1 digit for each level) before the index value, the limitation of the search is automatically obtained. For example, 3,4,4 trimethyl cis-pentene-2 has the index 747·08 on squalane at 70 °C (code = 3). Hydrocarbons are coded 1 and alkenes are coded 2, so the stored value is 13274708. Because the names of many chemical compounds are quite bulky, they were put in an auxiliary name file. The position in the name file ("address") must be stored together with the coded index. In this way the search file is small, and the names need only to be stored once. The standard deviation of the retention index must also be stored together with the coded index, because it is used in the matching criterion.

By putting the coded retention indices in ascending order, the binary search method can be used. This is much faster than sequentional search (for N data: $1 + \ln_2 N$ vs N comparisons).

The library structure is pictured in table 5-I. Each record in the search file consists merely of two words. The key word is the coded retention index. The

TABLE 5-I



other word contains the precision and the name address. The records in the name file have a variable length. Each record contains the name of the compound and one or more pointers to its retention values in the search file. To each pointer a label is attached that contains additional information such as literature source, instrumental conditions, etc. The effect of the structuring is that the entire search may be done in the search file. If one or more matching values have been found, the name of the compound and further information can be retrieved via the link address. Via the key addresses, other retention values of this compound can be directly retrieved, which is very useful for further investigations.

The search proceeds as follows: The library, stored on magnetic tape, is read into the computer store at the start of the processing. The retention index for identification is provided with an estimated precision and the coded conditions. The primary class must also be specified. For the relevant subclasses an a priori rating must be specified, which is used in the matching criterion (5.3). Finally, a threshold for the excess probability (5.4) must be specified. The index and the classifications are combined to one value, composed in a similar way to the coded values in the library. From this value, the precision of measurement and the specified threshold level, the range is calculated in which appropriate library values must be situated. The centre of the range is located by binary search. The matching criterion is applied to the values in the range, and those having a probability above the threshold, if any, are listed with the likelihood ratio to the most probable reference and the excess probability. Names and further information are printed out.

TABLE 5-II

a

index: 759·8, standard deviation 0·8, squalane, 50 °C, hydrocarbon compounds above threshold 0·5%:

759·4 2,3,3 trimethyl pentane; s.d. 0·1
likelihood ratio: 0·95
excess probability: 61·7%

760·1 2,3 dimethyl hexane; s.d. 0·1
likelihood ratio: 1
excess probability: 71·8%

761·4 2 methyl, 3 ethyl pentane; s.d. 0·1
likelihood ratio: 0·14
excess probability: 4·5%

b

index: 764·1, standard deviation 0·8, squalane, 70 °C, hydrocarbon compounds above threshold 0·1%

761·5 2,3 dimethyl hexane; s.d. 0·1
likelihood ratio: 0·007
excess probability: 0·12%

763·4 2,3,3 trimethyl pentane; s.d. 0·1
likelihood ratio: 0·90
excess probability: 38·2%

763·5 2 methyl, 3 ethyl pentane; s.d. 0·1
likelihood ratio: 1
excess probability: 45·4%

c

Combined results

— 2,3,3 trimethyl pentane
likelihood ratio: 1
excess probability: 70%

— 2 methyl, 3 ethyl pentane
likelihood ratio: 0·14
excess probability: 5%

— 2,3 dimethyl hexane
likelihood ratio: 0·007
excess probability: 0·12%

5.4. Examples

Retention data of hydrocarbons on squalane, taken from Rijks [5–6]), were compiled in the library. As a test, the index 759·8 measured by Tourres [5–7]) for 2,3,3 trimethyl pentane on squalane at 50 °C was introduced. The estimated standard deviation is 0·8. All hydrocarbon subclasses were given equal a priori rating. The probability threshold was set at 0·5%. The results of the search are given in table 5-IIa. There appears to be no clear-cut choice. Then the index 764·1 of this compound on the same phase but at 70 °C was used. Again a number of candidates is found (table 5-IIb). Combining both results, the ambiguity is resolved (table 5--IIc).

In practice usually a mixture of unknown compounds will be analysed. The precision estimate must take a possible peak shift from overlapping peaks into account [5–6]). If the same mixture is analysed on a second stationary phase, it is unknown which peaks do correspond to the same compound. Consider a hypothetical mixture which gives three peaks on squalane and on citroflex (table 5-III, data taken from Rijks [5–6])). Only one peak can be identified directly, assuming that the library is complete. For the other peaks more candidates are listed. If the results of the two analyses are combined by eliminating those references that do not occur on both lists, most peaks are uniquely identified. Only the presence of 3,4 dimethyl pentene-1 cannot be decided from the given data (in this case the peak areas will decide).

This example demonstrates that the combination of analyses enhances the identification. However, this identification is based on elimination: a candidate from one analysis shows to be absent because the other chromatogram is empty at the corresponding position. With increasing peak density the number of candidates that can be eliminated will decrease progressively. For complex mixtures the combination of analyses does not yield much additional information.

TABLE 5-III

squalane, 50 °C		citroflex, 50 °C	
index	candidates	index	candidates
637·5	3,4 dimethyl pentene-1	667·0	2 methyl hexane
	benzene		3,4 dimethyl pentene-1
	2,4 dimethyl pentene-1		4,4 dimethyl c-pentene-2
666·5	2 methyl hexane	729·5	3 methyl c-hexene-2
693·0	2,2 dimethyl t-hexene-3		2,5 dimethyl hexane
	3 methyl c-hexene-2	775·0	3 methyl heptane
			benzene

5.5. Structure-retention relations

The prediction of retention values is based on the assumption that contributions of structural elements are additive. Schomburg [5–8]) elaborated a system of increments for substituent groups on a given backbone structure. Fairly accurate predictions can be made for single substituents, but if more substituents are used, their interaction must be accounted for by second- and third-order corrections so that the system becomes complicated and obscure.

A more general approach is to define a set of structural units from which all molecules from a certain class of compounds can be assembled. If there are m such units, each molecule is described by an m-dimensional vector

$$\mathbf{n} = (n_1, \ldots, n_m);$$

n_j denotes the number of units j. If the retention contribution of unit j is denoted as c_j, the additivity assumption predicts the retention value of a compound as

$$R = \sum_{j=1}^{m} n_j c_j = \mathbf{n}^T \cdot \mathbf{c}. \tag{5.7}$$

The increment vector $\mathbf{c}$ can be calculated if for a sufficient number of members of the class the retention value is known. Let for p molecules $\mathbf{n}_k$, $k = 1, \ldots, p$, the retention value be R_k. This provides a set of p equations:

$$\mathbf{n}_k{}^T \cdot \mathbf{c} = R_k.$$

In matrix notation:

$$N \cdot \mathbf{c} = \mathbf{R}. \tag{5.8}$$

If the set is overdetermined ($p \geqslant m$), and the molecular description is not redundant, the increment vector $\mathbf{c}$ can be calculated by the least-squares method:

$$\mathbf{c} = (N^T \cdot N)^{-1} \cdot N^T \cdot \mathbf{R}. \tag{5.9}$$

We made some calculations on the basis of sets of structural units for aliphatic alkanes, described by Walraven [5–2]).

The four-digit set distinguishes primary, secondary, tertiary and quaternary carbon atoms: (n_p, n_s, n_t, n_q). This set is however redundant, as it is easily seen that for alkanes $n_p = 2 + n_t + 2n_q$. So three units are sufficient: $\mathbf{n} = (n_s, n_t, n_q)$. Calculation of the increment vector from a set of 25 alkanes, randomly chosen, showed that the reproduction accuracy (i.e. for compounds used in the calculation set) was about 20 index units. The prediction accuracy for another 25 alkanes was about 35 i.u. However, a comparabe prediction accuracy could be obtained from a calculation set of 10 alkanes.

A refined description set is the bond-type coding. Nine types of bonds exist

between carbon atoms in alkanes (primary–primary, primary–secondary, etc.). The reproduction accuracy was found to be about 5 i.u. and the prediction accuracy about 8 i.u.

An even more detailed set was constructed by differentiating between different neighbourhoods for each carbon atom (e.g. a tertiary carbon connected to two secondary carbons and one primary carbon). This set contains 69 units. Removing the redundant units, 35 units occur in alkanes up to C_9. This implies that at least 35 well-selected compounds are required for the calculation. The reproduction accuracy for a set of 45 compounds was about 0·8 i.u. and the prediction accuracy was 1·5 i.u.

These results show a basic dilemma: for an accurate prediction a large number of structural units must be distinguished, so that a large number of compounds is required for the calculation and few remain for a true prediction.

Takács et al.[5–9]) proposed a set of about 600 units! Although a number of them cannot occur in reality, the number actually used ($\pm$ 130) is larger than the number of alkanes measured at present. As the set of equations is underdetermined (and the description set contains a number of redundant units), it appears that most of the increments are arbitrarily assigned.

For other classes, such as alkenes, the greater structural variety necessitates the introduction of a larger number of structural units to attain a comparable accuracy of prediction. At present the number of available data is too small for predictions of interesting accuracy (i.e. < 5 i.u.). In these cases the more limited approach of Schomburg [5–8]) is still the appropriate method.

REFERENCES

[5–1]) D. A. Leathard and B. C. Shurlock, Identification techniques in gas chromatography, Wiley, London, 1970.
[5–2]) J. J. Walraven, Thesis Technological University Eindhoven, 1968, Ch. 5.
[5–3]) L. S. Ettre, Chromatographia **7**, 39, 1974.
[5–4]) M. Abramowitz and I. A. Segun (eds), Handbook of mathematical functions, Dover, New York, 1968.
[5–5]) L. S. Ettre, Chromatographia **6**, 489, 1973.
[5–6]) J. A. Rijks, Thesis Technological University Eindhoven, 1973, Ch. 4.
[5–7]) D. A. Tourres, J. Chromatog. **30**, 357, 1967.
[5–8]) G. Schomburg, Chromatographia **4**, 280, 1971.
[5–9]) J. Takács, C. Szita and G. Tarján, J. Chromatog. **56**, 1, 1971.

List of main symbols

A	peak area
a	coefficient
a_k	coefficient of kth term in a polynomial
b	coefficient
$\mathbf{b}$	vector, left-hand side of normal equations
c	coefficient
d	integer
E_k	discrepancy between measurement and fitted function at $t = t_k$
f	number of samples over the peakwidth (or, peakwidth expressed in number of samples)
F	F-test value
F_k	sampled filter weight function $f(\tau)$ at $\tau = k\,\Delta$
$f(\tau)$	filter weight function
$g(\tau)$	normalised peak model (unit area and centred in the origin)
G_k	sampled value of $g(\tau)$ at $\tau = k\,\Delta$
i, j, k	integer values
K	proportionality constant
m, n	integer values
m_n	nth central moment
$n(t)$	random noise signal (zero mean)
N	plate number/number of objects in a set
N_k	sampled value of $n(t)$ at $t = k\,\Delta$
p	parameter
$\mathbf{p}$	parameter vector
$\mathbf{\Delta p}$	correction vector for $\mathbf{p}$
$p(x)$	probability distribution of statistical quantity x
$p_n(t)$	polynomial of degree n
Rs	resolution
S, $S(\mathbf{p})$	sum of squares, for parameter vector $\mathbf{p}$
S/N	signal-to-noise ratio
t	time
t_R	retention time
w_f	width of the matched filter
w, w_p	peakwidth (distance between maximum and inflection point)
Y_k	sampled signal $y(t)$ at $t = t_k$
$y(t)$	signal, made up of peaks, baseline and noise
Z_k	filtered sampled signal at $t = t_k$
Z	matrix of normal equations (symmetrical)
Z^{-1}	inverse of Z
$z(t)$	filtered signal

z_{ij}	ijth element of Z
z_{ij}^{-1}	ijth element of Z^{-1}
γ_1	skew (eq. (2.5))
γ_2	excess (eq. (2.5))
Δ	sample interval
ε_A	combined error in the area (root of the sum of squared systematic error and variance)
ε_μ	combined error in the centre of gravity
μ	centre of gravity
μ_p	mean value of estimated quantity p
μ_p^0	true value of quantity p
η	normalised time ($\eta = (t - \mu)/w$)
σ, σ_p	standard deviation (of parameter p)
σ_y	mean noise amplitude (rms)
τ	dummy parameter/time constant in first-order system response

Summary

Data processing may be understood as the transformation of a chromatogram into operative information. This transformation is commonly effected in three steps, viz. extraction, identification and interpretation. The present work is confined to extraction and identification.

For the extraction our aim was to design a computer program which satisfies three requirements, viz. low detection limits, optimum accuracy and precision of the results and automatic processing. This combination of requirements is uncommon because so far high-quality data processing has demanded a considerable amount of judgement from the user.

The goal was achieved, on the one hand, by elaborating existing techniques and developing new methods for the program parts and, on the other hand, by integrating these parts into a program in such a way that information obtained from each part is used to adapt the subsequent parts more closely to the processed signal.

The main improvements in the parts are:

Application of matched-filter detection, which is known to yield an optimum detection. The threshold level is adjusted to the noise amplitude in the signal. The resolution limit at which a composite peak can be distinguished from a single peak is evaluated.

A method of peak-boundary location is described which yields correct boundaries on a sloping baseline. This eliminates an important source of error in subsequent baseline correction and in the area estimation.

A local baseline correction to each peak group is obtained by fitting a polynomial to the bracketing baseline sections.

The coordinate of the peak maximum is calculated by fitting a parabola to points around the top. An iterative procedure is described which adjusts the number of points to the noise amplitude and the peak shape, so as to minimize the sum of systematic and random errors.

For single peaks the area and the centre of gravity are calculated by numerical integration. An iterative procedure is developed which adjusts the integration limits for optimum accuracy and precision.

Overlapping peaks are separated by curve fitting. The peak model is selected according to the actual shape of the peaks.

An analysis of the systematic and random errors for a Gaussian peak showed that the accuracy and precision equal that of the best methods used so far.

Further advantages are that the program is easy in use, and that the results

are consistent, independent of the user's skill.

The main drawbacks are that large amounts of storage are used and that the processing time is long. Some possibilities are indicated for a more efficient programming. It is also assessed that a number of simplifications can be made which reduce the processing time to that of other programs, while maintaining a number of important advantages.

Identification based on comparison of the measured retention value with tabulated values is described. A matching criterion is proposed which takes the precision of the measured value and the tabulated values into account and allows for the introduction of further evidence. Applying this criterion yields both an absolute and a relative measure for the likelihood of matching compounds. Results of different analyses can be combined into a total score.

Data structuring for efficient search and direct retrieval of coherent information is described.

The possibilities and limitations of relations between structure and retention as an aid to identification are demonstrated.

Samenvatting

Het onderwerp van dit proefschrift is de verwerking van gaschromatografische gegevens. Men kan deze gegevensverwerking zien als een transformatie van het chromatogram tot bruikbare informatie. De transformatie vindt gewoonlijk in drie stappen plaats: van het geregistreerde signaal tot analytische gegevens (extraktie), vervolgens tot chemische gegevens (identifikatie) en tenslotte tot praktisch bruikbare informatie (interpretatie). In dit proefschrift komen alleen extraktie en identifikatie aan de orde.

Voor de gegevens extraktie stelden we ons tot doel om een computerprogramma te ontwerpen dat, uitgaande van het bemonsterde en gedigitaliseerde signaal, zonder verdere informatie over de kenmerken van het signaal, de analytisch relevante gegevens (piekoppervlakken, retentietijden, etc.) bepaalt met zo groot mogelijke nauwkeurigheid. De kombinatie van deze eisen is uniek: tot nu toe ging automatisering veelal ten koste van de kwaliteit van de resultaten en was voor uiterste nauwkeurigheid een veel aandacht kostende procedure nodig.

We hebben geprobeerd dit doel te bereiken door enerzijds voor de onderdelen van de verwerking, zoals piekdetektie, basislijnkorrektie en berekening van piekparameters, bestaande methoden te verfijnen of betere methoden te ontwerpen, en anderzijds door deze onderdelen zó in een programma te integreren dat de resultaten van elk onderdeel in de daaropvolgende delen worden benut voor een betere, aan het signaal aangepaste bewerking. Door de verwerking in stappen te doen, namelijk eerst globaal en vervolgens steeds gedetailleerder, krijgt elk chromatogram een zo goed mogelijk aangepaste behandeling.

De belangrijkste verbeteringen in de onderdelen zijn de volgende:

Voor piekdetektie wordt de "matched filter" methode gebruikt, waarvan bekend is dat hiermee optimale detektie wordt verkregen. De drempelwaarde voor de detektie wordt aangepast aan de ruisintensiteit in het signaal. De grens waarbij twee overlappende pieken nog apart kunnen worden gedetekteerd is berekend.

De piekgrenzen worden zodanig bepaald dat ze ook op een hellende basislijn korrekt zijn. Dit voorkomt fouten bij de basislijnkorrektie en bij berekening van het piekoppervlak.

De basislijnkorrektie wordt voor iedere piek of groep van overlappende pieken apart bepaald. Dit is nodig om ook bij een golvende of diskontinue basislijn de juiste korrektie te krijgen.

De plaats van de piektop wordt bepaald door een parabolische interpolatie tussen punten rond de top. Een iteratief mechanisme zorgt ervoor dat het aantal punten aangepast wordt aan de ruisintensiteit en aan de piekvorm (hoe sterker

de ruis hoe meer punten; hoe groter de asymmetrie hoe minder punten). Op deze manier wordt de som van de systematische en toevallige fouten verminderd.

Voor enkelvoudige pieken worden het oppervlak en het zwaartepunt berekend door numerieke integratie. De geldigheid van de resultaten is dus niet beperkt tot bepaalde piekvormen, zoals bij vele meetkundige benaderingen. De integratiegrenzen worden weer iteratief aangepast aan de ruisintensiteit en de werkelijke piekvorm, zodanig dat de som van de systematische en toevallige fouten wordt geminimaliseerd.

De gegevens van overlappende pieken worden door "curve fitting" berekend. Hierbij wordt de keuze van het meest geschikte piekmodel gemaakt op grond van de vorm van de pieken in het chromatogram.

Een analyse van de systematische en toevallige fouten voor een Gaussische piekvorm laat zien dat de bereikte resultaten tenminste die van de beste methoden tot nu toe evenaren. Verder verklaart deze analyse enkele bekende studies die experimenteel vonden dat de toevallige fouten omgekeerd evenredig zijn met de signaal–ruis verhouding en evenredig met de wortel uit het bemonsteringsinterval, zolang opeenvolgende ruisbijdragen ongekorreleerd zijn. Bovendien laat de analyse zien dat de toevallige fouten bij curve fitting exponentieel toenemen naarmate de scheiding tussen twee pieken afneemt. Dit betekent dat uit de som-curve van slecht gescheiden pieken weinig significante informatie kan worden verkregen.

Naast de grote nauwkeurigheid heeft het ontwikkelde programma twee belangrijke voordelen. Allereerst is het eenvoudig te gebruiken omdat voor verwerking geen extra informatie over het chromatogram nodig is. Verder zijn de resultaten consistent en niet afhankelijk van de kennis van de gebruiker.

Het programma heeft als nadeel dat het zeer veel geheugenruimte en rekentijd vergt. Er worden daarom enkele mogelijkheden opgesomd voor meer efficiënte programmering. Daarnaast wordt aan de hand van een ruwe analyse geconcludeerd dat het mogelijk is om door vereenvoudigingen de rekentijd terug te brengen tot die van andere programma's met behoud van een aantal essentiële voordelen.

Voor identifikatie door vergelijking van een gemeten retentiewaarde met getabelleerde waarden wordt een kriterium op basis van kansen voorgesteld. Hiermee kan zowel de nauwkeurigheid van de gemeten en getabelleerde waarden in rekening worden gebracht als de aanwezige extra informatie, bv. over het soort monster. Met het kriterium kan zowel een absolute maat voor de kans van potentiële kandidaat-stoffen worden berekend als een relatieve maat. Bovendien kunnen de resultaten van verschillende analyses worden gekombineerd tot één totaal-waardering.

Een ander aspect dat aan de orde komt is het struktureren van de tabellen in het computergeheugen zodat enerzijds het zoeken efficiënt kan verlopen en anderzijds op elkaar betrekking hebbende gegevens direkt kunnen worden teruggevonden.

Tenslotte wordt aan de hand van een voorbeeld duidelijk gemaakt hoe struktuur–retentie-verbanden als hulpmiddel voor identifikatie kunnen dienen.

Dankwoord

De direktie van het Natuurkundig Laboratorium van de N.V. Philips' Gloeilampenfabrieken ben ik zeer erkentelijk voor de gelegenheid die mij geboden is om het werk dat in dit proefschrift is beschreven uit te voeren aan de Technische Hogeschool Eindhoven en het in deze vorm te publiceren.
Verder wil ik Dr. D. Kroon bedanken voor de stimulerende begeleiding tijdens het werk, en Ir. G. Brouwer voor het kritisch doorlezen van het manuscript.

Levensbericht

Op verzoek van het college van dekanen volgt hier een kort levensbericht van de schrijver.

Hij werd geboren te Tienray op 31 juli 1948. Na het behalen van het H.B.S.-b diploma aan het R.K. Lyceum voor Jongens (Boschveld college) te Venray in 1965, begon hij met de ingenieursstudie scheikundige technologie aan de Technische Hogeschool te Eindhoven. Het afstudeerwerk in de groep Instrumentele Analyse werd gedeeltelijk uitgevoerd in het Instituut voor Instrumentele Analytische Chemie van de Tsechoslowaakse Akademie van Wetenschappen te Brno. In januari 1971 werd het ingenieursexamen afgelegd. Op 1 februari 1971 trad hij in dienst van de N.V. Philips. Hierna werd in de sectie Instrumentele Analyse van de Technische Hogeschool een aanvang gemaakt met het werk dat leidde tot dit proefschrift.

STELLINGEN

M. H. J. van Rijswick

3 december

I

Tegen het gebruik van de Gauss funktie als piekmodel bij curve fitting van niet gescheiden elutiekurven kan als bezwaar worden aangevoerd dat het model te eenvoudig is. Het hanteren van dit argument door degenen die loodlijnmethoden toepassen doet echter hypocriet aan.

J. J. M. Wijtvliet, Proefschrift T.H. Eindhoven, 1972, stelling 1.
B. Weimann, Chromatographia **7**, 472, 1974.
G. Schomburg, F. Weeke, B. Weimann, E. Ziegler, Chromatographia **7**, 477, 1974.

II

Het door Rijks et al. ontwikkelde systeem voor kraanloze serieschakeling van kolommen kan worden gebruikt als pyrolyse-detektor, door voor de tweede kolom het pyrolysesysteem op te nemen. De hoeveelheid informatie en de mogelijkheden voor identifikatie worden hierdoor aanzienlijk vergroot.

J. A. Rijks, J. H. M. van den Berg, J. P. Diependaal, J. Chromatog. **91**, 603, 1974.

III

De interpretatie van massaspectra van steroiden met patroonherkenningsmethoden, zoals gepresenteerd door Varzuma et al., kan worden verbeterd door een oneven aantal (> 1) beslissingsvektoren te trainen, elk vanuit een verschillend startpunt en te klassificeren bij meerderheid van stemmen.

K. Varzuma, H. Rotter, P. Krenmayr, Chromatographia **7**, 522, 1974.
N. J. Nillson, Learning Machines, McGraw-Hill, New York, 1965, Ch. 6.

IV

De konklusie van Matsushima en Enyo, dat de waterstofelektrodereaktie

$$H_2 + 2B \rightleftarrows 2H^+B + 2e$$

in zuur ($B = H_2O$) en alkalisch milieu ($B = OH^-$) volgens hetzelfde reaktiemechanisme verloopt wordt niet overtuigend gesteund door hun experimenten.

T. Matsushima, M. Enyo, Electrochimica Acta **19**, 125, 1974.

V

In het door McVitie en Wilson beschreven "stable marriage problem" is het zinvol om het begrip "bi-stabiele kombinaties" in te voeren voor paren van stabiele kombinaties die bij onderlinge partnerruil eveneens stabiel zijn. Men kan het volgende aantonen: bi-stabiele kombinaties zijn alleen mogelijk als beide leden van de ene soort tevreden zijn met hun huidige partner, terwijl van de andere soort beiden elkaars partner prefereren boven hun huidige partner.

D. G. McVitie, L. B. Wilson, Comm. of the ACM **14**, 484, 1971.

VI

Konsekwente toepassing van de richtlijnen van de Union Internationale des Associations d'Alpinisme voor moeilijkheids-klassifikatie bij rotsklimmen kan een einde maken aan de misvatting dat "artificiel" klimmen een graad moeilijker is dan vrij klimmen.

F. Wiessner, UIAA Schwierigkeitsbewertung und Routenbeschreibung, 1971.

VII

Gezien het belang dat grote groepen van de bevolking bij het inflatieproces hebben, is het gewenst dat er politieke partijen opkomen voor dit belang.

VIII

Voor wetenschappelijke publikaties dient een verjaringstijd van maximaal 10 jaar te worden aangehouden waarna de auteurs niet meer mogen worden aangevallen op hun fouten. Proefschriften moeten in dit opzicht als jeugdzonden worden beschouwd met een versnelde verjaringstijd.

Eindhoven, 3 december 1974 M. H. J. van Rijswick