Ultimate retention time accuracy in computer assisted chromatography

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.
Link to publication

ULTIMATE RETENTION TIME ACCURACY IN COMPUTER ASSISTED CHROMATOGRAPHY

J. J. M. Wijtvliet

ULTIMATE RETENTION TIME ACCURACY IN COMPUTER ASSISTED CHROMATOGRAPHY

door

Josephus Joannes Maria Wijtvliet
geboren te Den Haag.

Dit proefschrift is goedgekeurd door de promotoren

Prof.Dr.Ir. A.I.M. Keulemans

Prof.Dr. G. Guiochon

to Sue

CONTENTS

INTRODUCTION

Only in 1952 (Ref. 1) did gas chromatography present
itself as a powerful tool for the analytical chemist. Since
then the development has rapidly advanced. Presently gas
chromatography is the single most important and widely used
instrumental method of chemical analysis   (Ref. 2).

The development of gas chromatography has been extensive
for columns, injectors, detectors, and other components.
The original columns were all packed. Later came the capill-
ary column with its high resolving power   (Ref. 3). Very
recent is the introduction of the micropacked column with
the combined properties of the packed and the capillary
column   (Ref. 4).

The first gas chromatographs used the automatic titra-
tor (Ref. 1) as detector, which is quite insensitive. Then
came the katharometer (Ref. 5,6), with improved sensitivity
and ease of operation. Later came the broad group of ioni-
zation detectors (Ref. 7). The best known representative
is the flame ionisation detector (Ref. 8,9). The flame
became the most widely used because of its simplicity, high
sensitivity, reliability, and wide dynamic range. Only for
quantitative work is the katharometer superior over the flame
(Ref. 10). Presently there exist a host of detectors which
are all designed for specific purposes. Most of them exhibit
a high specific sensitivity to particular groups of components
or elements. One that should be mentioned is the electron
capture detector (Ref. 11,12). This detector is extremely
sensitive to electro-negative elements and hence to halo-
genated components. Unfortunately this detector has a limited
dynamic range and needs a highly skilled operator.

For a long time sample injection did not get the attention it needed. It was not until the capillary columns were developed that it became apparent that simple direct injection was not good enough. The main problem was how to introduce reproducibly an extremely small amount of more or less volatile matter. Streamsplitters (Ref. 13) brought some relief. And the new method of the falling needle works quite well for low volatile components which are dissolved in a volatile solvent (Ref. 14).

With the development of the gas chromatographic method came the development of GC data handling. For a long time the stripchart recorder was used for recording a chromatogram. Quantifying the results was an entirely manual operation. Retention time measurement was carried out by means of a stopwatch or from measuring the distance between peaks on the chartpaper. Integration was by triangulation, planimetry, cutting out and weighing, or other simple, inaccurate and time consuming methods. The first real improvement was the ball and disc integrator. Yet this method was still limited by the accuracy and two decade dynamic range of the recorder. A real breakthrough came with the introduction of the electronic integrator (Ref. 15). This instrument could accomodate a dynamic range of five decades. The electronic integrator has been continuously improved (Ref. 16). At present it is quite a sophisticated instrument which operates accurately and reliably.

The latest innovation in data handling has been the use of the computer. Two reasons for bringing this powerful tool into the field of gas chromatography are the capability of performing complex arithmetic and the ability to operate on a number of tasks simultaneously. At present the computer may be used for the following functions.

1. Peakdetection, calculating peakarea and peaklocation.
2. Instrument control.
3. Report generating.
4. Handling more gas chromatographs at the same time.

5. Peak deconvolution.

6. Calculating additional GC properties.

7. Calculating peaktop locations with extreme accuracies.

8. Handling the full dynamic range of any chromatograph.

Item 1 can be considered the standard data handling in gas chromatography. When this is carried out properly the computer may achieve the best results. Item 2 implies the possibility of full computer control of the gas chromatographic experiments (Ref. 22,23). This also will lead to more accurate results. Particularly the exclusion of manual operations eliminates the possibilities of human errors.

A lot of effort has been put into peak deconvolution (Ref. 19-23). It was thought that it should be possible to enhance the resolving power of a gas chromatograph by mathematical means. Success was only fair since it appeared not to be possible to describe a chromatographic peak properly by a mathematical model.

Until the application of computers to GC data handling the only information that was extracted from a gas chromatogram was peakarea and peaklocation. Sometimes the peak width was measured and the assymetry of the peak indicated. With proper software a computer may also calculate accurate peaksigmas, skew, kurtosis, excess, resolution and many more GC properties (Ref. 24,25).

This thesis attempts to be a contribution to the practical and theoretical backgrounds of gas chromatographic data handling. Peak detection and the calculation of area, retention time, peaksigma, skew, and kurtosis will be discussed. Accurate retention times will be particulary emphasized. This stems from the fact that the work for this thesis has been done in an laboratory which is particulary interested in component indentification by means of retention times. All problems that are encountered will be approached from the gas chromatographic point of view. This may appear to be a limitation. It should be noted however that extensive parts of this work apply to curve evaluation in other fields as well. No further introduction to gas chromatography will be given here as there already

exists an abundant bibliography on this topic (Ref. 26-30).

The first chapters of this thesis deal with the more theoretical aspects. Baseline determination, peakdetection, smoothing and accurate peaktop allocation are treated in detail. In the next chapters a practical computer program is described based on these theories. Finally some results are shown.

LITERATURE REFERENCES

1. A.T. James and A.J.P. Martin, Biochem. J., 50, 679 (1952).
2. L.S. Ettre, Anal. Chem., 43, no. 14, 20A (1971).
3. M.J.E. Golay, in V.J. Coates, H.J. Noebels, and I.S. Fagerson (eds), Gas chromatography, Academic Press, New York, N.Y., 1958, p.1.
4. C.A. Cramers, J. Rijks, and P. Bocek, J. Chromatogr. 65, 29 (1972).
5. N.H. Ray, J. Appl. Chem. (London), 4, 21, (1954).
6. N.H. Ray, J. Appl. Chem. (London), 4, 82, (1954).
7. J.F.J. Krugers, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1964.
8. I.G. McWilliam and R.A. Dewar, Eur. Symp. GC 2, 142 (1958).
9. L. Ongkiehong, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands. 1960.
10. M. Goedert and G. Guiochon, J. Chrom. Sci., 7, 323 (1969).
11. R.A. Landowne and S.R. Lipsky, Anal. Chem., 34, 727 (1962).
12. J.E. Lovelock, Anal. Chem., 35, 474 (1963).
13. C.A. Cramers, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1967, Chapter 2.
14. P.M.J. van den Berg and Th.P.H. Cox, Chromatographia, 5, 301, (1972).

15. Infotronics Corp., Houston, Tex., "CRS-1 Digital Chromatograph Integrator", Bull. CRS-102 (1962).
16. M. Bargain, Chim.Anal., 52, 164 (1970).
17. R.S. Swingle and L.B. Rogers, Anal.Chem., 43, 810 (1971).
18. J.E. Oberholtzer, Ph.D. Thesis, Purdue University, Lafayett, Ind. 1969.
19. A.A.A.M. Baaten, Graduation report, Eindhoven University of Technology, Eindhoven, The Netherlands, 1969.
20. H.M. Gladney, B.F. Dowden, and J.D. Swalen, Anal. Chem., 41, 883 (1969).
21. A.W. Westerberg, Anal. Chem., 41, 1770 (1969).
22. A.H. Anderson, T.C. Gibb, and A.B. Littlewood, Anal. Chem., 42, 434 (1970).
23. A. Baan, Graduation report, Eindhoven University of Technology, Eindhoven, The Netherlands 1972.
24. O. Grubner, Advances in Chromatography, 6 173 (1968).
25. E. Grushka, M.N. Myers, J.C. Giddings, Anal.Chem., 42, 21 (1970).
26. A.I.M. Keulemans, Gas Chromatography, Reinold, New York, N.Y., 1959.
27. A.B. Littlewood, Gas Chromatography, Academic Press, London, England 1970.
28. L. Szepesy, Gas Chromatography, Akadémiai Kiadó, Budapest, Hungary, 1970.
29. D.A. Leathard and B.C. Shurlock, Identification Techniques in Gas Chromatography, Wiley-Interscience, London, England, 1970.
30. P. Ambrose, Gas Chromatography, Butterworth, London, England, 1971.

CHAPTER 1
BASELINE DETERMINATION

1.1 INTRODUCTION

The baseline can be considered the one parameter that is most difficult to establish. Many techniques have been tried in this field (Ref. 1.1, 1.2). But inherent to its nature the baseline never can be determined with absolute confidence (Ref. 1.3). As it is, the baseline is hidden away under peaks, buried in noise, and obscured by drift. The accuracy of the baseline limits the final accuracy of virtually all other parameters (perhaps with the exception of retention data) to be established (Ref. 1.4). There are various ways to approach this problem. The baseline can be considered a straight or a curved line. The baseline can be established for the whole chromatogram at once or per peak or peakgroup. The actual influence of the baseline on the final results will be shown in Chapter 7.

1.2 STRAIGHT BASELINE

The following techniques will be considered:
1. Baseline over the whole chromatogram,
   a. Horizontal line
   b. Slanted line.
2. Baseline per peak or peakgroup,
   a. Horizontal line
   b. Slanted line.

1.2.1 Baseline over the whole chromatogram.

1.2.1.a Horizontal baseline.

Defining the baseline as a horizontal straight line through the whole chromatogram is the most simple approach. It is useful in those cases where the baseline is obviously straight and horizontal, as is the case in "clean" chromatograms with well resolved peaks. It is also useful when only retention times have to be calculated. Although it should be mentioned here that an inaccurate baseline affects the center of gravity. With an uncertainty in the baseline, only the peaktop can be determined accurately. This will be shown extensively in Chapter 7.

A horizontal baseline may also be selected if computer time is at a prime. It is also useful to choose the horizontal baseline method in such cases where the baseline is so obscured that no other method would be more reliable.

As simple as the determination of a straight horizontal line may seem, there are various methods:

- Average below average technique.

This technique will also appear when a threshold level has to be calculated. Essentially, first the average of all datapoints is calculated. Then the average of all datapoints with a value below that of the previous value. This process is then repeated one or two more times. The final average is said to be the baseline level. The whole process will be extensively discussed in Section 4 at the end of this chapter.

- Initial baseline extrapolation.

The initial baseline is the average of all datapoints from the start of a run until the start of the first peak. Or as an alternative, a predetermined number of datapoints is used. This initial baseline is then extrapolated horizontally.

–   Begin and end connection.

The baseline is here defined as the average of a given number of datapoints at the beginning and at the end of the chromatogram.

–   Connect baseline pieces.

Here it is first necessary to determine the peak boundaries. When this is done, all datapoints on the remaining pieces of baseline are averaged; this average is then the baseline.

1.2.1.b Slanted baseline.

A slanted baseline generally gives an improvement over a horizontal line. It is particularly good for chromatograms that exhibit a constant drift. Yet these cases are poor gas chromatography. A slanted baseline is also an improvement with respect to a horizontal line in the case of programmed gas chromatography. The programming may be temperature or flow. Yet in this case the true baseline is curved rather than straight. But if for some reason it is not feasable to establish a curved baseline then the slanted one is the next best.

There are various ways to determine a slanted baseline. These are similar to those techniques for a horizontal line:

–   Average below average technique.

First the average of all datapoints is calculated. Then a straight line is fitted through all points below the average. After this a straight line is fitted through all points below the previous fitted line. If necessary this may be repeated once again.

–   Begin and end connection.

A given number of datapoints at the beginning of the run are averaged and a given number at the end. A straight line that connects these two averages is said to be the baseline.

- Connect baseline pieces.

For this method peaklocations have to be established first. A straight line is then fitted through the remaining pieces of baseline.

1.2.2 Baseline per peak or peakgroup.

1.2.2.a Horizontal baseline.

To define a horizontal baseline per peak or peakgroup is rather simple. The begin and end points are averaged and this average represents the baseline level. It is also possible to average some points before the beginning and after the end of the peak.

1.2.2.b Slanted baseline.

A slanted baseline per peak or peakgroup is determined in about the same way as a horizontal line. Now the begin and end are connected by a straight line (Ref. 1.2.). Or a line may be fitted through some points before the start and after the end of the peak.

1.3 CURVED BASELINE

Just as for a straight baseline two different approaches may be taken. Usually a second, third, or fourth order function describes the baseline with sufficient accuracy. Or, in other words, using a higher order function hardly ever increases the accuracy of the baseline.

1.3.1 Baseline over the whole chromatogram.

The methods that can be applied here are applied in the same way as for a slanted straight baseline. Essentially a slanted line is a function of the first order.

It can be done in one of the following two ways:

- Average below average technique.

Calculate the average of all datapoints. Fit a curve through all points below this average. Fit a curve through all points below the previous curve. This may be repeated once more. It may also be useful to start with a low order function and to increase it with every step.

- Connect baseline pieces.

A curve is fitted through all points that are known to be part of the baseline.

1.3.2 Baseline per peak or peakgroup.

Here there is only one possibility. A curve is fitted through some points before the peakstart and after the peakend (Ref. 1.1, 1.5).

1.4 AVERAGE BELOW AVERAGE TECHNIQUE

1.4.1 Using all data.

The technique can best be explained in a step by step fashion. Refer also to Figure 1.1.
1. Calculate the average of all datapoints.
2. The result is called: average 1.
3. Calculate the average of all datapoints that have a value below that of average 1.
4. The result is called: average 2.
5. Calculate the average of all datapoints that have a value below that of average 2.
6. The result is called: average 3.
7. Repeat step 5 and 6 and obtain average 4 from all points below average 3.
8. Repeat step 7 as often as required, every time increasing the "average number".
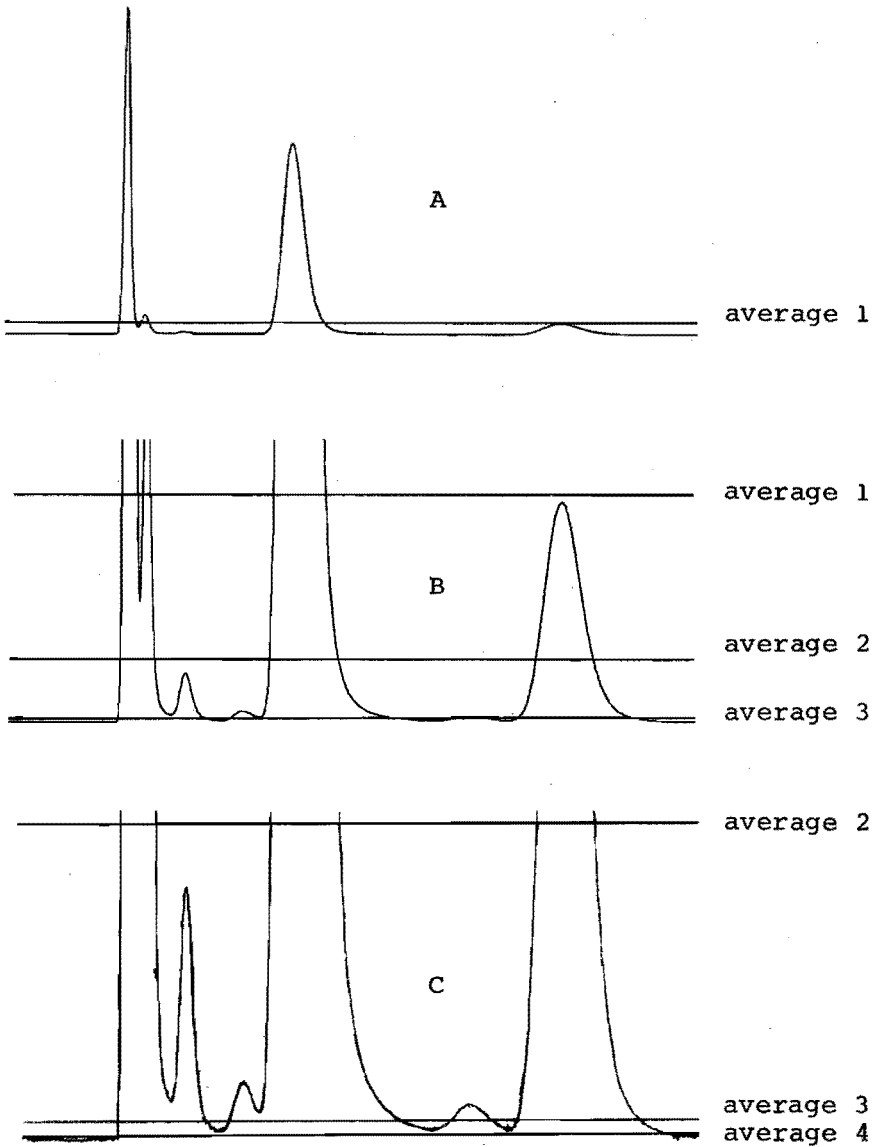
Figure 1.1. Ilustration of the average below average technique.

A: Chromatogram drawn without scale expansion.

B: 20 x expanded scale.

C: 100 x expanded scale.

Customarily average 2 is used as threshold level, and
average 3 as baseline. Although it may be at times better
to use average 4 as baseline. A criterion for this might
be the percentage of datapoints that is below average 3
with respect to the percentage that is below average 4.
Another method is described in Section 5.2.3. This method
proved in practice to be the more reliable one.

1.4.2. Using part of the data.

For the technique as outlined above it is not neces-
sary to use all datapoints. It is in most cases quite
possible to only take every second, third or even tenth
datapoint. As long as the number of datapoints is suffi-
ciently large, the result is hardly different. In practice
every fifth datapoint with a minimum of 250 points gave
results that were virtually identical to the results that
were obtained when using all datapoints.

1.5.   LITERATURE REFERENCES

1.1.   A. Baan', Graduation report, Eindhoven University
        of Technology, Eindhoven, Netherlands, 1972, chapter 3.
1.2.   R.D. McCullough, J. Gas Chromatography, 5, 635 (1967).
1.3.   J. Novák, J. Gelbicóvá-Ruzicková, S. Wicar and J.Janak,
        Anal.Chem., 43, 1996 (1971).
1.4.   P.C. Kelly and W.E. Harris, Anal. Chem., 43, 1170 (1971).
1.5.   J.W. Frazer, L.R. Carlson, A.M. Kray, M.R. Bertoglio,
        and S.P. Perone, Anal. Chem., 43 1479 (1971).

CHAPTER 2
PEAK DETECTION

2.1 INTRODUCTION

This chapter will deal with the detection of start-
points, endpoints, tops, valleys, and shoulders of peaks
or peakgroups; although the utmost accuracy will not be
pursued here, particularly not for the peaktops.

Together with baseline determination, the peakdetection
is the most important task for every gas chromatography
program. Once this task is performed, the gas chromatogram
is fully described and all subsequent computations are
only done to make the representation of the results clear.
It is somewhat debatable if separation of fused peaks, e.g.
by means of curve fitting, could also be called a major task.
Essentially this is a task of the gas chromatograph and
not of the computer, even though the computer may enhance
the separation of peaks (Ref. 2.1).

A major problem that is encountered in peakdetection
is pushing the detection limit to as low a level as possible.
This means detecting very small peaks, or more generally
detecting the peakboundaries as carefully as possible.
Pushing the detection limits involves fighting noise and
drift. Noise (most particularly: spikes) will obscure small
sharp peaks. Drift will obscure broad peaks. Drift will
also make it difficult to obtain the utmost accuracy in
detecting the peakboundaries. This is, of course, provided
that it is necessary to obtain the utmost accuracy. Since
any considerable increase in accuracy will have to be paid
for, it is desirable that the required accuracy can be
specified to some extent. In other words: if one is only

interested in a somewhat rough estimation of the chromato-
gram one should not spend much time in obtaining the best
possible detection levels. On the other hand if it hardly
costs anything to obtain high accuracies one might as well
always do so. Flexibility has its price as well.

2.1.1 Assortment of techniques

Peakdetection can basically be done in two different
ways:
1. Using threshold levels.
2. Using derivatives of the signal.
Then still, various methods can be distinguished:
1. Fixed or dynamic criteria.
2. A rough estimation followed by a routine to find the
   accurate information or the direct approach.
These different techniques can be applied in virtually
every possible combination. Although some are more useful
than others; the merits of these various approaches will
be discussed here.

2.2   THRESHOLD LEVELS OR DERIVATIVES

2.2.1 Threshold levels

Here threshold level means a certain level above the
baseline. Any part of the chromatogram above this level is
said to belong to a peak. As a result, single peaks can not
be distinguished from groups of peaks. This is pictured in
Figure 2.1.

Since a threshold level is essentially a level above
the baseline, its usefulness is also tied to the accuracy
of the baseline. This can be demonstrated best in those
cases where the baseline jumps or drifts unexpectedly.
Refer to Figure 2.2 and Figure 2.3.
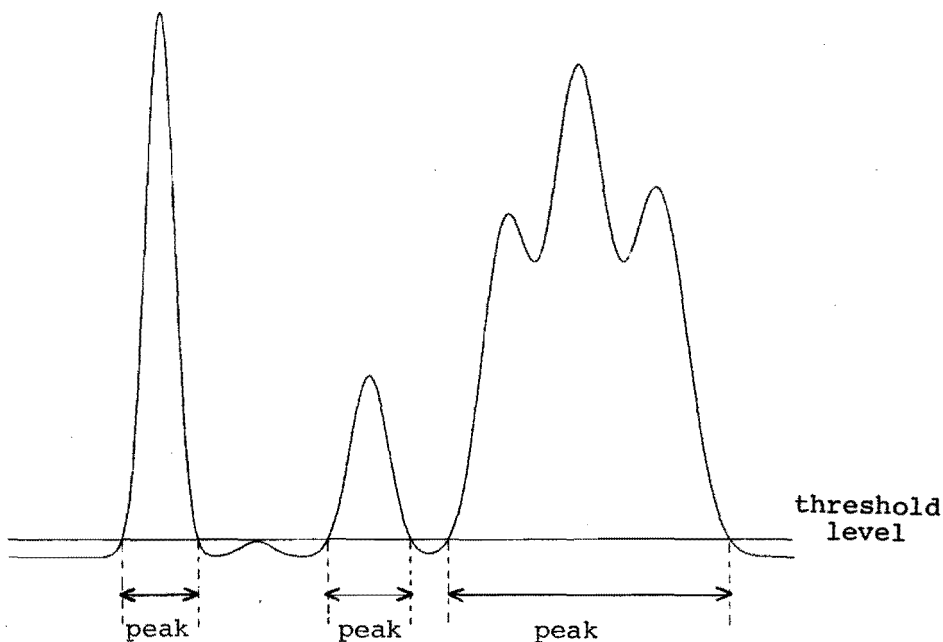
Figure 2.1.   Peak detection through a threshold level.
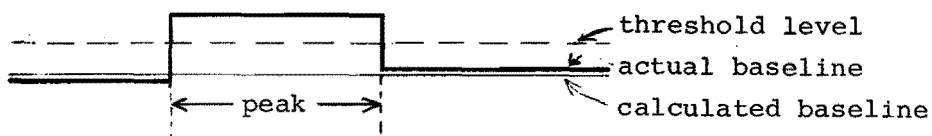
Figure 2.2.   The threshold level method recognizes a
baseline jump as a peak.
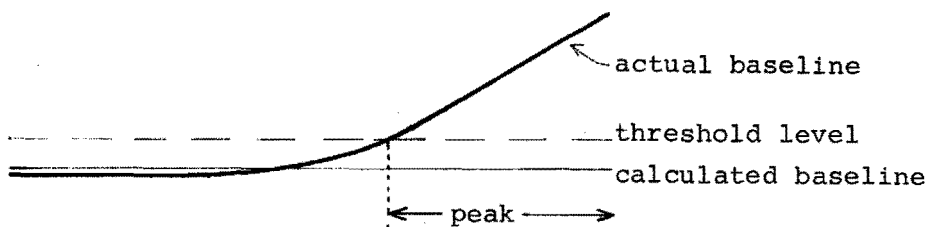
Figure 2.3.   The threshold level method recognizes base-
line drift as a peak.

2.2.2 Derivatives

The use of derivatives is an alternative approach. Actually the use of derivatives also implies the use of threshold levels. But generally it is now a dual level above and below zero, and of course taken with respect to the derivative used. In Figure 2.4 the derivatives of a Gaussian peak are given.

original peak

first derivative

second derivative

third derivative
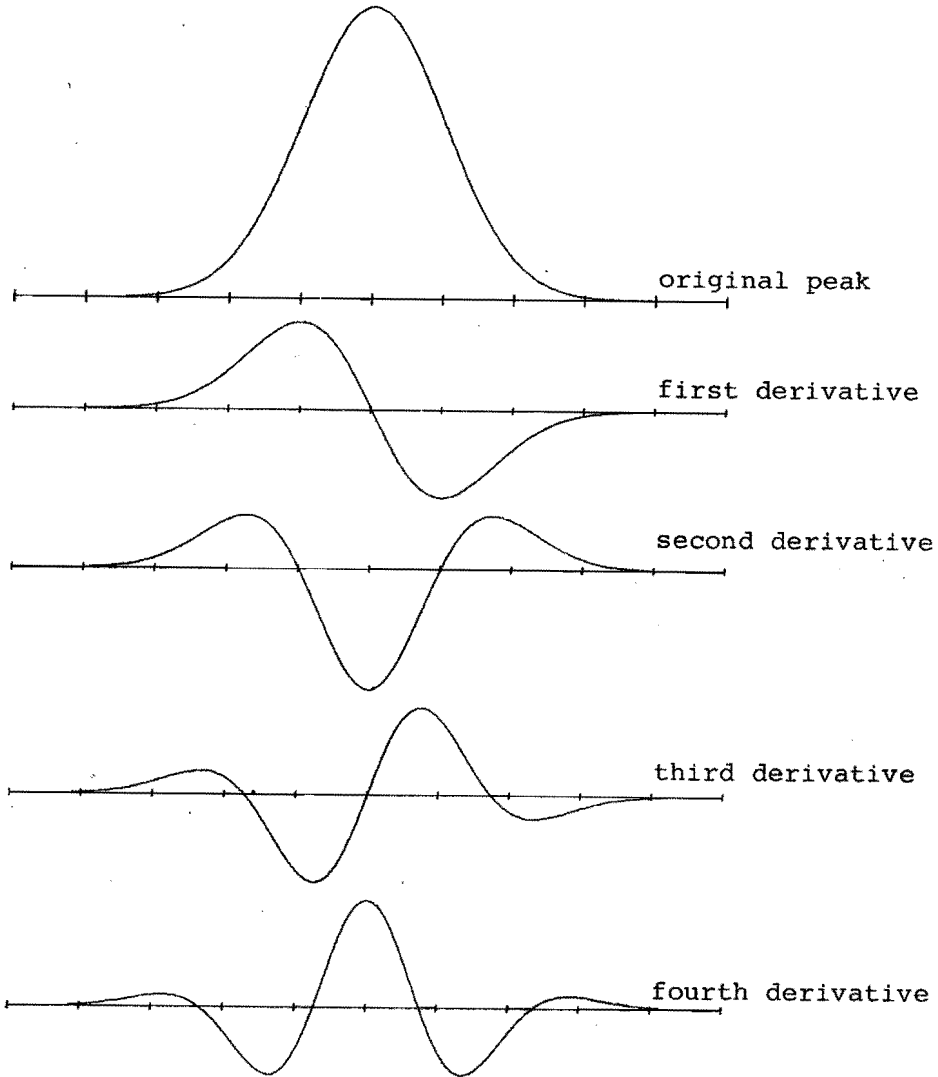
fourth derivative

Figure 2.4. A Gaussian peak and its derivatives.
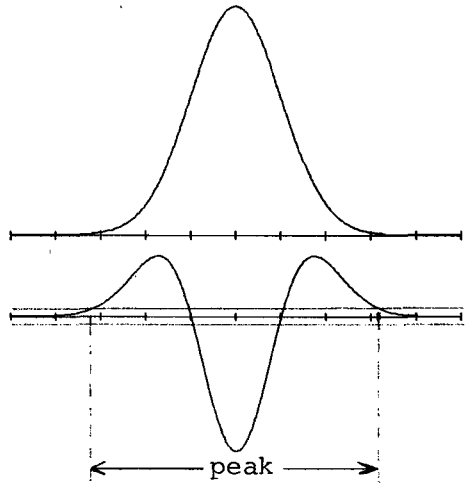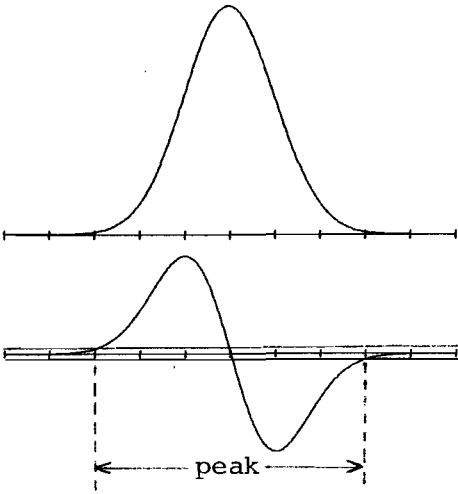
Figure 2.5.a. Peak detection through the first derivative.

Figure 2.5.b. Peak detection through the second derivative.

Actual baseline

First derivative
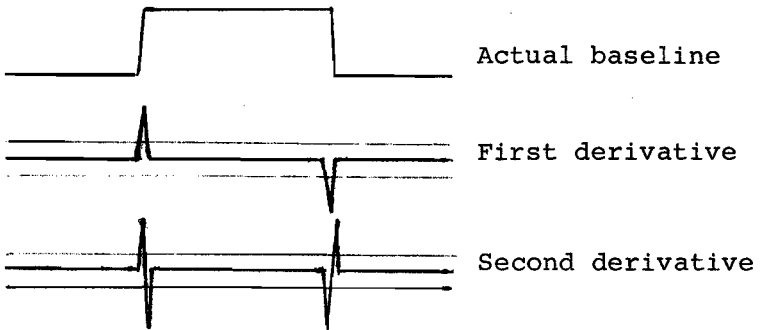
Second derivative

Figure 2.6. Effect of a baseline jump on the peak detection through derivatives.

Actual baseline

First derivative
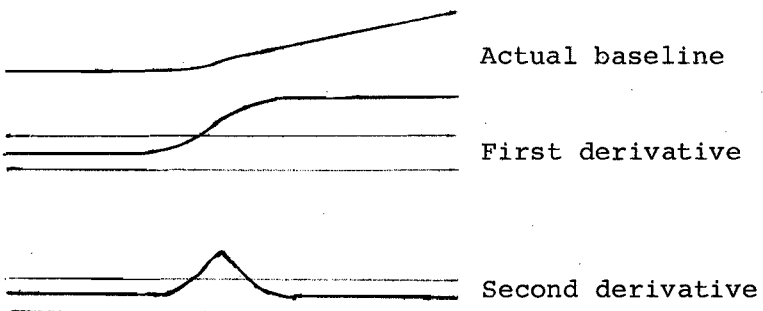
Second derivative

Figure 2.7. The effect of baseline drift on the first and second derivative.

The most commonly used derivatives are the first and second. Their use is displayed in Figures 2.5.a and b. The advantages over the simple threshold method are obvious. A jump is hardly noticed. Refer to Figure 2.6. A jump in the baseline can be very effectively rejected as a peak, since the peak detection logic requires a certain pattern in the fluctuations of the derivatives (Ref. 2.2).
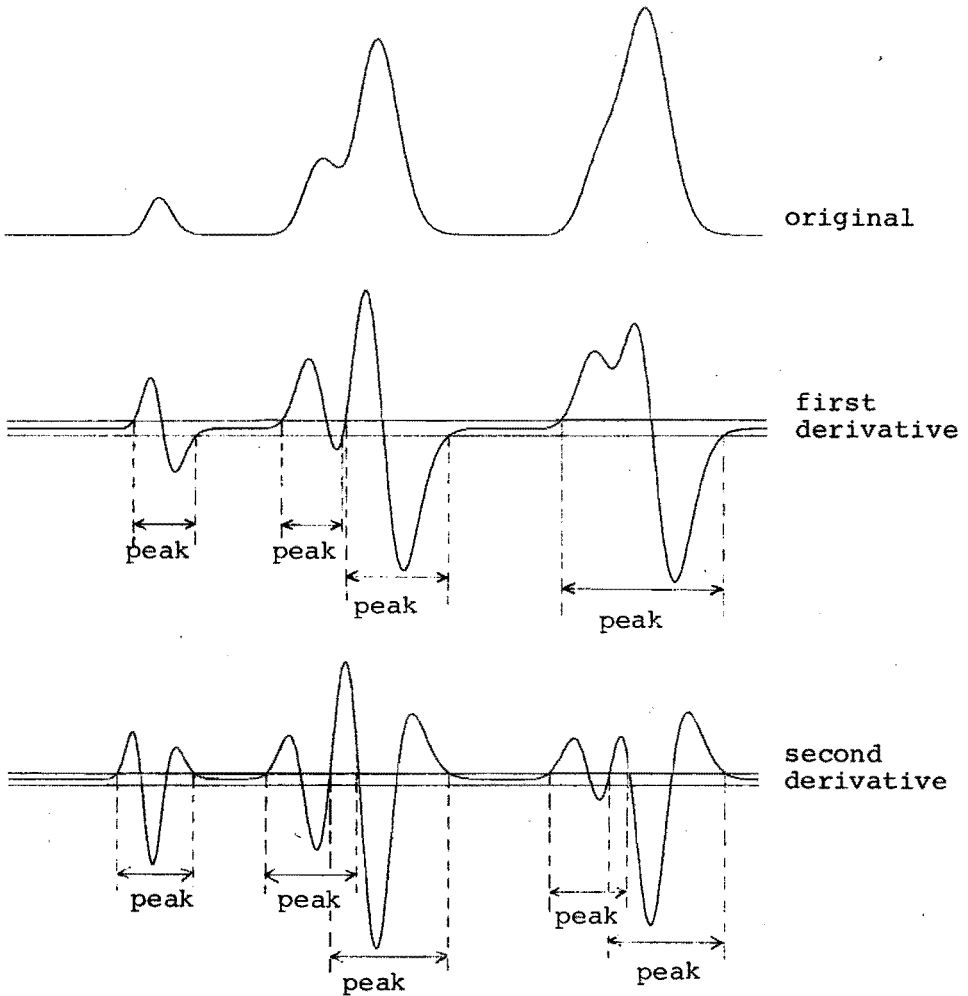
Figure 2.8.   Detecting unresolved peaks by means of derivatives.

Drift may not be handled so well by the first deriva-
tive, but there are no problems for the second derivative.
This is shown in Figure 2.7. Derivatives also work well on
groups of unresolved peaks as is pictured in Figure 2.8.
Thus seemingly the second derivative works best.

The major set back of the use of derivatives is their
sensitivity to noise, as can be shown clearly from Figure
2.9, where a noisy Gaussian peak is shown with the deriv-
atives. This implies that low and wide peaks are particu-
larly difficult to detect when using derivatives. Another
disadvantage is that the use of derivatives tends to be
essentially more time consuming in computer programs.
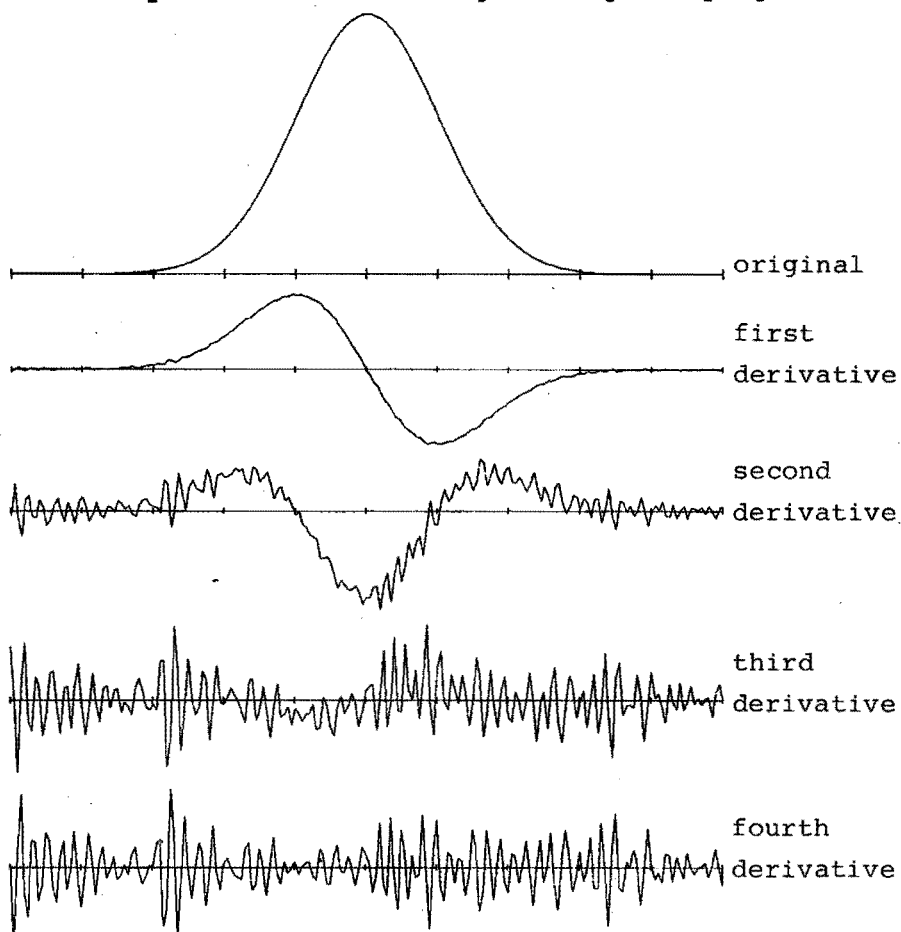
Figure 2.9. Gaussian peak with random noise and the
associated derivatives. Signal to noise ratio: 5000.

2.2.3 Threshold level and derivatives combined.

A combination of both methods may be able to solve many problems. It would be very possible to first detect peaks and peakgroups by means of a threshold level, since this method tends to overlook only the very smallest peaks. Then the derivative technique may be used to divide peakgroups into individual peaks and reject spurious peaks.

2.3 ESTABLISHING A THRESHOLD LEVEL

2.3.1 Possible techniques.

There are three ways to establish a threshold level for peak detection.

1. Average below average.
   The threshold level is the average of all datapoints that are below the average level. The technique is already extensively described in 1.4.

2. Noise dependent.
   First the noise on the baseline is calculated over the initial baseline. This noise level is then multiplied by a certain factor (usually between 2 and 10) to obtain the threshold level.

3. External.
   The threshold level is given a fixed value by the program-user. This value may be an absolute figure or relative, i.e. as a fraction of the signal span.

Method 2 makes two assumptions. First, that the initial baseline is long enough to obtain a reliable noise-figure. Second, that the noise level is constant over the whole chromatogram.

Method 3 is a typical "manual" method. It requires human judgement before processing the GC-data. This is not desirable because it may lead to errors very easily. Method 1 has been chosen since it does not have the disadvantages of the other two. Extensive use has also proven the reliability of the method. It was found only desirable to have method 3 available as an option to handle special cases.

2.4 APPLICATION OF A THRESHOLD LEVEL

2.4.1 Detecting peaks, or peakgroups.

Basically every point above the threshold level is assumed to be part of the peak. Refer to Figure 2.1. Only one exeption will be made. If a peak apparently consists of no more than two datapoints then this peak will be considered a spike and be rejected.

2.4.2 Detecting tops and valleys.

Essentially every maximum in a peakgroup will be called a peaktop, and every minimum, a peakvalley. Or in other words: every time the sign of the first derivative goes from positive to negative indicates a top, and the valleys are indicated by a change from negative to positive.

However, this technique does not work well because now even the smallest noise spike will be recognized as a peak. This problem can be avoided through the introduction of the so called "followfactor". In order to use this it is assumed that the first derivative is obtained by subtracting subsequent datapoints. Every time the first derivative changes sign, this has to be confirmed before it is recognized as a real top or valley. This confirmation takes place if after a change in sign, the sign stays the same for at least n points, where n is called the followfactor.
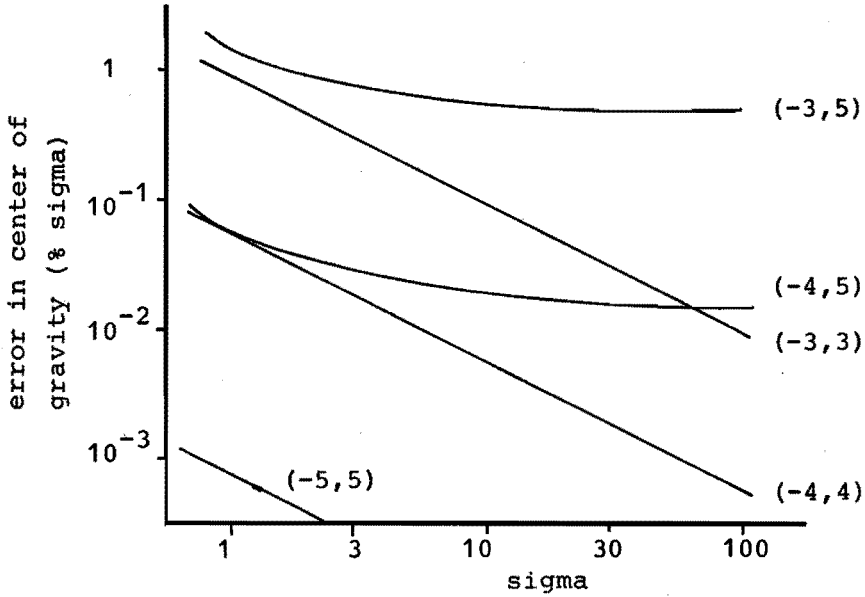
Figure 2.10. The error in the center of gravity as a function of the peak boundaries and the peaksigma. The peak boundaries are given in multiples of the peaksigma.
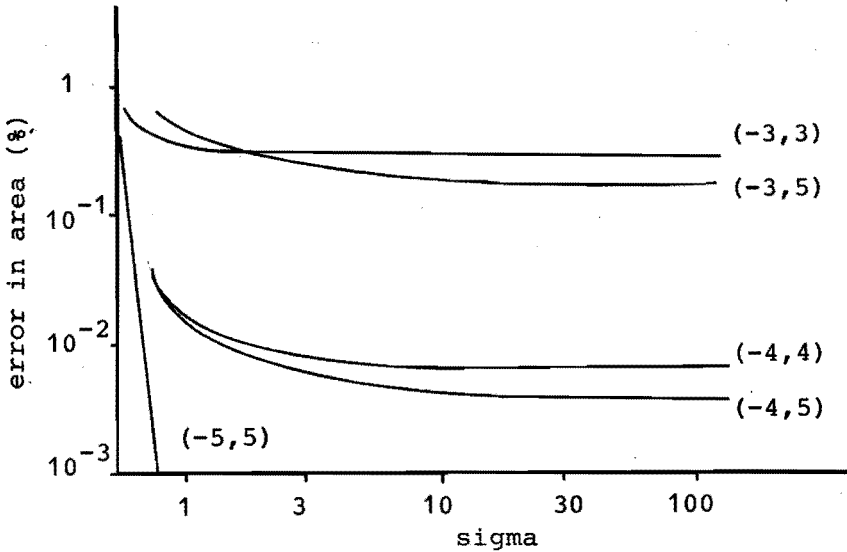
Figure 2.11. The influence of the peak boundaries and the peaksigma on the area determination.
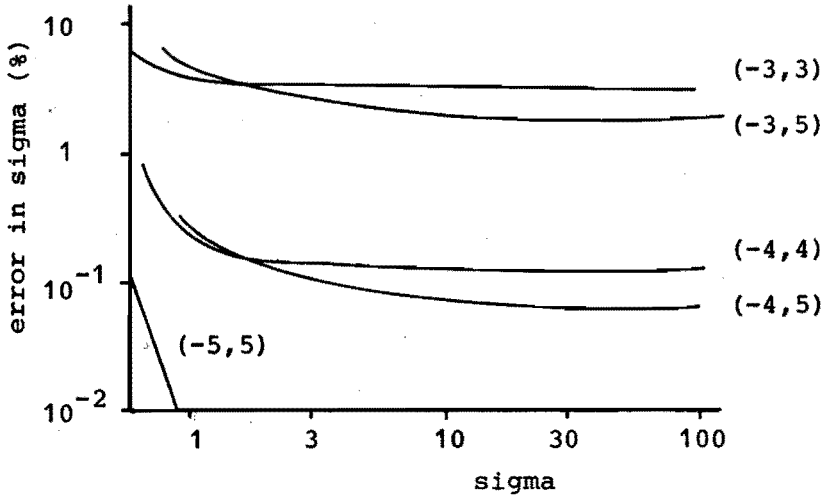
Figure 2.12. Error in the sigma as a function of the peak boundaries and the peaksigma.
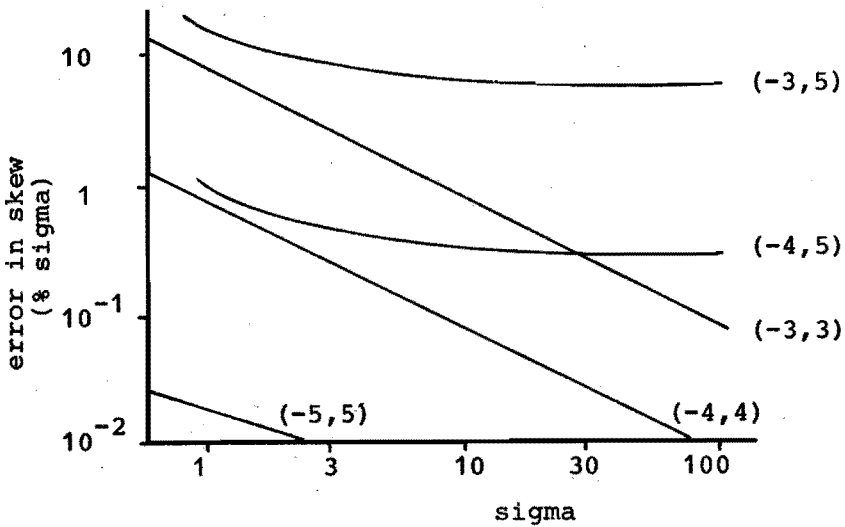
Figure 2.13. Error in the skew as a function of the peak boundaries and the peaksigma.
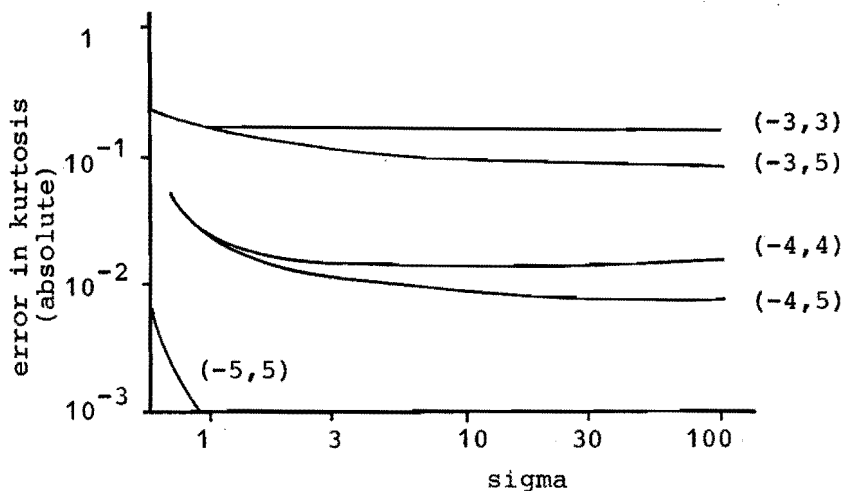
Figure 2.14. Error in the kurtosis as a function of the peak boundaries and the peaksigma.

2.5 EFFECTS OF INACCURATE PEAK BOUNDARIES

By inaccurate peak boundaries is meant peak boundaries that are too close to the top. The effect of this has been tested on a computer generated peak (Ref. 2.3). The results are shown in Figures 2.10 through 2.14. In these graphs the peak boundaries are expressed as a number times the peak sigma.

From these graphs it becomes clear that inaccurate peak boundaries deteriorate only slightly the accuracy of the final results. It is interesting to note that the center of gravity and the skew are only affected if the peak boundaries are asymmetrical.

2.6 LITERATURE REFERENCES

2.1 S.M. Roberts, Anal. Chem., 44, 502 (1972).

2.2 A. Baan, Graduation report, Eindhoven University of Technology, Eindhoven, The Netherlands, 1972, Chapter 1.

2.3 S.N. Chesler and S.P. Cram, Anal. Chem., 43, 1922 (1971).

CHAPTER 3
PEAKTOP LOCATION

3.1 INTRODUCTION

Chapter 2 explained how peak detection can be done.
Peak detection implied the detection of peaktops. Detec-
tion of a peaktop is defined here as a fairly good first
estimate of the location of a peaktop; usually the data-
point closest to the actual top. Yet peakdetection is not
considered the same as the calculation of an accurate peak-
top location. One method to calculate the accurate peaktop
location is fitting a curve through the top part of the
peak and then calculating the position of the top of this
curve. Once the top part of the peak can be described by
a well-defined function then it becomes an easy mathemati-
cal problem to find the maximum of this function.

3.2 CURVE FITTING THE PEAKTOP

3.2.1 Gaussian

Usually a gas chromatographic peak is assumed to ap-
proach a Gaussian peak. Therefore it would seem best to
fit a Gaussian curve to the peaktop. Unfortunately the
deviations from a Gaussian curve can be quite extensive
and are not well defined. It should also be mentioned that
fitting a Guassian curve is a slow iterative process. It
will be shown that a good alternative is available. There-
fore the method of fitting a Gaussian function has been
abandoned.

3.2.2 Parabola

The top of a Gaussian peak closely resembles a para-
bola. A next best and easy to use approach is fitting a
parabola to the top. In this case it is possible to cal-
culate the peaktop straight forward. The formula for this
can be derived as follows.

Assume the parabolic function as:

$$y = ax^2 + bx + c \qquad (3.1)$$

The first derivative is:

$$y' = 2ax + b \qquad (3.2)$$

The second derivative is:

$$y'' = 2a \qquad (3.3)$$

In the maximum of the parabolic function:

$$y'_m = 0 = 2ax_m + b \qquad (3.4)$$

or:

$$x_m = -\frac{b}{2a} \qquad (3.5)$$

For a parabola fitted through a number of datapoints the
y, y', and y'' can be calculated for any value of x, using
tables as published by Savitsky and Golay (Ref. 3.1). For
a point $x = t$, where t is near the top one can write:

$$y'_t = 2at + b \qquad (3.6)$$

$$y''_t = 2a \qquad (3.7)$$

From Equations 3.6 and 3.7 can be derived:

$$b = y'_t - y''_t \cdot t \qquad (3.8)$$

If Equations 3.7 and 3.8 are substituted in Equation 3.5,
then:

$$x_m = -\frac{y'_t - y''_t \cdot t}{y''_t} = t - \frac{y'_t}{y''_t} \qquad (3.9)$$
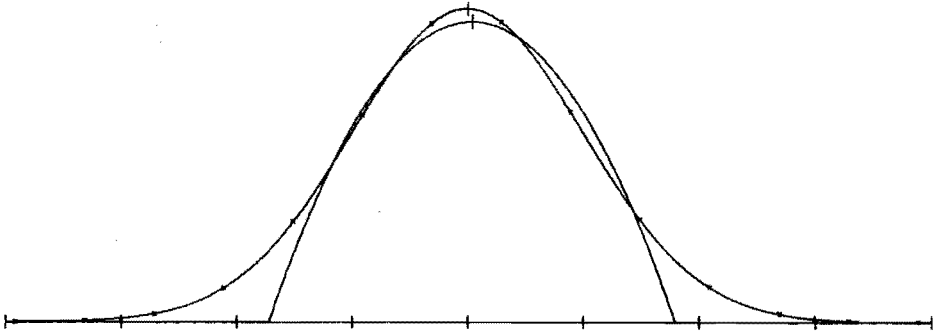
Figure 3.1   Gaussian curve with a parabola through the
             top five points.

The values of y' and y" are obtained using a number of
datapoints to the left and right of t. The total number
of datapoints, including t itself, is called n. It is now
said that the maximum is determined using a parabola fitted
through n datapoints. In Figure 3.1, where n = 5, the true
tops of the Gaussian and the parabola are indicated. Since
the number of datapoints to the left and to the right of
t are always taken equal it is clear that n always will
be an odd number.

3.3 SYSTEMATIC ERRORS

    If the whole process is to work well as outlined here
then t has to be taken as close to the true maximum (x_m)
as possible. Also n should be small compared to the number
of datapoints that describe the peak. The last limitation
arises from the fact that only the top part of a Gaussian
curve closely resembles a parabola. Since curve-fitting
techniques are applied, it is obvious that the calculated
result will differ from the actual value. It is therefore
important to know how big these discrepancies are, and, if
possible, correct for them.

3.3.1 Estimation of errors

The parabola fitting method is first applied to a true Gaussian curve without any noise. This way the basic performance can be tested.

When the true top of the Gaussian peak coincides with a datapoint, then there can be no discrepancies because of symmetry. These discrepancies occur when the true top does not coincide with a datapoint.

Let s be the number of datapoints over one sigmawidth of the Gaussian peak. The difference between the true top and the closest datapoint will here be called Δd. The discrepancy between the true top location and the calculated one will be called Δc. The Δc and Δd are expressed as parts of the basic data interval, u. Refer to Figure 3.2. The Δc is a function of Δd, s, and n. The influence of n is shown in Figure 3.3 for s = 5 and various values of n. The influence of s is demonstrated by Figure 3.4.
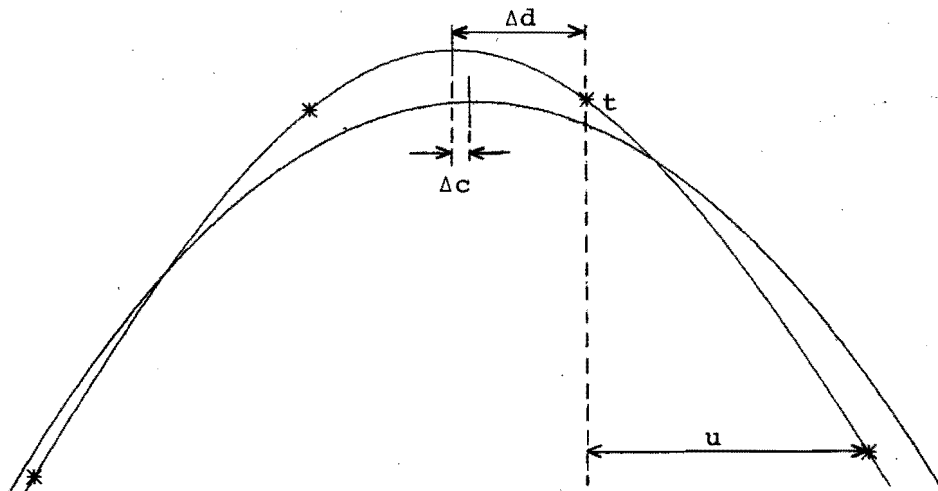
Figure 3.2. Enlargement of the top part of Figure 3.1, n = 5, s = 1.6, Δd = 0.49, Δc = 0.065.

The average error $\overline{\Delta c}$ can be calculated as follows:

$$\overline{\Delta c} = \sqrt{\frac{\sum\limits_{k=1}^{100} \Delta c_k^2}{99}} \qquad (3.10)$$

Where Δc is calculated for 100 different Δd-value's;
the Δd-values being equally spaced over the range 0 to 1.
Refer to Figure 3.5.
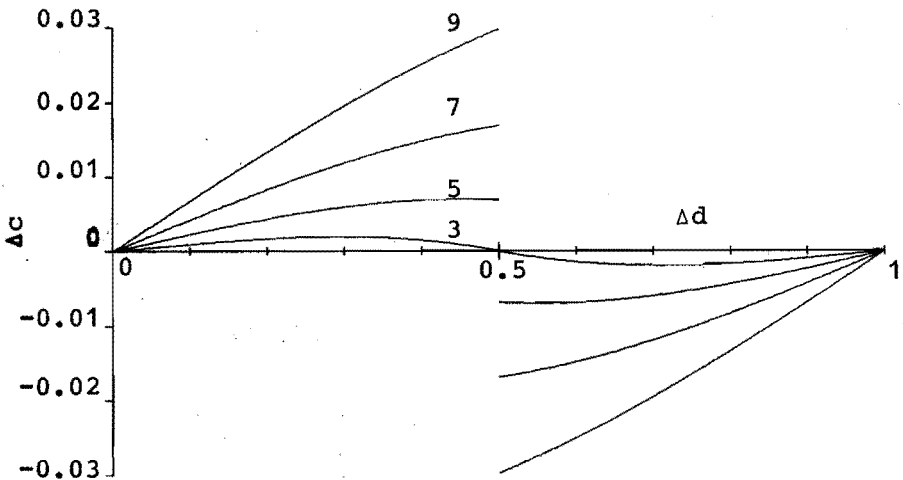
Figure 3.3 Difference (Δc) in between the true top and
the calculated top, for sigma = 5, and n as indicated.
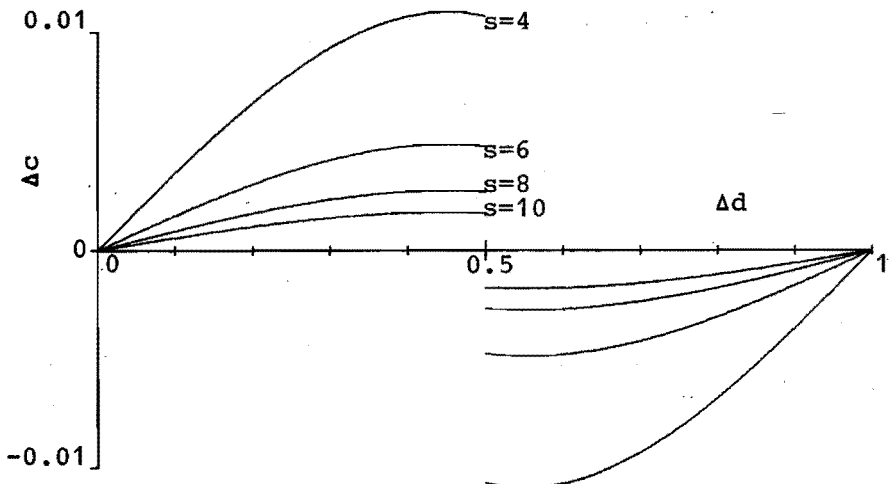
Figure 3.4 Influence of Δd and sigma on Δc, for a 5-point
parabolic fit.
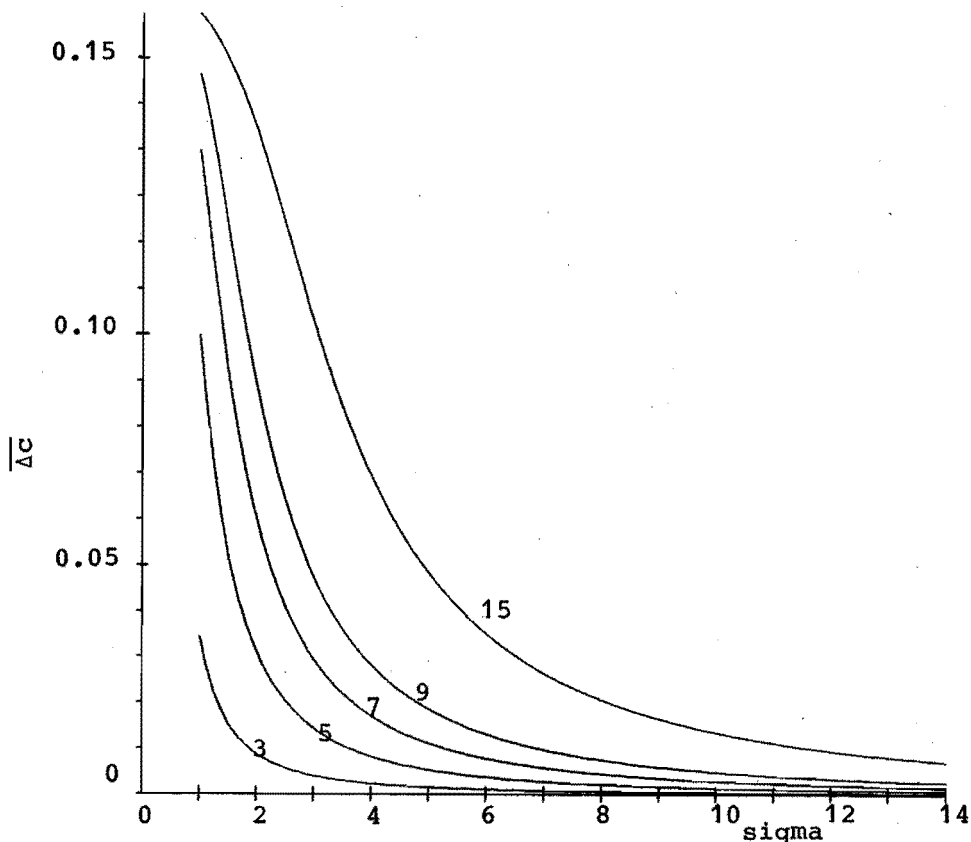
Figure 3.5   The influence of sigma on $\overline{\Delta c}$, at values of n
as indicated in the plot.

3.3.2 Reduction of systematic errors.

The Figures 3.3 through 3.5 show that Δc is quite
small, yet the curves for n = 5 are somewhat unusual in
their discontinuity at Δd = 0.5. In order to understand
this better it is good to consider Equation 3.9 again.
Assume t being the datapoint closest to the actual top, then
the factor y_t' / y_t'' can be considered a correction factor.
This correction factor should be equal to the distance
from t to the actual top. If this factor is ignored then
Equation 3.9 becomes:

$$x_m = t \qquad (3.11)$$

The error curve in this case will be as in Figure 3.6.
Comparing Figure 3.3 with Figure 3.6 shows clearly that
for $n \geqslant 5$ the correction factor is not large enough. This
can be improved by adding a constant; Equation 3.9 then
becomes:

$$x_m = t - (1 + p) \frac{y'_t}{y''_t} \tag{3.12}$$

It is shown that Δc and hence p is dependent on n and s.
In fact it looks very much like $p = f(1/s^2)$, except for a
high value of n combined with a low value of s. This could
transform Equation 3.12 into:

$$x_m = t - (1 + q/s^2) \frac{y'_t}{y''_t} \tag{3.13}$$

This new equation for x_m gives rise to a new Δc and hence
to a new $\overline{\Delta c}$. This new $\overline{\Delta c}$ is dependent on q. This dependence
is shown in Figure 3.7 for various values of n and a peak-
sigma of eight. Figure 3.7 shows that the $\overline{\Delta c}$ may be strongly
improved with a correct choice of q.
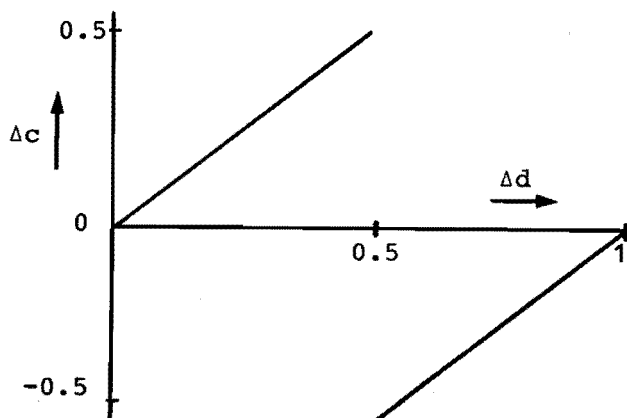
Figure 3.6 Error curve for no fitting at all (n = 1).

One should bear in mind that a q-value of zero represents the results of Equation 3.9, where no correction factor was used. Figure 3.8 shows that there is only a slight dependence left on s. Although the sensitivity to a slight change in q is higher for lower s-value.

A list of experimentally found q factors is given in Table 3.1. Applying these q factors to the curves of

Table 3.1

Correction factor q
as a function of n

n	q
3	0.10
5	0.47
7	0,98
9	1.68
11	2.55
13	3.61
15	4.86
17	6.30
19	7.93
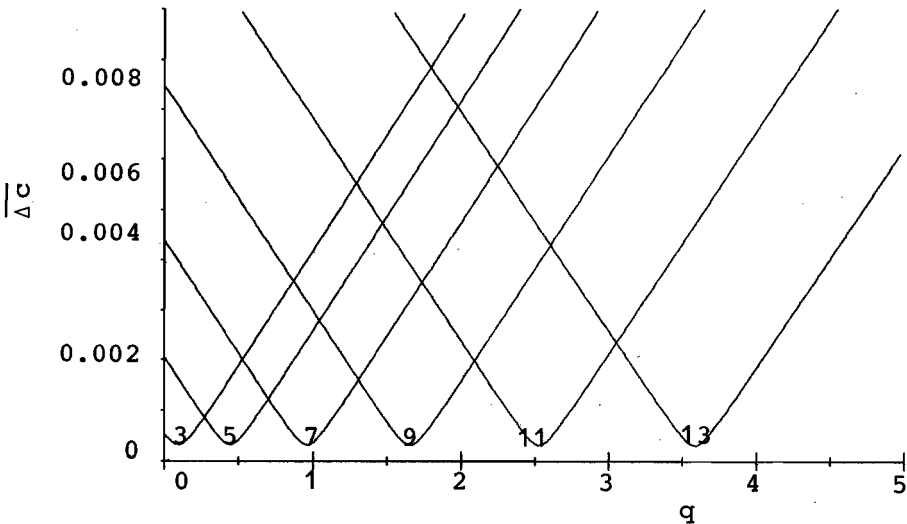
Figure 3.7. Influence of q on $\overline{\Delta c}$ at sigma = 8 and for n-values as indicated in the plot.
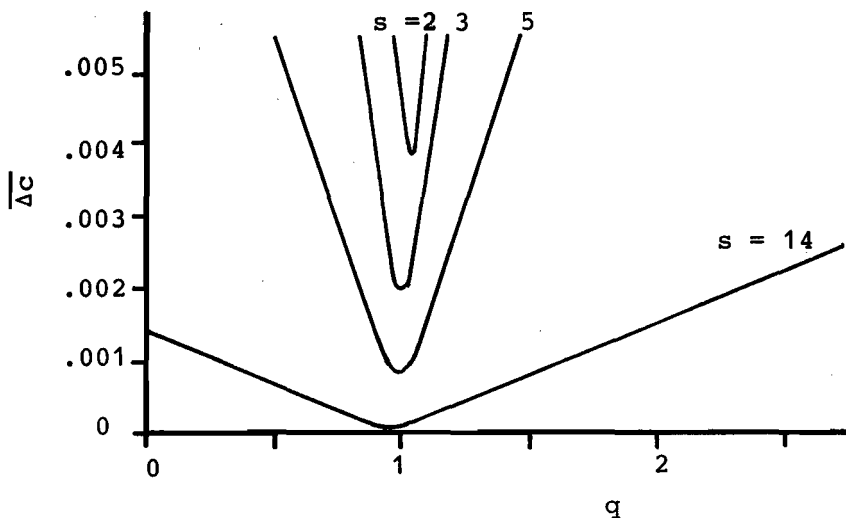
Figure 3.8 Influence of sigma and correction factor q on the average error $\overline{\Delta c}$, for a 7-point fit.
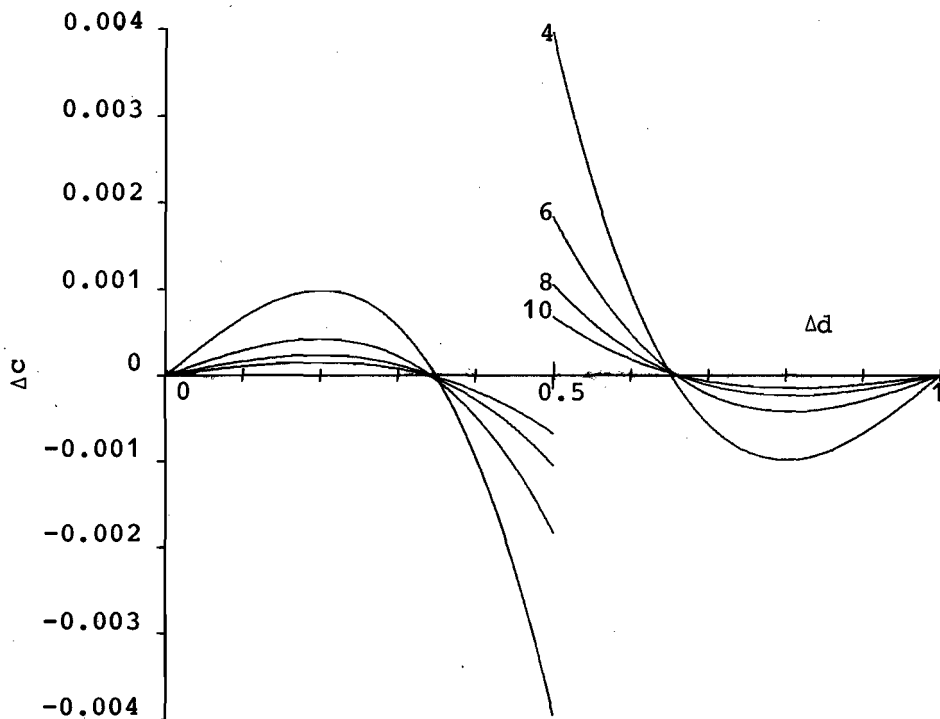
Figure 3.9. Δc as a function of Δd, after applying the appropriate q factor; $n = 5$, and the s-values are indicated in the plot.

Figure 3.4 and Figure 3.5 yields Figure 3.9 and Figure 3.10.
Both Figures 3.9 and 3.10 show the improvement in Δc and
$\overline{\Delta c}$. For $\overline{\Delta c}$ this improvement may not seem so strong for the
combination of high n and low s-values. However one should
realize that these combinations represent unrealistic cases.
If the ratio of n and s is less than five then the use of
the correction factor q yields an improvement in $\overline{\Delta c}$ from
one to two orders of magnitude.
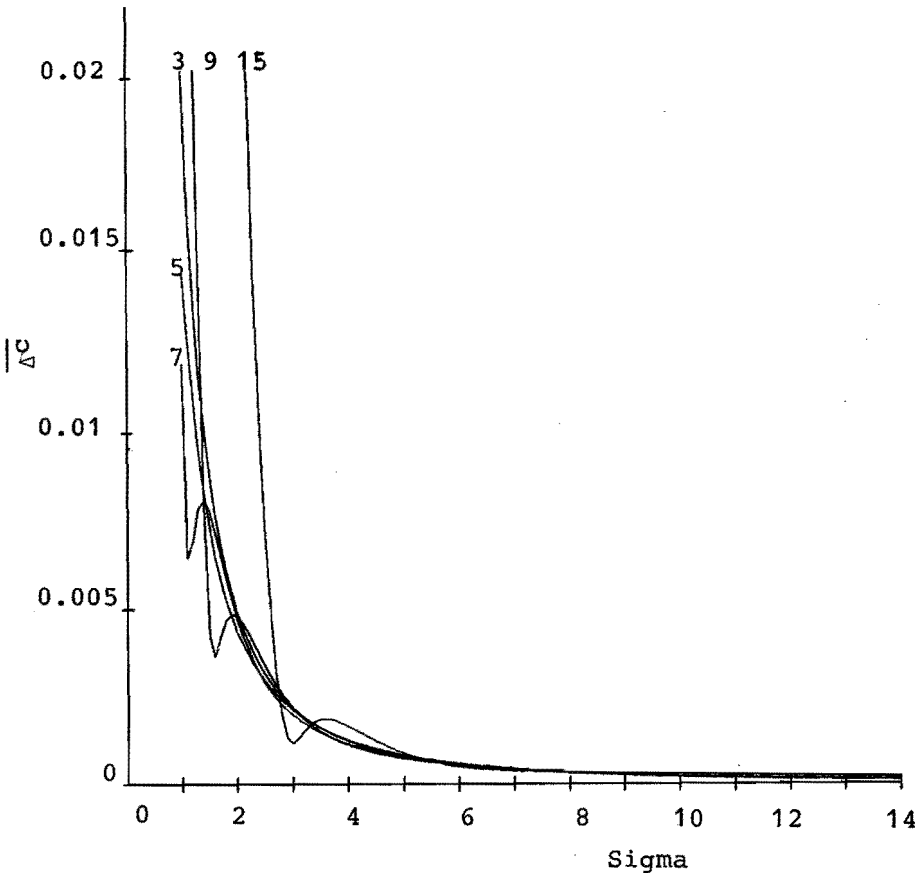
Figure 3.10. $\overline{\Delta c}$ as a function of sigma, for the n-values
as indicated in the graph and after applying the appropriate
q factors.

3.4 RANDOM NOISE ERRORS

After applying the correction factors, the parabola fitted through the top of a Gaussian peak gives very good results. It does not seem to make any difference whether three or more points are used for fitting the parabola. Therefore one would be inclined to use only the three point fit, since this is the shortest formula and hence requires the least amount of computer storage and time.

Unfortunately this only is true for Gaussian peaks. Gas chromatographic peaks differ on two counts. One is noise and the other is asymmetry. Particularly noise may impair the accuracy of the peaktop detection. In order to investigate this effect, a computer program has been written which generates Gaussian peaks with random noise super-imposed. The signal to noise ratio R is defined here as:

$$R = \frac{\text{height of Gaussian peak}}{\text{r.m.s. of noise amplitude}} \qquad (3.14)$$
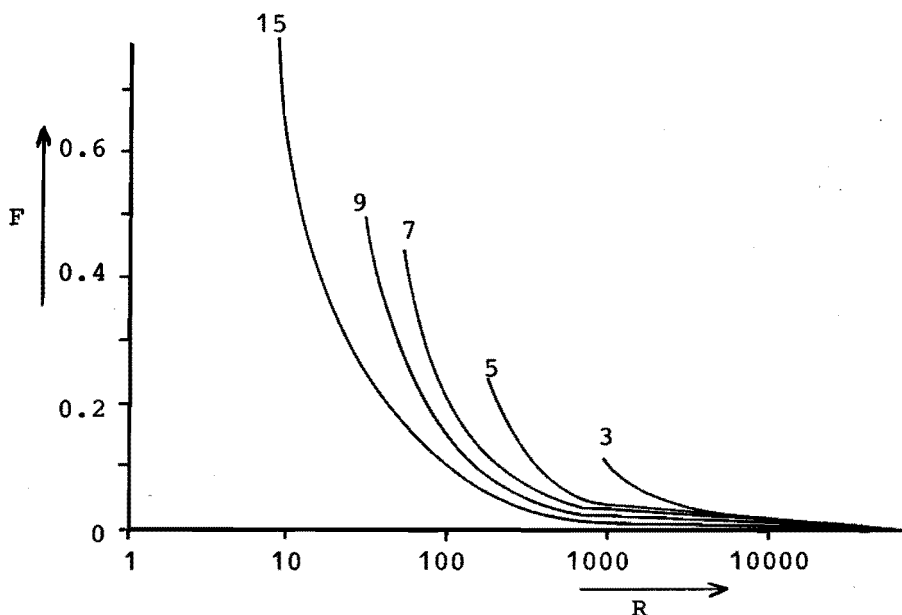
Figure 3.11   Standard deviation F of the error in the peaktop location due to noise for n-values as indicated.

The Gaussian peak was generated such that the actual top coincided with a datapoint. For every R this was done 50 times. The standard deviation F was calculated from:

$$F = \sqrt{\dfrac{\sum\limits_{i=1}^{50} \Delta c_i^{\;2}}{49}} \qquad (3.15)$$

The result is shown in Figure 3.11. It becomes now obvious that the best results are obtained using the largest value of n. However it should be realized still that this is only true for pure Gaussian peaks. The asymmetry has also an important effect on the peaktop allocation.

3.5 EFFECT OF ASYMMETRY

So far all the methods that have been used could be quantified quite well. This is a result of the fact that these techniques are well defined. Asymmetry however can not be described so well. There exists a large range of sources for asymmetry, out of which the following are worth while mentioning:
- Overloading phenomena (Ref. 3.2).
- Low injector temperature.
- Interaction with support phase.
- Insufficient separation.
- Detector response.
- Amplifier response.

The exact effects of most of these phenomena on the peak-shape are not known. Consequently it is not possible to describe real gas chromatographic peaks correctly by a mathematical model. Thus it is not possible to calculate the effects on the peaktop detection. Hence these effects can only be quessed. For the peaks that look symmetrical it seems fair to assume that the errors are in the order of magnitude as in the case of Gaussian peaks.

That is, provided that the parabolic fit is applied
only over the very top part of the peak and that the
correction factor q is used. Thus the curves of the
remaining error will be similar to the ones pictured
in Figure 3.5.

If the parabola goes over a large part of the GC
peak then the error will increase. The situation wor-
sens if the asymmetry becomes stronger. Since the asym-
metry can not be exactly described it is also impossible
to calculate the errors made in peaktop allocation. It
should be noted however that strong asymmetry points to
poor gas chromatographic conditions, which in turn cannot
provide accurate retention data and hence will not be
dealt with.
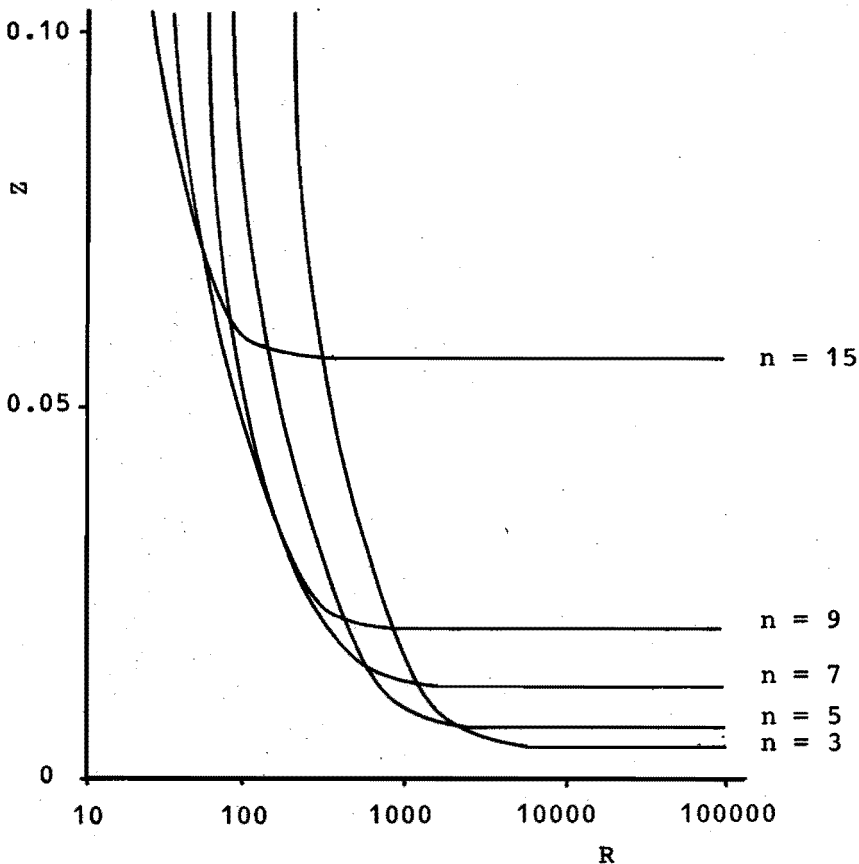
Figure 3.12. Overall error Z as a function of the signal
to noise ratio R and the width of the parabola n.

3.6 COMBINED EFFECTS OF NOISE AND ASYMMETRY

In gas chromatography noise and asymmetry will always be present. Both introduce errors in the peaktop allocation. The errors from noise and asymmetry are independent and can be combined into one overall error Z.

$$Z = \sqrt{\Delta c^2 + F^2} \tag{3.16}$$

Equation 3.16 holds only if the error from asymmetry indeed is equivalent to Δc, which only will be true if the asymmetry is not strong. The overall error for a peaksigma.

signal to noise ratio, R
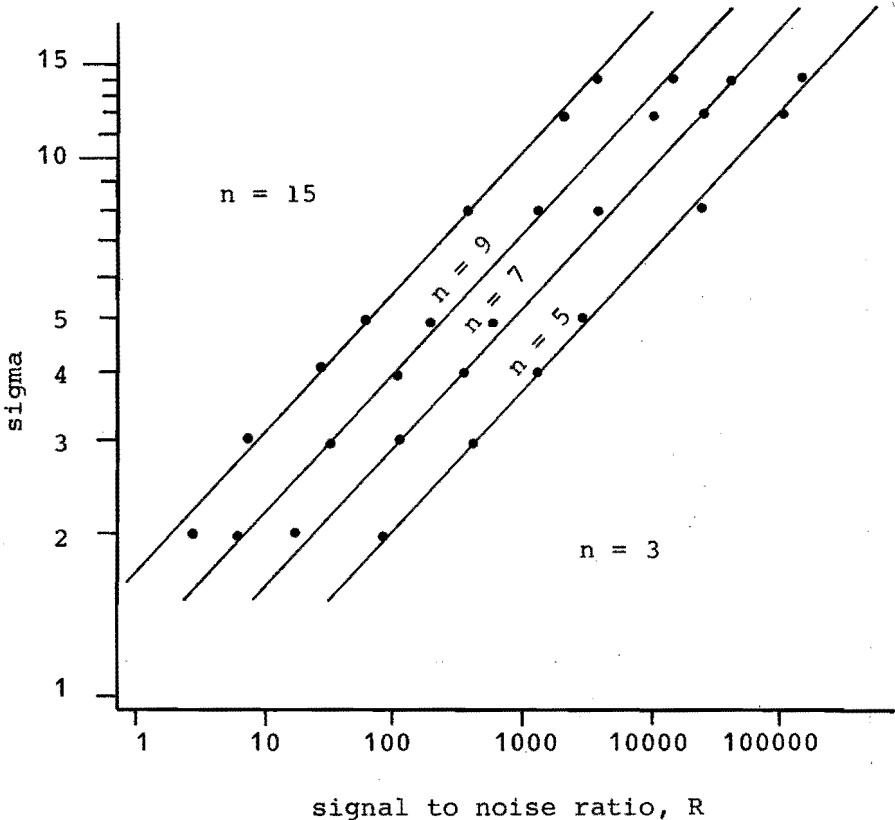
Figure 3.13. Optimum value of n as a function of sigma and R. Here n is the number of datapoints for fitting a parabola, and R the signal to noise ratio. The lines in the graph separate areas with optimum n-values.

of five is shown in Figure 3.12. From this picture it becomes clear that the required number of fitting points depends on the signal to noise ratio. In Equation 3.16 is also shown that Z depends on Δc, which in turn depends on the peaksigma, as was shown in Figure 3.5.

If Figure 3.12 is generated for a series of different peaksigma's, then the curve-intersections can be shown in another graph. After connecting these points Figure 3.13 evolves. This picture finally shows clearly how many points to use for a parabolic fit in order to get optimal results.

3.7 LITERATURE REFERENCES

3.1. A Savitsky and M.J.E. Golay, Anal. Chem., <u>36</u>, 1627 (1964).

3.2. C.A. Cramers, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1967, Chapter 1.

CHAPTER 4

SMOOTHING

4.1 INTRODUCTION

Any signal that can be measured has random fluctuations. These random fluctuations are called noise. The ratio of signal-amplitude to the noise-amplitude is called signal to noise ratio. This signal to noise ratio puts a limit on the accuracy with which the signal can be measured. It is therefore desirable to improve the signal to noise ratio. The technique of doing so is called smoothing or filtering. This can be done in a digital or analog way. The most commonly known method of smoothing is a RC filter. This is one type of hardware solution. In this thesis we are more interested in what can be done with software.

Of course it is possible to simulate a RC filter, but this filter has a disadvantage. It introduces a timelag in the signal. In other words: the signal is distorted. With this a major problem in smoothing techniques is encountered. They all introduce some form of signal distortion. Noise however can also be considered to be signal distortion. So it can be said that smoothing decreases the distortion due to noise, but introduces a new kind of distortion. As long as the balance of the two types of distortion is positive then smoothing is favorable. Unfortunately the two types of distortion are of a totally different kind and can not be easily compared. But, starting with a light smoothing and going to a heavier smoothing, it will be clear that there must be an optimum to the degree of smoothing that can be applied. This optimum depends on the noise level, peakshape, data rate, and smoothing function.

4.2 SMOOTHING FUNCTIONS

A survey of available smoothing functions is given by
Savitsky and Golay (Ref. 4.1). In general the possible
functions range from the moving average technique to com-
plete curve fitting. The first is the most simple, the
second a more complex technique. But the moving average
gives a much poorer result than curve fitting, provided
that a correct function is used. The moving average tech-
nique can be upgraded by using a different function e.g.
triangular, parabolic, binomial, Gaussian, or others (Ref.
4.2, 4.3).

The mathematical description of the smoothing process
is:

$$Y'_j = \frac{\sum_{i=-m}^{m} C_i Y_{j+i}}{N} \qquad (4.1)$$

$$N = \sum_{i=-m}^{m} C_i \qquad (4.2)$$

Y_j = datapoint j.

Y'_j = datapoint j after smoothing.

C_i = smoothing coefficient i.

N = normalizing value.

m = (smoothing width - 1)/2.

Thus the C-values describe the smoothing function. The
smoothing width is expressed in a number of datapoints.
All smoothing functions that will be described here are
symmetrical, therefore it will only be necessary to give
the C-values from zero through m. The various types of
smoothing that will be discussed are expressed in Tables
4.1. a and b. It should be noted here that a smoothing
width of three points usually is not enough to define the
smoothing function well. This is particularly true for the
parabolic function. All smoothing functions are basically
unlimited in extending their width.

Table 4.1.a

Coefficients for smoothing with a moving average, a tri-
angular function, and ·a parabolic function.

	smoothing width	N	m									
			0	1	2	3	4	5	6	7	8	9
moving average	3	3	1	1								
	5	5	1	1	1							
	7	7	1	1	1	1						
	9	9	1	1	1	1	1					
	11	11	1	1	1	1	1	1				
	13	13	1	1	1	1	1	1	1			
	15	15	1	1	1	1	1	1	1	1		
	17	17	1	1	1	1	1	1	1	1	1	
	19	19	1	1	1	1	1	1	1	1	1	1
triangular	3	4	2	1								
	5	9	3	2	1							
	7	16	4	3	2	1						
	9	25	5	4	3	2	1					
	11	36	6	5	4	3	2	1				
	13	49	7	6	5	4	3	2	1			
	15	64	8	7	6	5	4	3	2	1		
	17	81	9	8	7	6	5	4	3	2	1	
	19	100	10	9	8	7	6	5	4	3	2	1
parabolic	5	35	17	12	-3							
	7	21	7	6	3	-2						
	9	231	59	54	39	14	-21					
	11	429	89	84	69	44	9	-36				
	13	143	25	24	21	16	9	0	-11			
	15	1105	167	162	147	122	87	42	-13	-78		
	17	323	43	42	39	34	27	18	7	-6	-21	
	19	2261	269	264	249	224	189	144	89	24	-51	-136

Coefficients for smoothing with a binomial function and a Gaussian function

	smoothing width	N	m 0	1	2	3	4	5	6	7	8	9
binomial	3	4	2	1								
	5	16	6	4	1							
	7	64	20	15	6	1						
	9	256	70	56	28	8	1					
	11	1024	252	210	120	45	10	1				
	13	4096	924	792	495	220	66	12	1			
	15	16384	3432	3003	2002	1001	364	91	14	1		
	17	65536	12870	11440	8008	4368	1820	560	120	16	1	
	19	262144	48620	43758	31824	18564	8568	3060	816	153	18	1
Gaussian	3	10	8	1								
	5	20	8	5	1							
	7	101	27	22	11	4						
	9	106	22	19	13	7	3					
	11	91	15	14	11	7	4	2				
	13	109	15	14	12	9	6	4	2			
	15	190	22	21	19	16	12	8	5	3		
	17	286	30	29	26	21	18	14	10	6	4	
	19	401	37	36	33	29	25	20	15	11	8	5

4.3 SMOOTHING EFFICIENCY

It has been pointed out in the introduction of this chapter that smoothing has two aspects: noise reduction and signal distortion. In this section the ability to reduce noise of the various smoothing functions will be investigated (Ref. 4.4.)

Noise must be generated in order to evaluate the noise-reducing power of a smoothing function. This is done by generating normally distributed numbers by computer. These figures represent random noise with a certain r.m.s. amplitude. After applying the smoothing function the r.m.s. amplitude can again be calculated. The ratio of the two will here be defined as the smoothing efficiency E.

$$E = \frac{\text{r.m.s. before smoothing}}{\text{r.m.s. after smoothing}} \quad (4.3)$$

The dependency of E on the smoothing function and the smoothing width is graphed in Figure 4.1. From this figure it is obvious that the moving average is the most efficient and the parabolic and binomial are the least efficient of the functions studied.

In order to compare the efficiency better it is useful to define a relative efficiency e.

$$e_E = \frac{\text{width function X (E)}}{\text{width moving average (E)}} \quad (4.4)$$

Equation 4.4 gives e for function X at a smoothing efficiency E. The smoothing efficiency has always the highest value for the moving average function. Consequently e is always smaller than one. Any e-value can be derived from Figure 4.1. The e-values for the various smoothing functions are pictured in Figure 4.2.
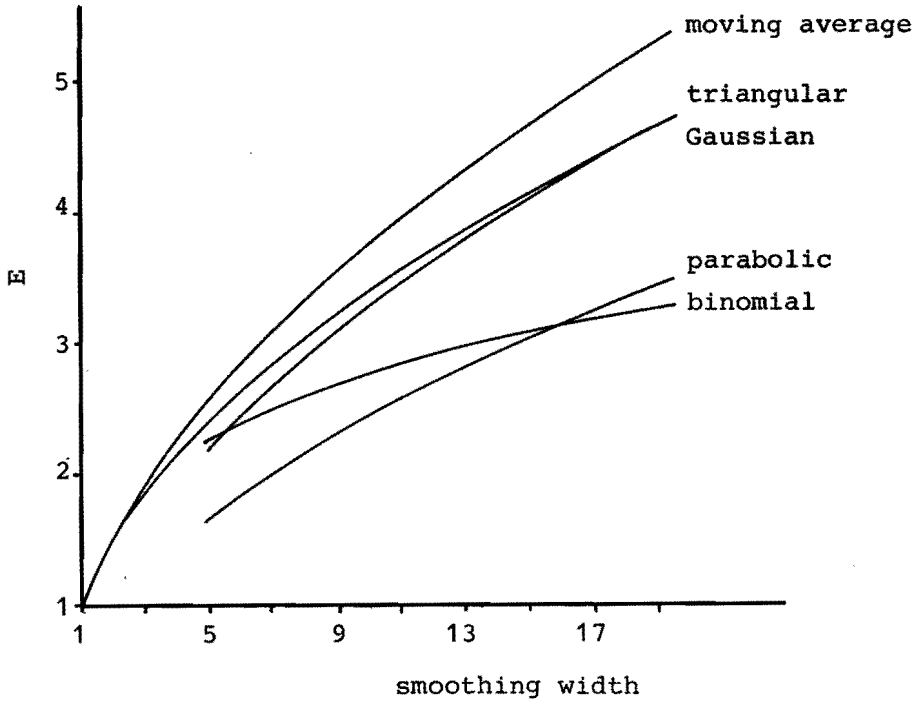
Figure 4.1. The smoothing efficiency E as a function of
the smoothing width and for various smoothing functions.
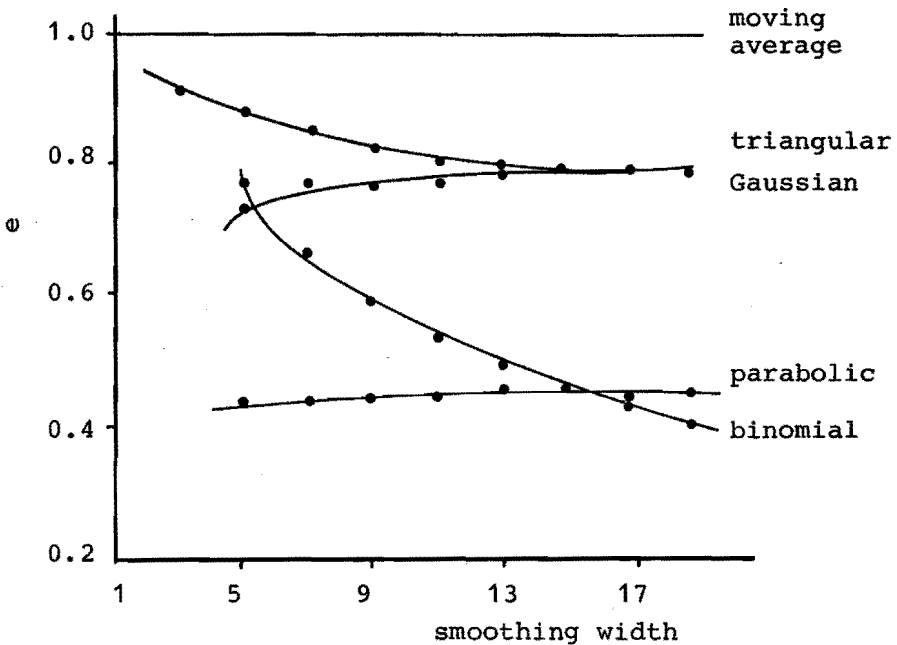
Figure 4.2. The relative smoothing efficiency e as a
function of the smoothing width and for various smoothing
functions.

Since the points on every curve may come from various combinations of smoothing width, sigma, and relative smoothing efficiency (e) it is quite surprising that the graphs consist of simple lines. This means that given a required smoothing it is possible to see directly from the graphs which type of smoothing is best. Of course, provided that it is known how much distortion is allowed.
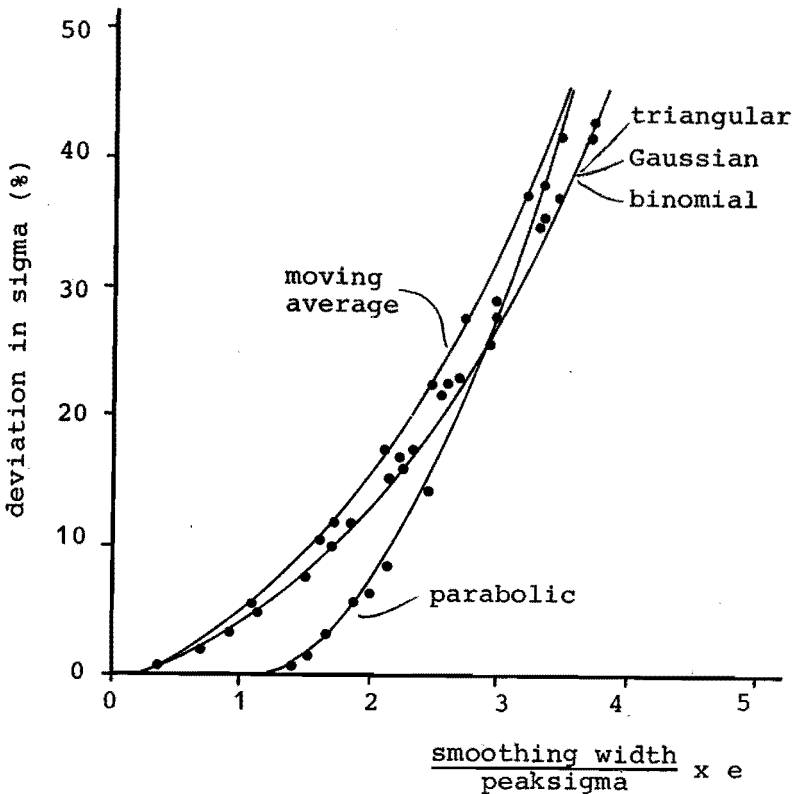
Figure 4.3. The influence of smoothing on the sigma of a Gaussian peak.

4.4 SMOOTHING DISTORTION

It is quite unfortunate that the distortion due to filtering cannot be expressed as simply and as explicitly as the distortion due to noise. This results from the fact that the actual signal has many more important aspects then just the r.m.s. amplitude.

In the introduction it has been explained that smoothing affects the signal. Here the signal could be defined as: gas chromatographic peaks. This definition is still too vaque though since the exact shape of a GC peak cannot be well described. Therefore the influence of smoothing shall be indicated on true Gaussian peaks.

The shape and location of a gas chromatographic peak in general can be described by the zero through the fourth moments (Ref. 4.5, 4.6). Distortion of a peak could then be described in terms of changes in these moments. For a Gaussian peak and symmetrical smoothing functions it is then obvious that there will be no effect on the zero, first, and third moment. Only the second and fourth moment (and consequently sigma and kurtosis) may be affected.

Another means of quantifying the distortion is through calculation of the error signal. The error signal here being defined as the difference of the original signal and the smoothed signal. The amplitude of this error signal is a measure of the introduced distortion. The influence of the various smoothing functions on the sigma is pictured in Figure 4.3, the influence on the kurtosis in Figure 4.4. The error signal is shown in Figure 4.5. On the horizontal axis of these three graphs is shown the ratio of the smoothing width to the sigma of the Gaussian peak being smoothed, multiplied by the relative smoothing efficiency.
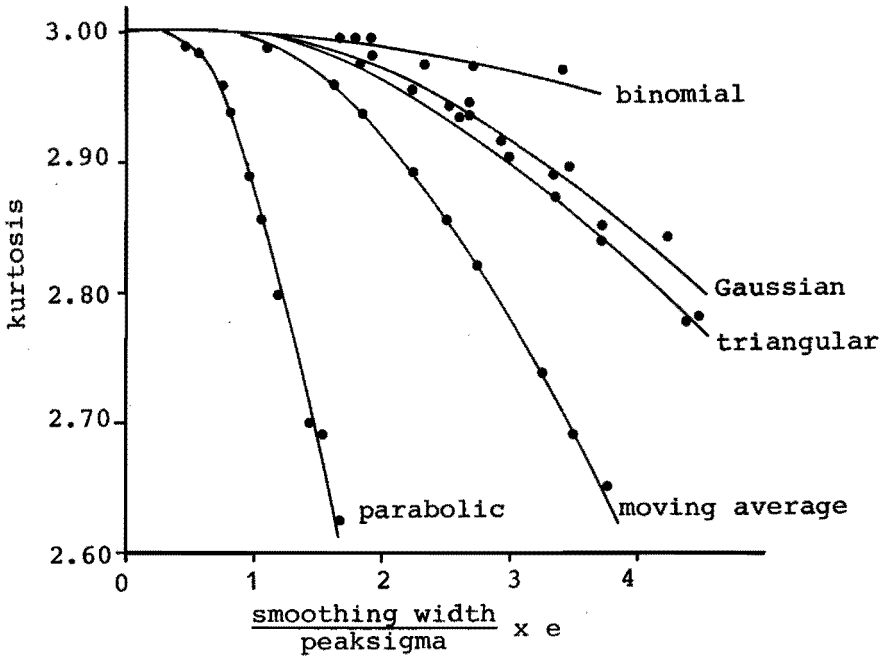
Figure 4.4.   The influence of smoothing on the kurtosis.
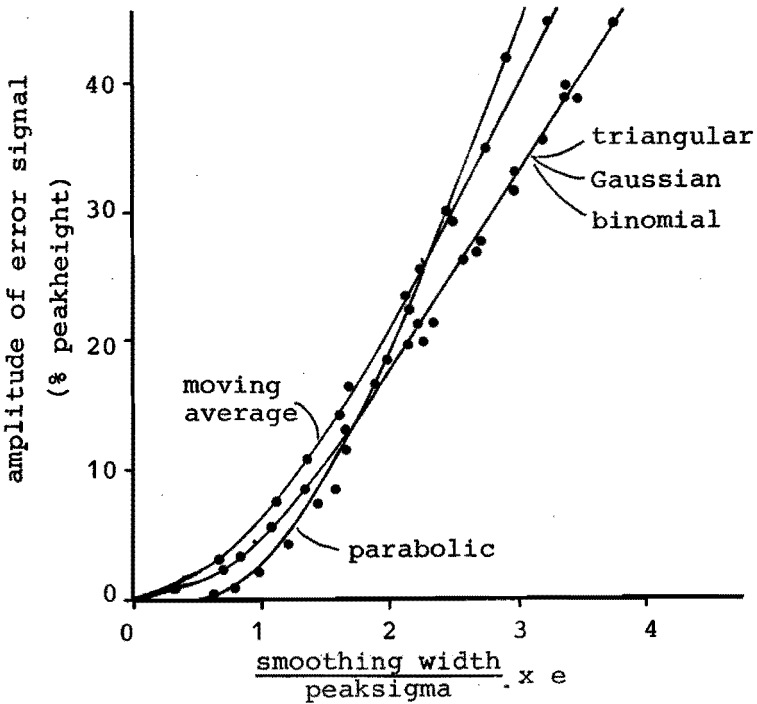
Figure 4.5.   The amplitude of the smoothing induced error
signal.

4.5 FURTHER EVALUATIONS

It is now possible to compare the various smoothing functions on their ability to reject noise. It is now also possible to quantify the smoothing-induced distortion, given a Gaussian peak. The effect on strongly asymmetrical peaks is still unknown. But this is less important, as long as it is realized that the area never will be impaired.

One factor of importance is that little is known as to how smoothing affects peak detection routines.   In general it is to be expected that peak detection will be more reliable if the noise level goes down, as long as the peaksigma is not increased too much (Ref. 4.7, 4.8).

Looking at Figures 4.3, 4.4, and 4.5 it may be concluded that the parabolic smoothing is often the best choice. Except when the kurtosis is studied and expect when a very heavy smoothing is necessary.

Practical points worth while noting here are:
- Applying a moving average can be programmed to run much faster on a computer than any of the other functions.
- It is not absolutely necessary to apply the functions exactly as they are expressed in Tables 4.1.a and b.
- It may be possible to change the factors slightly so that they may be programmed as bit shifts. The latter greatly increases the execution speed while only slightly impairing the smoothing function.

4.6 SPIKE REMOVING

Commonly noise is divided into three categories: drift, noise, and spikes. This is illustrated in Figure 4.6. Drift is the low frequency component of noise while the high frequency component is still called noise. Spikes appear as sharp lines that project above the band of normal noise. Yet spikes are also noise.

When a smoothing routine is applied, the noise amplitude
will be reduced and so will the height of the spikes. One
could say that the "wrinkles" of the noise are being
"ironed out". For a spike this would mean lowering and
broadening. This however could make a spike appear to be
a gas chromatographic peak, which would be highly undesir-
able. Therefore, it would be very useful if spikes could
be recognized as such, and then completely eliminated.

The recognition of spikes can be explained using Figure
4.7. In this figure, x is the location of a datapoint that
is being checked. The points a and b are preceding points,
while c and d are following points. Point p is the inter-
section of the line a-b with the position of x. Point q
follows similarly for the line c-d. Point b' is the ampli-
tude of b projected at location x. Datapoint x is now con-
sidered a spike if its amplitude is outside the range of
the points p, q, b', and c'. If a datapoint is found to
be a spike then it will be replaced by the average value
of the preceding and the following two datapoints. In the

drift

+

noise

+

spikes.

=

actual baseline
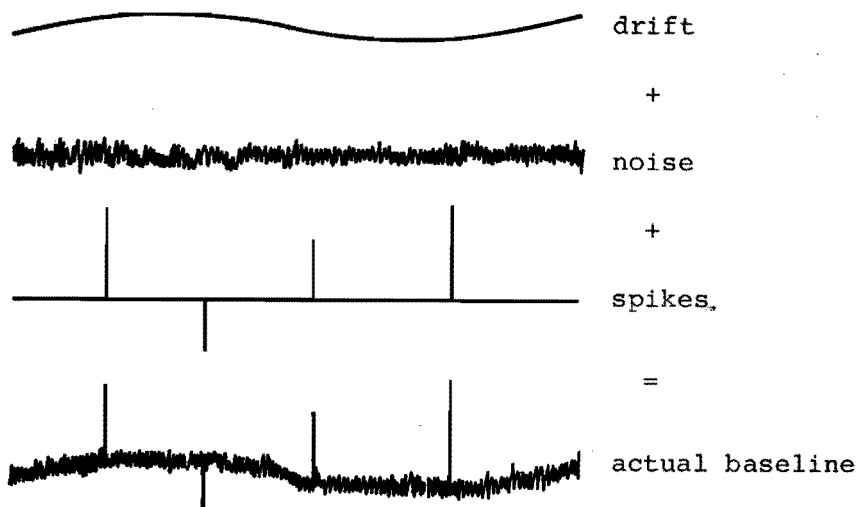
Figure 4.6. Different kinds of noise.
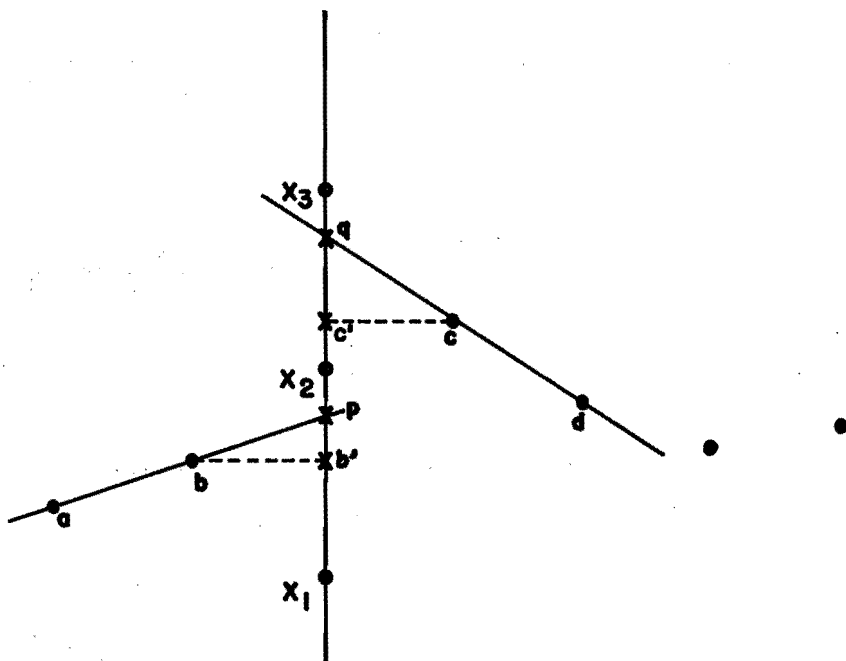
Figure 4.7. Spike removing.

figure as it is drawn, x_1 and x_3 are considered spikes while x_2 is not. In practice this appeared sometimes to be too sensitive. An increase of the range of p, q, b', c' with a fraction of the total data span would decrease this sensitivity. A value of 0.02% of the data span worked very well in practice.

A flowchart of a computer program that performs this type of spike removing is described in Section 5.2.1. In practice it appeared occasionally useful to apply the spike removing two times. An example of the capabilities of spike removing and smoothing is given in Figure 4.8. This figure shows a part of an actual chromatogram together with the results of spike removing and parabolic smoothing.

52
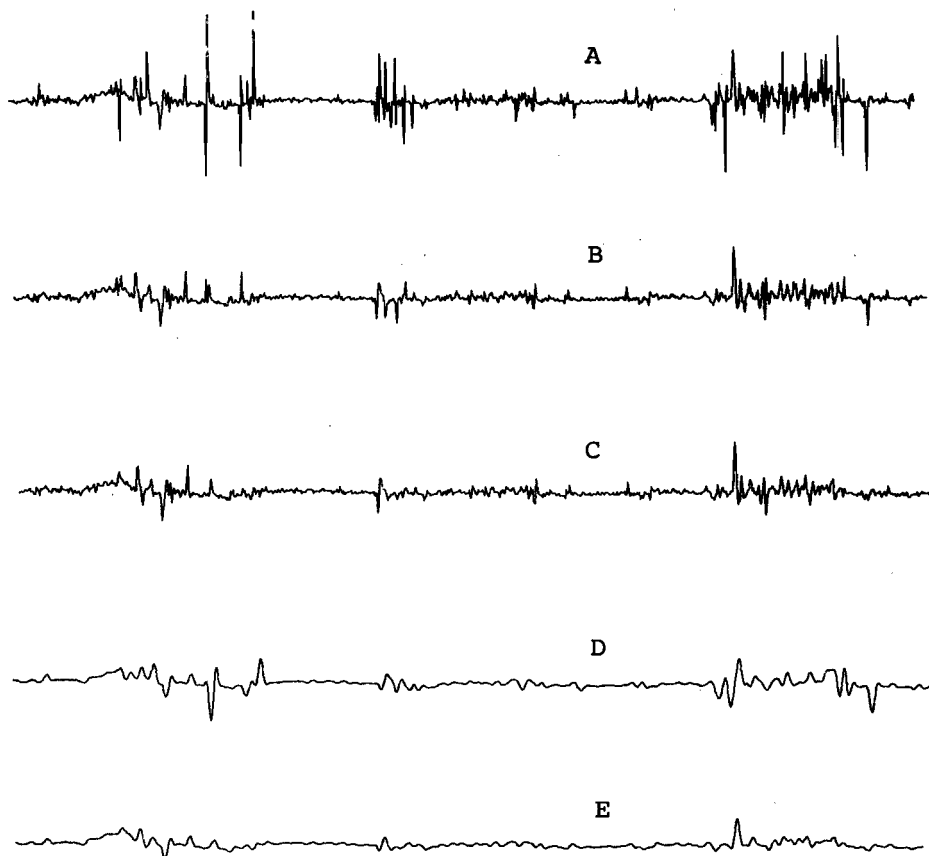
Figure 4.8. Effects of spike removing and smoothing on an actual gas chromatographic baseline.

curve	spike removing	smoothing
A	none	none
B	single	none
C	double	none
D	none	17 point
E	double	17 point

note: Double spike removing means performing the spike removing two times.

CHAPTER 5
SOFTWARE

5.1 INTRODUCTION

Based on the considerations as laid down in the pre-
vious chapters a computer program has been written for
handling gas chromatographic data. This program, called GCX,
will be described in this chapter. The aim of this program
is to deduce as much information from the raw gas chro-
matographic data as possible. In pursuing this goal the
following considerations were made:

- The program should run with as little user supplied
information as possible.

- The program should correct wherever possible for
human or instrumental errors.

- The program should handle gas chromatograms of all
kinds without problems, provided the gas chromatographs
are not temperature (or otherwise) programmed.

- From all chromatographic parameters to be calculated
particularly the retention time should be calculated with
the highest possible precision.

- It should not take more than one hour for an operator
of a gas chromatograph to learn to use the program.

- An abnormal condition occuring during the processing
of the data should be mentioned in the output.

- Any message that appears in the output should be
understandable by any gas chromatograph operator.

- The software should be modular so that it can easily
be changed and updated.

- Execution times should be kept short. However, this
should not happen at the expense of any of the other points
mentioned here.
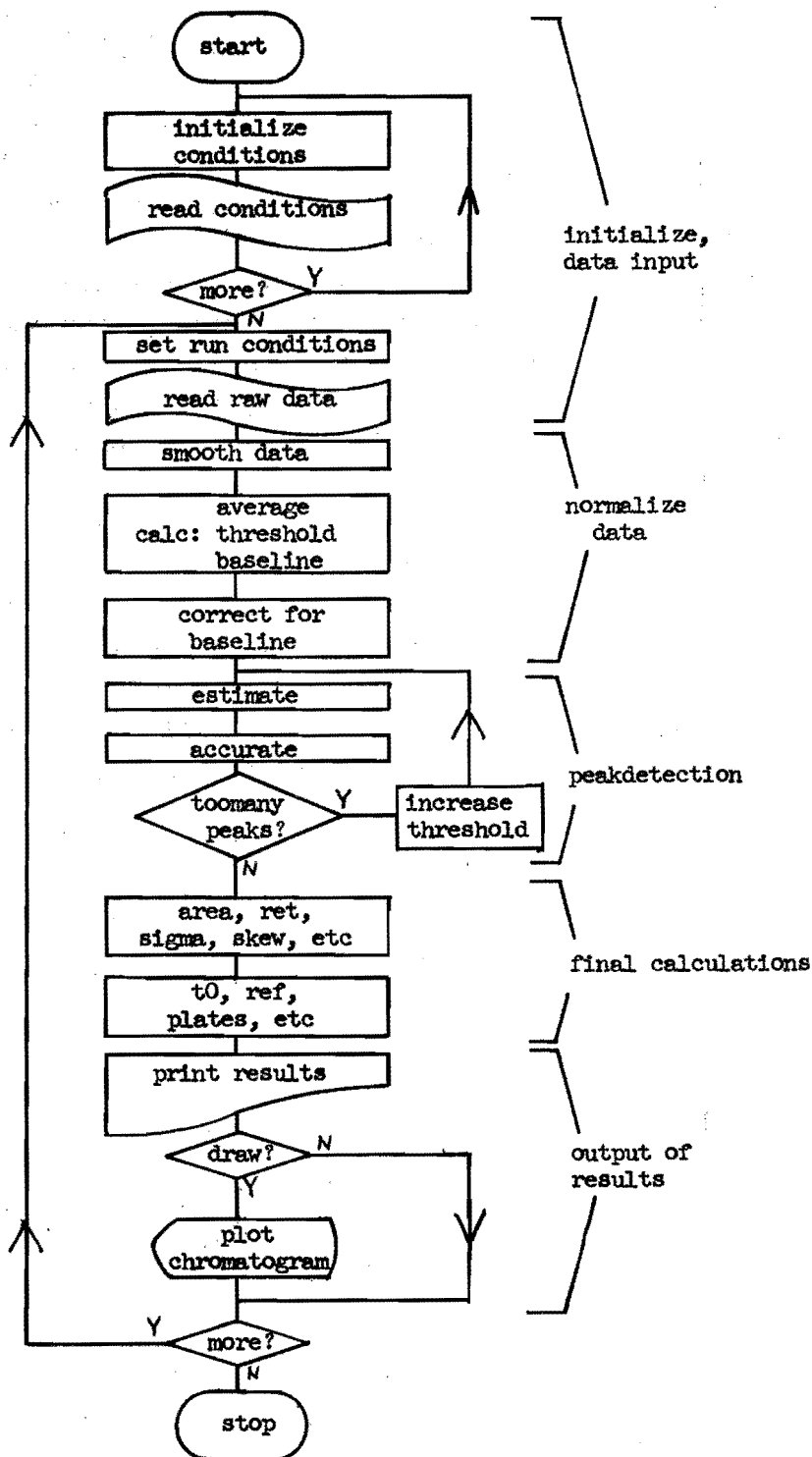
Figure 5.1. Simplified flowchart of the GC-program.

- The output should be in a clear and convenient format.
- The user should be able to select special options for handling unusual or unexpected types of chromatograms.

Several of these considerations stem from the fact that out laboratory has always been particularly interested in high retention data. This implies that most work is done on isothermal gas chromatography. It should be noted however that this does not mean that it is definitely impossible to handle data from programmed gas chromatography. In practice it showed that these chromatograms were processed quite well. Only with high programming rates did the rising baseline impair the obtained accuracies.

A simplified flowchart of the final program is shown in Figure 5.1. The complete listing of the program is about 30 pages printed output. Because of its length this listing is not given here. Instead flowcharts of the relevant parts of the program will be given with additional information and explanation. Initialisation and data input will be discussed after a description of the remainder of the program.

5.2 DATA NORMALISATION

Data normalisation consists of the following routines:
- Smoothing
- Calculating levels
- Correcting for baseline.

Smoothing consists of two steps: spike removing and actual smoothing. The calculation of the various levels could be in part considered to belong to peakdetection, however the baseline level is required to normalize the data.

5.2.1 Spike removing

Refer for a flowchart of this routine to Figure 5.2. This routine is designed according to the theory of Section

4.3. Before the actual routine is entered, the first three datapoints are made equal to the average of the next four datapoints. Also the last two datapoints are made equal to the average of the three preceding ones. This seemed to be the only real effective way to get rid of start and stop transients of the data acquisition system.
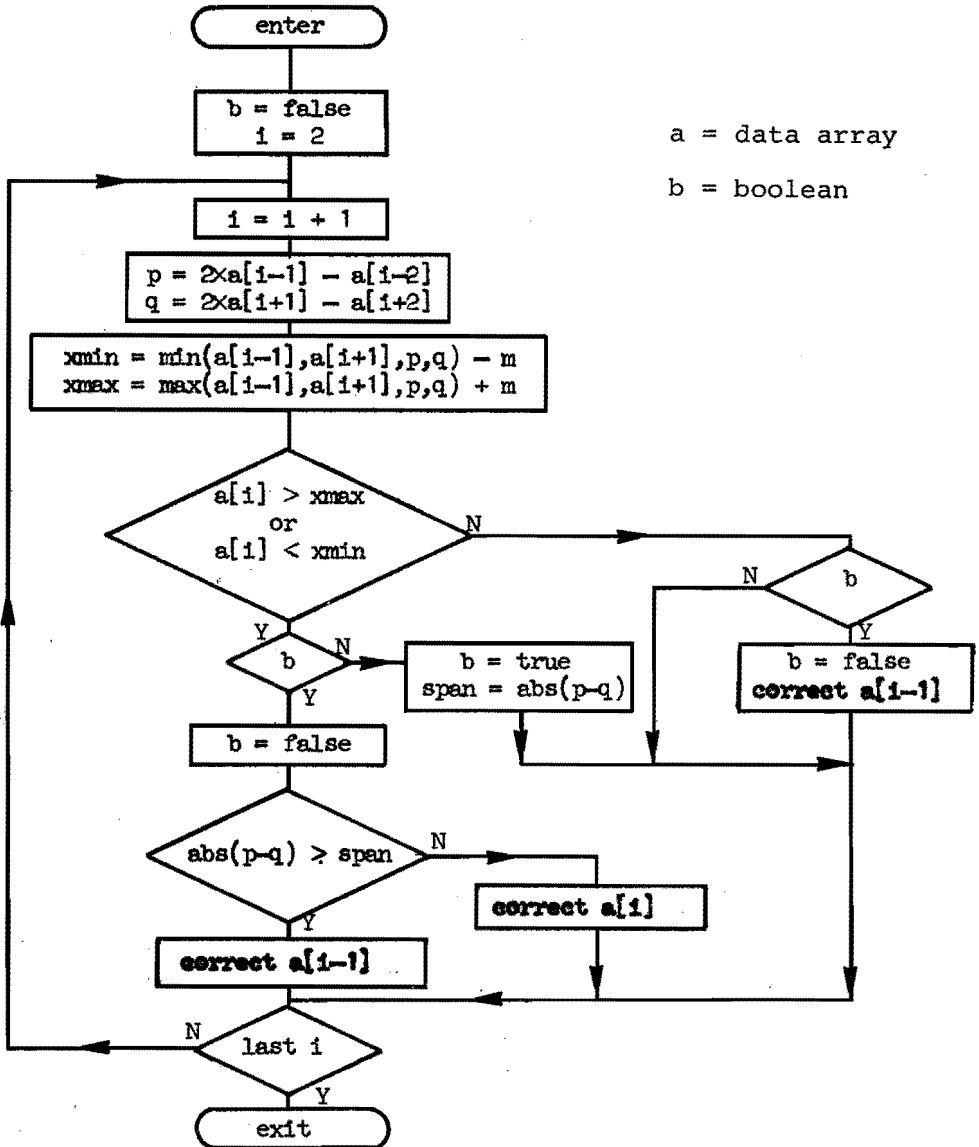
Figure 5.2. Flowchart of the spike removing routine.

5.2.2 Smoothing

The smoothing is performed by the parabolic function as described in Chapter 4. The initial smoothing width can be selected with a maximum width of nine datapoints. After smoothing half the data the smoothing width is increased by two. A heavier smoothing may be obtained by repeated smoothing. In case of heavy smoothing (more than seven points) the parabolic smoothing is preceded by a five point Gaussian function. This is done to avoid negative dips adjacent to sharp peaks.

5.2.3 Calculating levels

The following initial levels are recognised:
- Average
- Threshold
- Baseline
- Low baseline.

These levels are calculated according to the average below average technique as explained in Section 2.4.

The average below average technique was selected since this method uses the data on a statistical basis rather than preselecting a part of the data. Practice showed that this method worked quite accurately and very reliably. Consequently no attempts have been made to incorporate other methods.

For the calculation of the average every i^{th} datapoint is used, whereby i is chosen such that in total about 300 datapoints are used. For each consecutive level, i is then decreased by one until i is two.

In addition to the mentioned levels a "special" baseline is calculated. This special baseline is the average of the first and the last m datapoints. Normally m is 20, but can also be user selected.

The final baseline will be selected according to the flowchart in Figure 5.3. This final baseline has in exten-
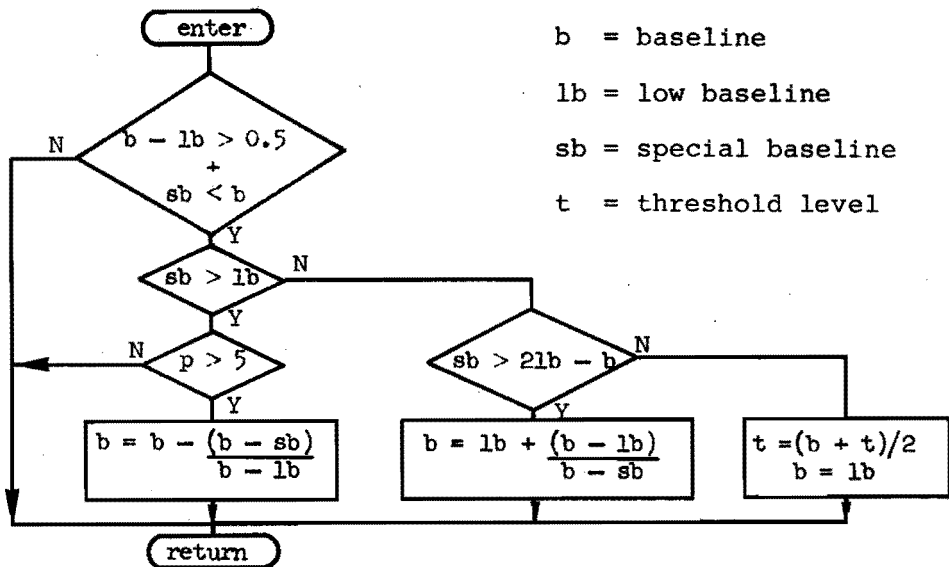
Figure 5.3. Flowchart of the baseline selection.

sive experimentation proved to work very well. In only
about ten percent of the processed chromatograms could a
slight improvement be obtained by selecting the "special"
baseline, or the "low" baseline.

As is shown in this flowchart the threshold level may
also be changed by the program. It is also possible to
select a final threshold level different from the calcu-
lated one.

5.2.4 Baseline correcting

Once the final baseline has been established, this
value will be subtracted from all datapoints and from any
other value that is related to the signal amplitude, e.g.
the other levels, overflow value, etc. In doing so the
signal and all its properties become a deflection from zero
and the baseline itself becomes essentially equal to zero.
This greatly simplifies all following calculations.

5.2.5 Solventpeak skimming

One problem which quite often is encountered in gas chromatography is that the components of interest are diluted with a large amount of solvent. In order to inject sufficient sample, the column will be overloaded with solvent. This in turn gives rise to an offscale peak with a very long tail. This tail is often as long as half the chromatogram or even longer. Even though this tail is essentially part of a peak and not the baseline, it is mostly regarded as being a baseline as far as the other peaks are concerned. Therefore this tail and preferably the whole solvent peak should be removed. This is done in a routine which is flowcharted in Figure 5.4. This routine applies a technique known as "skimming" (Ref. 5.1).

In short the main actions are as follows. First the leading edge is recognized, followed by detecting the beginning of the trailing edge. The solvent peak will then be completely eliminated until the trailing edge and with the exception of the leading edge. This leading edge now shows up as a very sharp and tall peak. From the start of the trailing edge (f in the flowchart) a straight skimming line is calculated that connects f with only one other point (touch in the flowchart). "Touch" is chosen such that all other points are above this line. That part of the skimming line that runs from "f" to "touch" is said to represent the actual solvent peak in that range. The part below this line is then eliminated. Now "touch" gets the function of "f" and a new touching or skimming point will be found. This will be continued until the signal is back at the original baseline or until the end of the chromatogram is reached.

The program will apply peak skimming only upon a users request. The routine is actually located in between the routines for smoothing and level calculating.

enter

find lead

calc. initial baseline
correct for baseline
calc. initial noise

i = lead + 3

i = i + 1

trail start — N → a[i] = 0

Y

f = trail

x = 0
slope = 0
b = a[f]

x = x + 1
y = slope × x + b

y < 0
or
f + x ≥ last — Y

N

a[f+x] < y — N

Y

slope = (a[f+x]−b)/x
touch = f + x — Y ← a[f+x] > 0

N

b = b + noise/2
i = f − 1

i = i + 1
a[i] = a[i] − b − (i−f)×slope

i = touch − 1 — N

Y

f = touch

end chromatogram — N

Y

exit

a   = data array.
lead = start of leading
       edge of solvent.
trail= start of trailing
       edge of solvent.
slope= slope of the
       skimming line.

Figure 5.4.   Flowchart of the skimming routine.
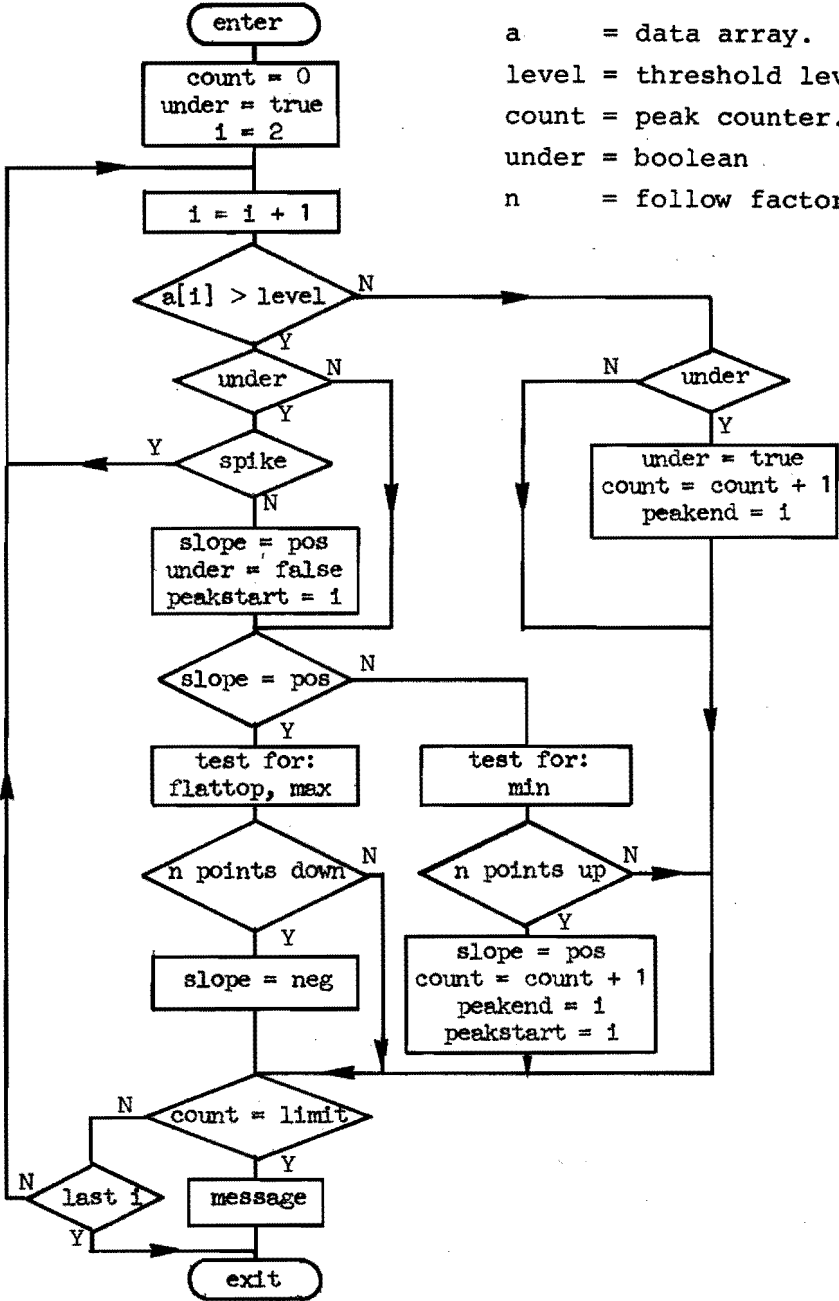
62

Figure 5.5. Flowchart of procedure ESTIMATE.

5.3 PEAK DETECTION

Peak detection is essentially based on the use of
threshold levels as calculated according to Section 5.2.3.
In a first step the peak locations are estimated. In a
second step the estimated positions are made more accurate.
The procedures that perform these functions are called
ESTIMATE and ACCURATE. After the actual peak detection the
noise and drift of the pieces of baseline in between the
peaks are calculated. Finally a check is made if there
were not too many peaks. If so, a few variables are changed
and the whole peak detection is done again.

5.3.1 Procedure ESTIMATE

The functions of this procedure are:
- Estimate the start, the top and the end of each peak.
- Check if the peaktop is flat.
- Reject spikes.
- Count the number of peaks.
- If 99 peaks are detected give a warning and stop
  searching for more peaks.

For the flowchart refer to Figure 5.5. An explanation of
this routine can be found in Section 2.4.

5.3.2 Procedure ACCURATE

As indicated before, this procedure evaluates the start
and end points of peaks. For a flowchart refer to Figure 5.6.
This is done by extending the peak boundaries. Before this
is actually done the routine determines first the maximum
limit for extending the peak boundaries. This limit is
a function of the peakheight and peakwidth, as deter-
mined by ESTIMATE. The area in between the estimated peak
boundaries and the limit could be called a "window" in
which the accurate peak boundary is to be found. If the
window of one peak overlaps the window of an adjacent peak

$$p = en[j] + 0.5 \times b + 4 \times b \times level/top[j] + 2$$

$$q = st[j+1] - 0.3 \times b - 5 \times b \times level/top[j+1] - 2$$

$$r = en[j] + \left(\frac{0.8 \times top[j]}{top[j]+top[j+1]}\right) + 0.1 \times (st[j+1]-en[j])$$

$$b = st[j+1] - en[j]$$

st[j] = startpoint of peak j

en[j] = endpoint of peak j

top[j] = height of peak j

Flowchart:

enter → j = 0 → j = j + 1 →

st[j+1] = en[j]?
- Y → (exit loop path)
- N → calc: p, q

p > q?
- N → (skip)
- Y → calc: r

r < q?
- Y → p = q
- N →

r < p + r > q?
- Y → p = r
- N →

trigger(en[j],p)

r ≥ p + p > q?
- Y → q = en[j]
- N →

trigger(st[j+1],q)

peakshape OK?
- N → discard peak
- Y →

last j?
- N → (loop back to j = j + 1)
- Y → exit
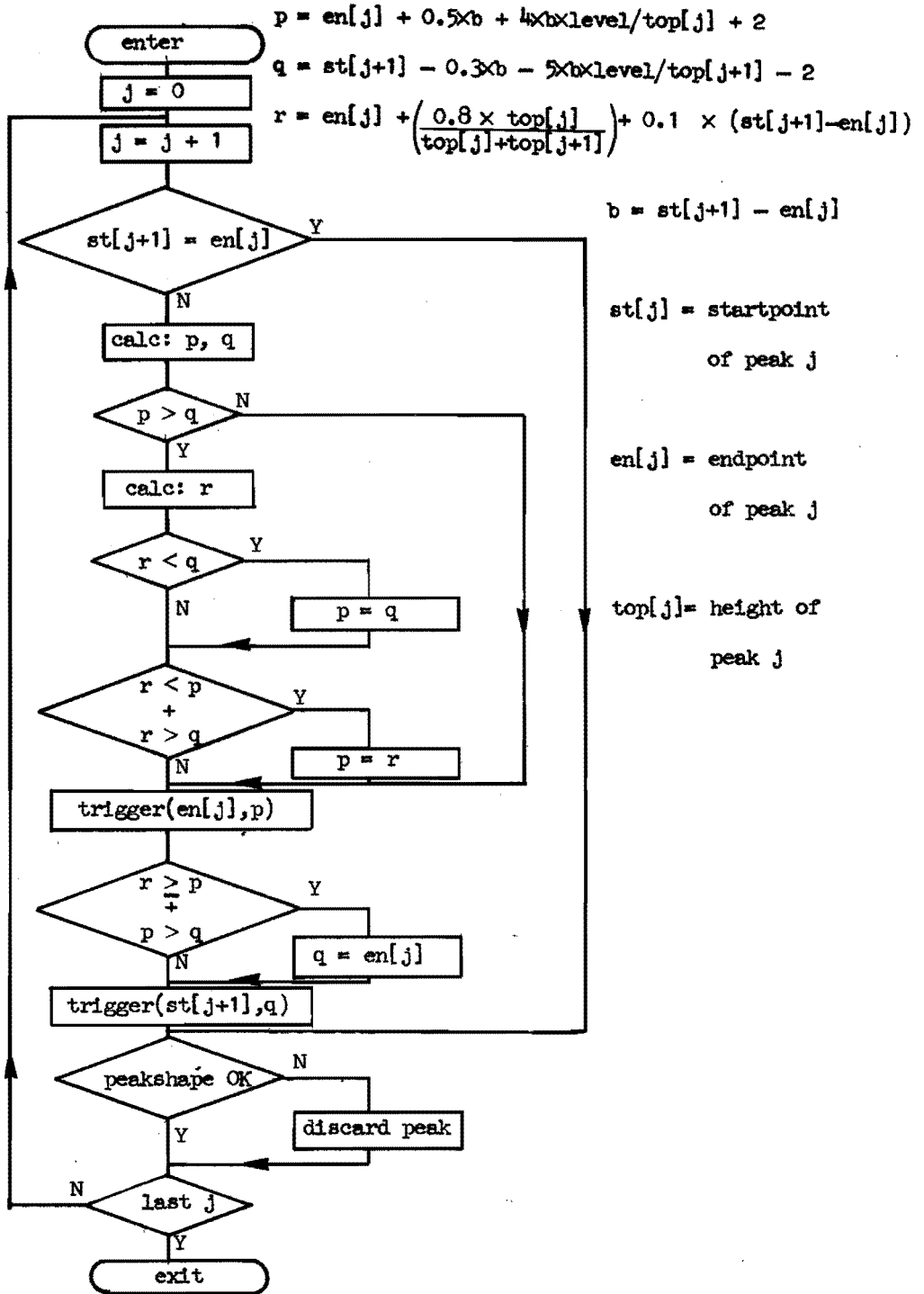
Figure 5.6.   Flowchart of procedure ACCURATE.

then the space in between these peaks will be divided. This will be done such that each window is proportional to the height of the corresponding peak.

Once the window is calculated, procedure TRIGGER is called. Refer to Figure 5.7. In this routine the peak boundary will be extended until one of the following conditions is met:

   - The signal drops below zero.
   - A valley is detected.
   - The edge of the window is reached.

After this ACCURATE checks the peakshape. Peaks that

a = data array.

$$\left.\begin{array}{l}\text{start}\\\text{finish}\end{array}\right\}\text{procedure arguments}$$
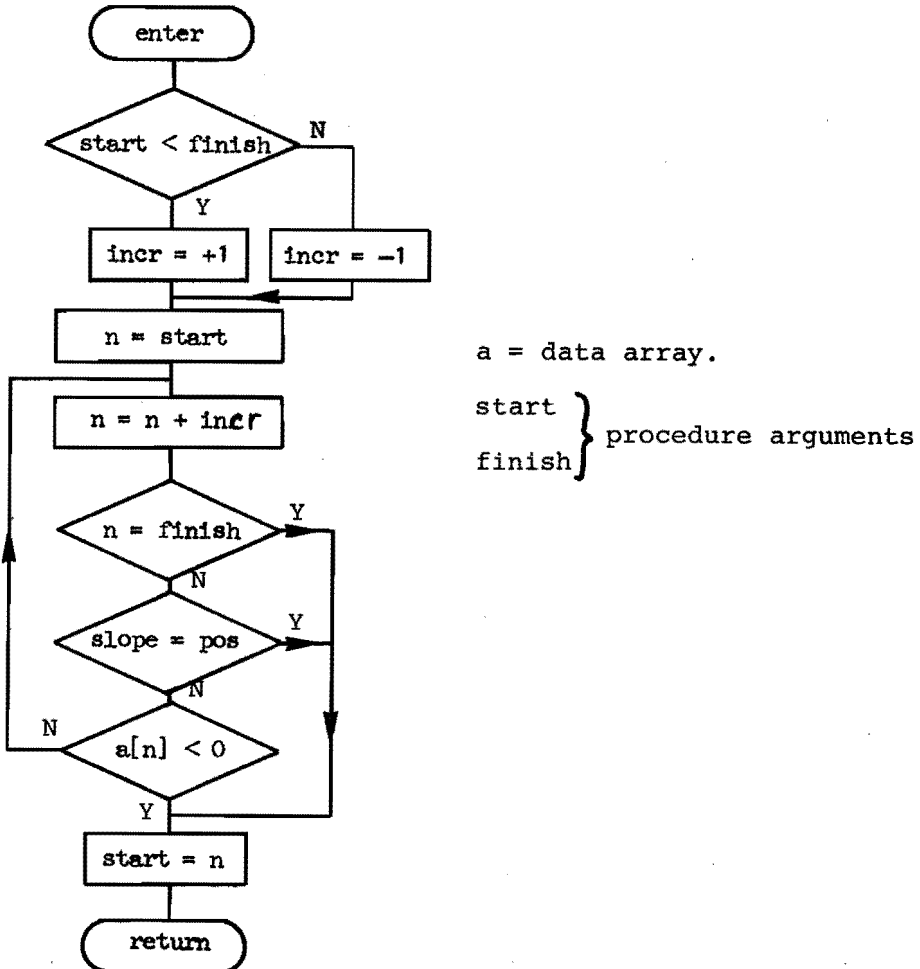
Figure 5.7.   Flowchart of procedure TRIGGER.

are too narrow (less than six datapoints) or that are com-
pletely skewed (top within two datapoints of the peak
boundary) will be discarded.

5.3.3 Test on too many peaks.

After completion of ACCURATE the program calculates
the average drift and noise of the remaining baseline. A
new level is calculated. It is called "skiplevel" and is
at the threshold level plus four times the noise level.
"Skiplevel" is used later in the program.

Having calculated these, a check is made if ESTIMATE
had found too many peaks. If so, the program will enter the
routine flowcharted in Figure 5.8 and return to ESTIMATE.
As indicated in the flowchart, the value of the threshold
level is changed. The value of "follow" is increased if
the program enters the routine more than once. One could
say that this routine readjusts the peak detection sensi-
tivity if the peak detection proved to be too sensitive.
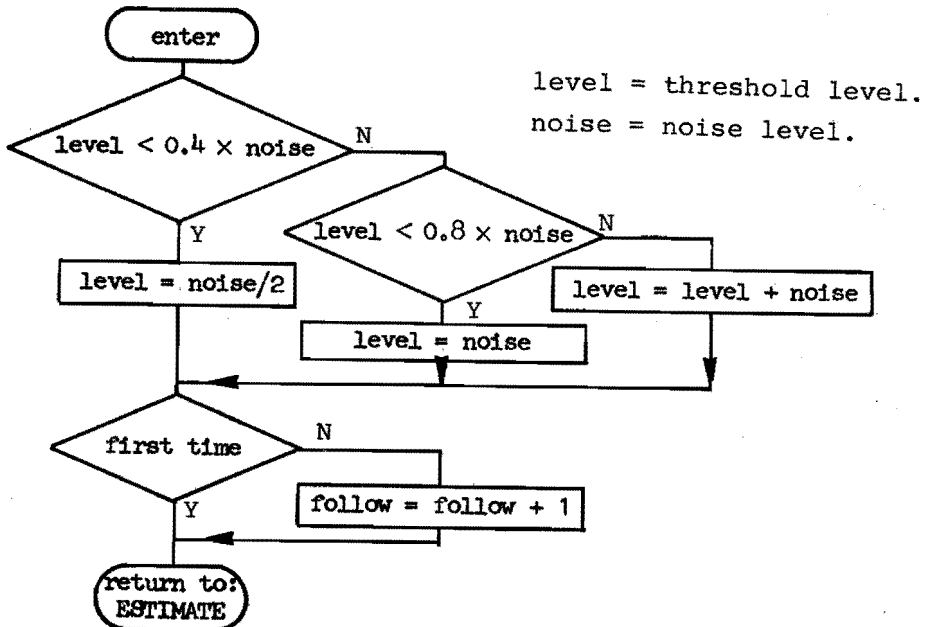
Figure 5.8. Flowchart of the threshold recalculation.

5.4 FINAL CALCULATIONS

It does not seem to be useful to flowchart this part of the program. Only the formulas will be given that are used in the final calculations.

Area:
$$A = \sum_{i=start}^{end} a_i \qquad (5.1)$$

Peaktop: according to Chapter 2.

Center of Gravity:
$$G = \frac{\sum_{i=start}^{end} i a_i}{A} \qquad (5.2)$$

Sigma:
$$s = \frac{\sum_{i=start}^{end} (i-G)^2 a_i}{A} \qquad (5.3)$$

Skew:
$$W = \sum_{i=start}^{end} (i-G)^3 a_i \qquad (5.4)$$

$$w = W/s^3 \qquad (5.5)$$

Kurtosis:
$$K = \frac{\sum_{i=start}^{end} (i-G)^4 a_i}{A} \qquad (5.6)$$

$$k = K/s^4 \qquad (5.7)$$

Plates:
$$P = (G/s)^2 \qquad (5.8)$$

Resolution:
$$R_j = (G_{j+1} - G_j)/j \qquad (5.9)$$

The amplitude of the ith datapoint is indicated by a_i. The subscript j indicates the jth peak.

The average plate number is calculated from all those peaks that are higher than the skiplevel (refer to Section 5.3.3). Also excluded for the calculation of

CONTENTS OF CONDITIONTAPE:
+710706   +23   +4   +4.00

- - - - - - - - - - - - - -

DATE OF MEASUREMENT: JULY * 6 - 1971   RUNNUMBER   23   GCX 31   15-9-1972

NUMBER OF POINTS   1678
IN TOTAL REMOVED SPIKES   0
PARABOLIC SMOOTHING WIDTH 3
MAXIMUM VALUE   +18659
MINIMUM VALUE   -1274

              LEVEL   LOWDATA PERCENTAGE   PEAKS HIGHER

AVERAGE   -697.29   74.74   5
THRESHOLD   -1238.12   54.11   9
BASELINE   -1262.15   28.78   6
LOW BASE   -1268.94
SPEC BASE   -1271.03   BSPEC IS +20.00 SEGMENTS ARE: -1272.05 -1270.00

THE FINALLY SELECTED BASELINE IS AT -1268.94
NOTE: BEFORE FURTHER PROCESSING ALL DATA WILL BE CORRECTED FOR BASELINE..........

THRESHOLD LEVEL AFTER BASELINE CORRECTION   24.04
BASELINE DRIFT AND NOISE (R.M.S.)   7.19   4.44
THE SKIPLEVEL IS, 52.00

FOR VALLEY DETECTION A CONTINUITY OF   3 POINTS IS REQUIRED
THE BASIC TIME INTERVAL IS   .50
THE PROGRAM WILL AT A LATER POINT FIND A SUITABLE REFERENCE PEAK
DETAILED PEAK INFORMATION WILL BE PRINTED OUT
THE CHROMATOGRAM WILL BE PLOTTED.
THE BEGIN AND END POINTS OF THE PEAKS WILL BE INDICATED IN THE PLOT.
THE CHROMATOGRAM WILL ALSO BE PLOTTED ON A X 10 SCALE.
THE X AND Y DIMENSIONS OF THE COMPLETE GASCHROMATOGRAM ARE DIVIDED BY   4.000   2.000

Figure 5.9. Page 1 of a typical output of program GCX version 31.

PEAK NUMBER	START TIME	END TIME	WIDTH	PEAK HEIGHT	RETENTION TIME	GRAVITY CENTER	....AREA.... ABSOLUTE	RELATIVE		SIGMA	PLATES	RES.	K'	REMARKS
1	76.5	87.0	10.5	19928	81.762	81.909	149946	49.557	15.677	1.52	2916	5.95	N.A.	NONE
2	87.0	103.5	16.5	1173	90.416	90.938	11870	3.923	1.241	2.57	1256	7.44	.11	NONE
3	103.5	126.5	23.0	150	109.219	110.043	2255	.745	.236	3.43	1031	8.04	.34	NONE
4	126.5	145.0	18.5	32	137.218	137.608	655	.216	.068	3.78	1326	6.82	.68	NONE
5	145.0	223.5	78.5	11800	161.749	163.360	302574	100.000	31.634	6.18	699	13.71	.98	NONE
6	223.5	265.0	41.5	112	247.452	248.116	4488	1.483	.469	7.76	1822	5.91	2.03	NONE
7	269.0	357.0	92.0	4825	291.712	293.986	216985	71.713	22.686	10.19	832	18.21	2.57	NONE
8	443.0	514.0	71.0	109	477.500	479.588	7551	2.496	.790	13.59	1246	9.65	4.84	NONE
9	559.0	726.5	167.5	3244	607.372	610.727	260162	85.983	27.200	17.63	1200	N.A.	6.43	NONE

APPROXIMATE NUMBER OF THEORETICAL PLATES FOR THIS COLUMN   1005

PEAK NUMBER	...RETENTION TIME OF THE PEAKTOP... 1 POINT	3 POINT	5 POINT	7 POINT	9 POINT	11 POINT	15 POINT	OPTIMAL	NET	LOG	RELATIVE RETENTION	RELATIVE AREA
1	82.00	81.762	81.774	81.790	N.A.	N.A.	N.A.	81.762	N.A.	N.A.	N.A.	15.677
2	90.50	90.369	90.383	90.400	90.416	90.424	N.A.	90.416	8.654	.23719	1.6464	1.241
3	109.00	109.150	109.166	109.191	109.219	109.250	109.316	109.219	27.456	1.43664	5.2237	.236
4	137.00	137.250	137.274	137.216	137.194	137.218	137.315	137.218	55.455	1.74394	10.5507	.068
5	161.50	161.713	161.725	161.737	161.749	161.760	161.823	161.749	79.987	1.90302	15.2179	31.634
6	247.00	247.250	247.287	247.323	247.378	247.430	247.452	247.452	165.690	2.21930	31.5234	.469
7	291.50	291.653	291.660	291.667	291.675	291.685	291.712	291.712	209.950	2.32212	39.9440	22.686
8	477.50	477.500	477.500	477.500	477.500	477.500	477.500	477.500	395.738	2.59741	75.2911	.790
9	607.50	607.364	607.361	607.351	607.349	607.360	607.372	607.372	525.610	2.72066	100.0000	27.200

Figure 5.10.   Page 2 and 3 of a typical output of program GCX version 31.

INFORMATION BASED ON CENTERS OF GRAVITY

PEAK NUMBER	....CENTERS OF GRAVITY....		GRAVITY VS PEAK-TOP		
	ABSOLUTE	RELATIVE	DIFF.	DIFF/SIGMA	RATIO
1	81.909	.000	N.A.	+.146	+.0963	1.00170
2	90.938	9.029	1.7074	+.522	+.2032	1.00577
3	110.043	28.134	2.3202	+.824	+.2404	1.00754
4	137.608	55.699	10.5328	+.390	+.1032	1.00284
5	163.360	81.451	12.4025	+1.611	+.2606	1.00996
6	248.116	166.208	31.4300	+.664	+.0856	1.00260
7	293.986	212.077	40.1040	+2.274	+.2231	1.00789
8	479.588	397.680	72.2015	+2.088	+.1537	1.00437
9	610.727	528.819	100.0000	+3.355	+.1903	1.00552

GENERAL STATISTICAL INFORMATION

PEAK NUMBER	GRAVITY CENTER	RELATIVE AREA	........SIGMA........			OVERALL	....SKEW....		CURTOSIS	
			FIRST	SECOND	RATIO					
1	81.91	49.557	1.46	1.58	.925	1.52	+.?	+.2452	+16.9	+3.1852
2	90.94	3.923	2.17	2.88	.755	2.57	+24.9	+1.4744	+298.9	+6.8908
3	110.04	.745	3.16	3.65	.866	3.43	+36.3	+.9012	+639.2	+4.6291
4	137.61	.216	3.89	3.67	1.060	3.78	-5.2	-.0955	+508.2	+2.4730
5	163.36	100.000	5.55	6.64	.836	6.18	+368.4	+1.5607	+14492.4	+9.9332
6	248.12	1.483	8.05	7.59	1.073	7.76	-45.1	-.0966	+9917.4	+2.7352
7	293.99	71.713	9.56	10.68	.894	10.19	+1011.2	+.9556	+60202.0	+5.5828
8	479.59	2.496	13.66	13.52	1.010	13.59	+462.8	+.1846	+89287.2	+2.6211
9	610.73	85.983	16.88	18.23	.926	17.63	+4180.9	+.7627	+476868.1	+4.9341

Figure 5.11.   Page 4 and 5 of a typical output of program GCX version 31.

the average plate number is the first peak.

For t_0, the first peak in the chromatogram higher than the skiplevel is used. For the reference peak, the last peak higher than twice the skiplevel is selected. However, the user may specify any other peak to become reference peak by giving the approximate net retention time. The program then selects a peak that is within 1.5 times s of this time.

5.5 OUTPUT AND RESULTS

The actual output is mostly a matter of formatting. This is so dependent on the computer and the language used, that it is not of general interest to describe the programming in detail. An example of the output of program GCX is given in Figures 5.9 through 5.11.

5.6 INITIALISATION AND DATA INPUT

It will be clear now that at various points the program has options built in which can be user selected. For any option the user has to supply a set of two figures. The first figure is the option number, the second figure gives a corresponding value. For each option the program has a default value which is used if the user does not supply a value or supplies an obviously erroneous value. Table 5.1 lists some of the options, with the default values and the range within which a user may select a value.

Before the list of options, the user has to supply a date of measurement and a run number. If there are more gas chromatograms to be processed then every option list has to be followed by the figure 99, thereby indicating that another run follows. Only the last option list should not be followed by 99.

After entering the so called "condition tape", the

Table 5.1

List of options for program GCX, with default values
and the maximum range.

option number	description	default value	maximum range
1	Follow value (datapoints)	3	1-10
2	Starting time (sec)	0	∞
3	Data interval (sec)	0.5	0.1-100
4	Scale factor for plot	2	1-10
5	Net time for reference peak (sec)	1)	1-∞
6	Threshold level (as fraction of tallest peak)	2)	0-1
7	Smoothing width	5	1-21
13	Type baseline: 1=normal baseline 2=low baseline 5=aut. baseline 10=special baseline	5	1-10
14	Number of datapoints to be used for the special baseline	20	1-∞
16	Minimum peakwidth (datapoints)	6	1-∞
23	Plot the chromatogram	1	b 3)
24	Plot with x10 sensitivity	1	b
25	Indicate peakboundaries in the plot	1	b
29	Punch the results on paper tape	0	b
30	Apply solvent peak skimming	0	b

Notes: 1) Default value: the last peak in the chromato-
gram which is taller than the skiplevel.
2) Default value: threshold level calculated
according to the average below average technique.
3) "b" means: boolean. The value 1 means "true",
and the value 0 means "false".

papertapes with the digitized chromatograms are read in.
The physical end of each tape indicates the end of a chromatogram. The number of tapes with chromatograms must be
equal to the number of option lists.

5.7 MISCELLANEOUS

In addition to the calculations that are mentioned so
far the program performs a few more functions. One of these
functions is to check at various processing stages if a
peak is a real gas chromatographic peak or a spurious peak
caused by drift or noise. A peak will be discarded if any
of the following criteria is met:
- The peakwidth at threshold level is less than three
  datapoints.
- The total peakwidth (as calculated by ACCURATE) is
  less than "narrow". The value of "narrow" is normally
  six datapoints, but may be user selected.
- The peaktop is less than two datapoints from either
  end.
- The plate number of the peak differs more than a
  factor of four from the average plate number.
- The height of the peak, less the average of begin and
  end is less than the threshold level.

The last three criteria are only checked for peaks that
are lower than the skiplevel.

Another function is printing messages at the various
steps of the data processing if anything goes wrong. These
messages are fairly extensive and clear to any gas chromatograph operator. If any error message was printed, then
also a general warning is printed on that part of the output where the operator is most likely to look first. Table
5.2 lists some of the possible messages.

Table 5.2. Selection of computer generated messages, with the associated actions.

message	result
- *The year (or month, or day) is entered incorrectly.*	1,2,A
- *Maybe the sequence month-day was mixed up.*	1,2,A
- *The condition tape ended incorrectly.*	2,A
- *It is not allowed to specify a condition index xxx.*	3,2,A
- *It is not allowed to set a boolean other than by means of a "0" or a "1".*	3,2,A
- *The maximum number of runs is xxx.*	3,2,A
- *The initial baseline was only xxx long, processing of this chromatogram has been terminated.*	4,5,B
- *No front edge could be found in the skimming procedure.*	4,6,B
- *No trailing edge could be found in the skimming procedure.*	4,6,B
- *A new threshold level is calculated in order to lower the number of estimated peaks.*	7,2,C
- *No reference peak could be found.*	2,B
- *A scale factor of xxx does not seem to be useful, and has been changed into 8.*	2,B
- *Something has gone wrong with the input.*	5,B

Resulting action:
1. No date of measurement will be printed.
2. Continue processing data.
3. Skip the rest of the condition tape.
4. This applies only to solvent peak skimming.
5. Start processing the next chromatogram.
6. Process this chromatogram without skimming.
7. Return to the peak detection routines.

Resulting general warning messages:
A. *Something was not right, while setting the conditions.*
B. *Check the full output for detailed error messages.*
C. *The execution time has deteriorated through a program corrected irregularity.*

5.1 LITERATURE REFERENCE

5.1 A. Fozard, J.J. Frances, and A.J. Wyatt, Chromatographia, 5, 130 (1972).

CHAPTER 6
HARDWARE

6.1 INTRODUCTION

An extensive assortment of hardware was used to test
out the theories and software as described in this thesis.
This hardware may be divided into five groups:

1. Gas chromatographs
2. Gas chromatographic columns
3. Data acquisition systems
4. Data handling systems
5. Miscellaneous equipment.

The gas chromatographs were of assorted makes and models
and used a wide selection of columns. For acquisition of
data two systems have been assembled and used. For data
handling two different computers have been used.

All these systems will be coded so that they can
easily be referenced in the other chapters.

6.2 GAS CHROMATOGRAPHS

The different gas chromatographs will be indicated by
a capital G and a running number.

G1. An in-house assembled instrument, consisting of the
following components:
- Becker pressure controllers (Becker, Delft, Holland).
- Becker thermostat type 1452 SH.
- Becker flame ionisation detector type 5002 AC.
- A "homemade" injector block without splitter.
- A "homemade" amplifier based on a Analog Devices

model 310 J parametric amplifier (Analog Devices, Cambridge, Mass.).

G2. Same as G1 except for the thermostat and pressure controllers, these were:
- Colora Ultrathermostat NB 33369 (Colora, Lorch-Württ, Germany).
- Negretti pressure controllers (Negretti and Zambra Ltd., London, England).

G3. Another in-house assembled instrument, consisting of the following components.
- Negretti pressure controllers.
- Becker thermostat type 1452 D.
- Becker flame ionisation detector type 5002 A.
- Atlas electrometer amplifier model DC 60 CH (Atlas MAT, Bremen, Germany).
- A "homemade" injector of the falling needle type (Ref. 6.1).

G4. In-house assembled system composed off:
- Becker pressure controllers.
- Prototype thermostat (Ref. 6.2, 6.3). The main characteristics of this thermostat are a stability of $0.005^{o}C$ and a maximum temperature gradient of $0.01^{o}C$.
- A "homemade" flame ionisation detector as described in Ref. 6.4.
- Injector and amplifier as in G1.

G5. Hewlett-Packard gas chromatograph type 5750 (Hewlett-Packard, Palo-Alto, Cal.). This instrument was modified to accomodate a falling needle injector.

G6. In-house assembled system composed off:
- Tamson double waterbath thermostat type TEVMB 70/230 (Tamson, Zoetermeer, Holland).
- Texas Instruments quartz high precision pressure controller (Texas Instruments Ltd., Bedford, England).
- Becker flame ionisation detector.
- Hamilton splitter inlet (Hamilton, Whittier, Cal.).
- Atlas electrometer amplifier model DC 60 CH.

G7. Hewlett-Packard gas chromatograph type 7600 A equipped
with an electron capture detector.

6.3 COLUMNS

The different kinds of gas chromatographic columns
will be indicated by a capital C and a running number.

C1. Packed column.
   - 4 m long, 2 mm inner diameter, copper.
   - 10% squalane on Anakrom 70/80 mesh (Analabs Inc.,
     North Haven, Conn.).
C2. Packed column.
   - 2 m long, 2 mm inner diameter, copper.
   - 12% squalane on Chromosorb S 100/110 mesh (Applied
     Science Laboratories Inc., State College, Pa.).
C3. Micro-packed column.
   - 15 m long, 0.8 mm inner diameter, glass.
   - 2% squalane on Chromosorb W, acid washed, 180-200μ.
C4. Micro packed column.
   - 10 m long, 0.8 mm inner diameter, glass.
   - 15% squalane on Chromosorb W, 180-200μ.
C5. Capillary column.
   - 20 m long, 0.25 mm inner diameter, glass.
   - SE-30.
C6. Capillary column.
   - 30 m long, 0.25 mm inner diameter, glass.
   - OV-101.
C7. Capillary column.
   - 20 m long, 0.5 mm inner diameter, glass.
   - SE-30.
C8. Capillary column.
   - 25 m long, 0.4 mm inner diameter, glass.
   - SE-30
C9. Capillary column.
   - 100 m long, 0.25 mm inner diameter, stainless steel.
   - Squalane.

6.4 DATA ACQUISITION SYSTEMS

Two data acquisition systems have been used. Both were assembled in-house. A block diagram is shown in Figure 6.1. The two systems will be indicated as D1 and D2.

D1. The main components were:
- Timer: H/P model 5233 L counter/timer (Hewlett-Packard, Palo-Alto, Cal.).
- D.V.M.: Dynamco model DM 2025 digital voltmeter (Digital Measurements Ltd., Mychett, Andershot, Hants, England). The following features were used: 20 msec conversion speed, 20 msec RC filter, ranges 0.19999 V or 1.9999 V.
- Serializer: Dynamco model DM 5021 punch drive unit.
- Punch: Friden model SP 2 paper tape punch (Friden, Rochester, N.Y.), with a maximum speed of 20 char/sec.

D2. In this setup the main features are:
- Timer: assembled from components, the main parts being a crystal controlled oscillator, type B-902 (Ebauches, Neuchatel, Switzerland) and five Delta Electronics model TE 2 counter units (Delta Electronics, Zierikzee, Holland).
- D.V.M.: the same as the one used in D1, except the model number is DM 2020-A and the ranges are 0.39999 V and 3.9999 V.
- Serializer: assembled from components, mainly Fairchild integrated circuits. (Fairchild, Moutain View, Cal.).
- Punch: Facit model 4070 paper tape punch. (Facit AB., Solná, Sweden). The maximum speed is 75 char/sec.

The two data acquisition systems basically operate the same way. After depressing a start button, the timer unit starts sending pulses to the D.V.M. at the preset rate. The pulse interval may be set at 0.2, 0.5, 1, 2, 5, or 10 seconds. Each pulse initiates the D.V.M. to take a reading
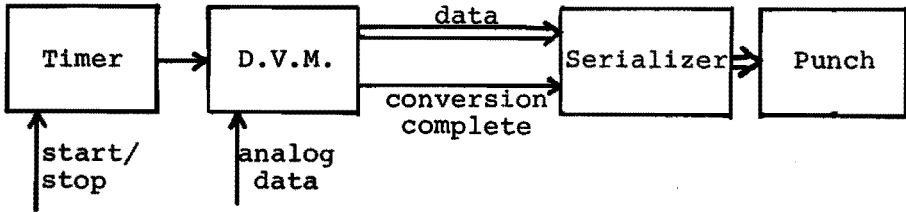
Figure 6.1   Blockdiagram of the data acquisition system.

of the analog signal. When the reading is taken, the D.V.M. sends a conversion-complete signal to the serializer. The measurement is then serialized and punched on paper tape.

In addition to the differences that are listed above, it should be mentioned that D1 cannot operate at the highest speed (0.2 sec. interval) due to the slow punch. System D2 has thumbwheel switches which can be set at a certain time, when this time is reached a stop signal is automatically generated. System D1 can only be stopped by pushing a stop button. The stop signal stops the timer and thereby terminates a GC-run.

In essence system D2 is an updated version of system D1. System D2 is more reliable and more convenient to operate.

6.5 DATA HANDLING SYSTEMS

Data handling has been done by two computer systems. One system was an Electrologica X-8 (Electrologica, Rijswijk, Holland). The other system was a Philips P9200 (Philips Gloeilampenfabrieken, Eindhoven, Holland). The EL X-8 was used for handling all the paper tapes that were produced by the data acquisition system as well as for most other calculations. The P9200 was used for some calculations that were made to substantiate the theories on smoothing.

The EL X-8 is a general purpose computer. It has a
48 K 26 bit word 2.5 μsec core memory. The system has a
drum, three magnetic tape drives, three paper tape punches,
three paper tape readers, a plotter and a line printer.
The system runs under a monitor that allows full multi-
programming and uses the drum as virtual core memory. All
programs for the EL X-8 were written in ALGOL 60.

The P9200 is a timesharing system. It consists of a
P9205 as communication computer and a P9202 central com-
puter with a 32 K 16 bit word 0.96 μsec core memory. The
peripheral equipment comprises a line printer, two disk
drivers, a paper tape reader, and a paper tape punch. Most
programs for the P9200 were written in ALGOL 60 and a few
in BASIC.

6.6 MISCELLANEOUS EQUIPMENT.

In addition to the components and systems that are
described so far, it may be worthwhile to mention the
following points.
- Samples were injected with Hamilton syringes
  (Hamilton, Whittier, Cal.) or S.G.E. syringes
  (Scientific Glass Engineering PTY Ltd., Melbourne,
  Victoria, Australia.)
- Nitrogen was always used as carrier gas, except for
  the long glass micro-packed columns which were run
  with hydrogen.
- Nitrogen and hydrogen were supplied by Loosco
  (Loosco, Amsterdam, Holland).

6.7 LITERATURE REFERENCES

6.1 P.M.J. van den Berg and Th.P.H. Cox, Chromatographia, 5, 301 (1972).

6.2 J.H.M. Wijtvliet, Report on Automation of a Gas Chromatograph, Technical High School, Heerlen, The Netherlands, 1972, Chapter 4.

6.3 G.L.M. Ramaekers, Graduation Report Eindhoven University of Technology, Eindhoven, The Netherlands, 1972, Chapter 4.

6.4 J.F.J. Krugers, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1964, Chapter 1.

CHAPTER 7

RESULTS

7.1 INTRODUCTION

The data handling program GCX has been extensively tested. A total of approximately 500 chromatograms has been processed. In addition a series of computer generated chromatograms have been tested. A few experiments showed, that it was entirely feasible to process data from other sources, e.g. NMR and mass spectrometry.

The single most important result is that the program never really failed. The program responded very well on errors in the input data. Very often errors were fully corrected. In all other cases the program skipped the error or stopped. Virtually any irregularity was reported through some message in the output. In most cases these messages did not leave the user any doubts as to the actual source of the irregularity.

The major goal of the computer program was the accurate determination of peaktop locations. It can be said now that this goal has been fully reached. The obtained accuracies surpassed all expectations. It was partially surprising that the reproducibility of retention times was much better than was thought to be possible.

The center of gravity of peaks could be calculated with the same accuracy as the peaktops. However this was done with gas chromatograms with very good baselines and well separated peaks. Any deviations in the baseline have their immediate effect on the center of gravity. Also poor resolution and inaccurate peak boundaries impair the accuracy.

The other quantitative results underscored the statement that the baseline is the most important parameter to

be established. A slight error in the baseline determination shows up strongly in the calculations of the area, skew, and kurtosis. Also the accuracy of the peak boundaries affects the results (Ref. 7.1).

Poor resolution is entirely a matter of gas chromatography and not of data handling. Therefore the effects of poor resolution on the accuracies of the various gas chromatographic properties will not be discussed here.
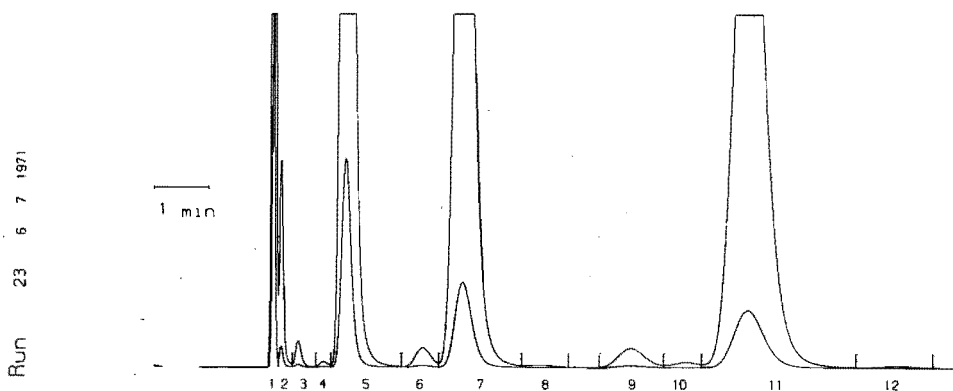

Figure 7.1.   Chromatogram of the hydrocarbons methane through hexane, after processing by program GCX. The small peaks stem from impurities in the sample.
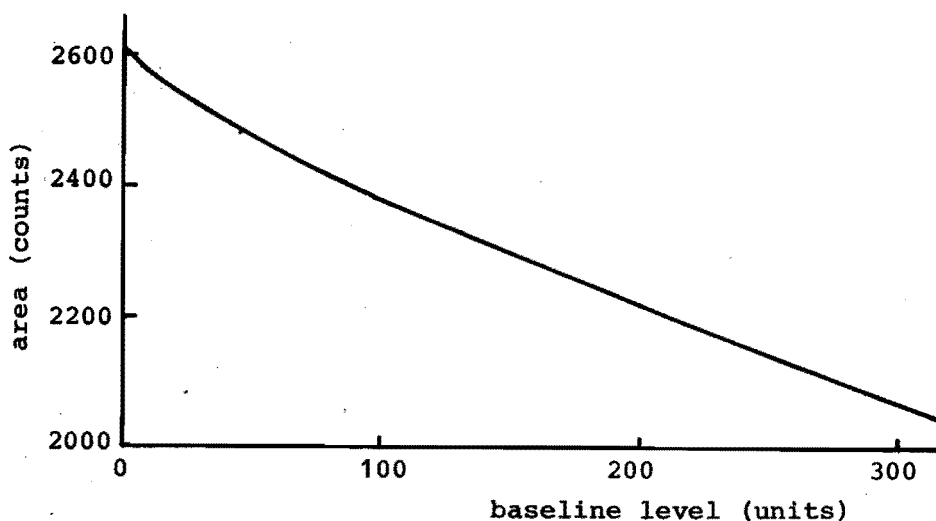

Figure 7.2.   Area as a function of the baseline level. The peakheight is 3246 units.

7.2 INFLUENCE OF THE BASELINE

It is difficult to give general data concerning sensitivity of the center of gravity, area, sigma, skew and kurtosis to an inaccurate baseline. For a symmetrical peak the center of gravity and skew are completely independent of the baseline. For asymmetrical peaks all properties depend on the type and the extent of the asymmetry, both of which can widely change.

It is still desirable to have an idea about the effects. Therefore a chromatogram was taken from the series of Section 7.4. This chromatogram is shown in Figure 7.1. The properties of the hexane peak (peak nr. 11) were then calculated as a function of the baseline. The results are shown in Figures 7.2 through 7.6. All quantities are given as they are calculated by program GCX. The peak height was 3246 units. The peaktop was at 607.5 sec.
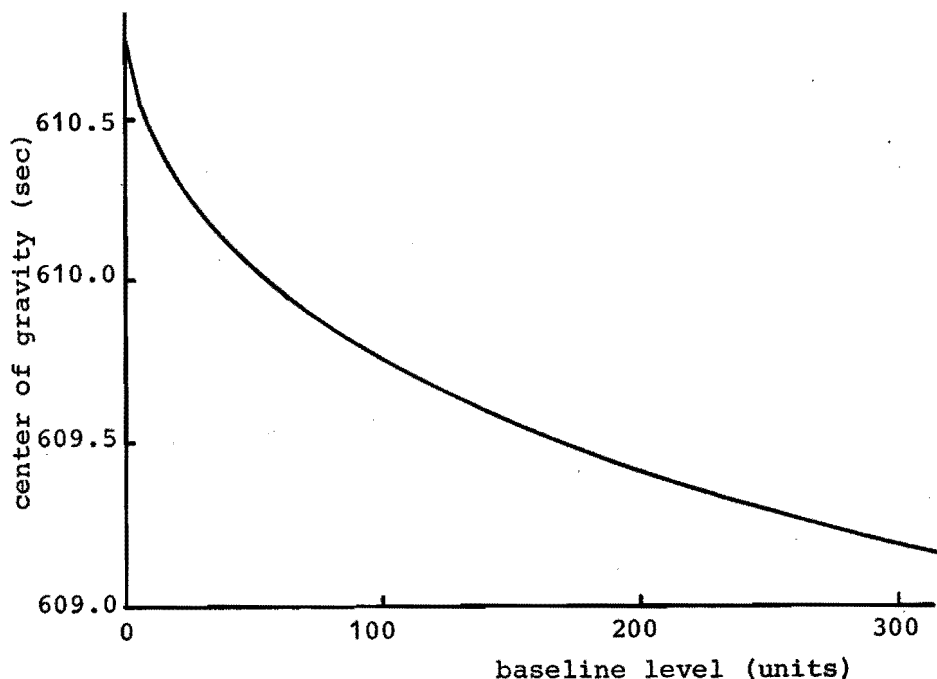

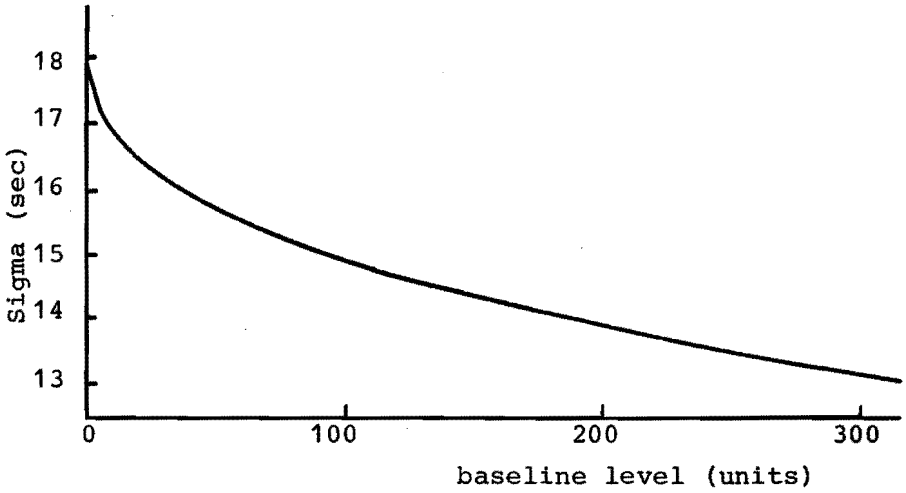Figure 7.3. Center of gravity as a function of the baseline level. The peakheight is 3246 units.

Figure 7.4. Sigma as a function of the baseline level.
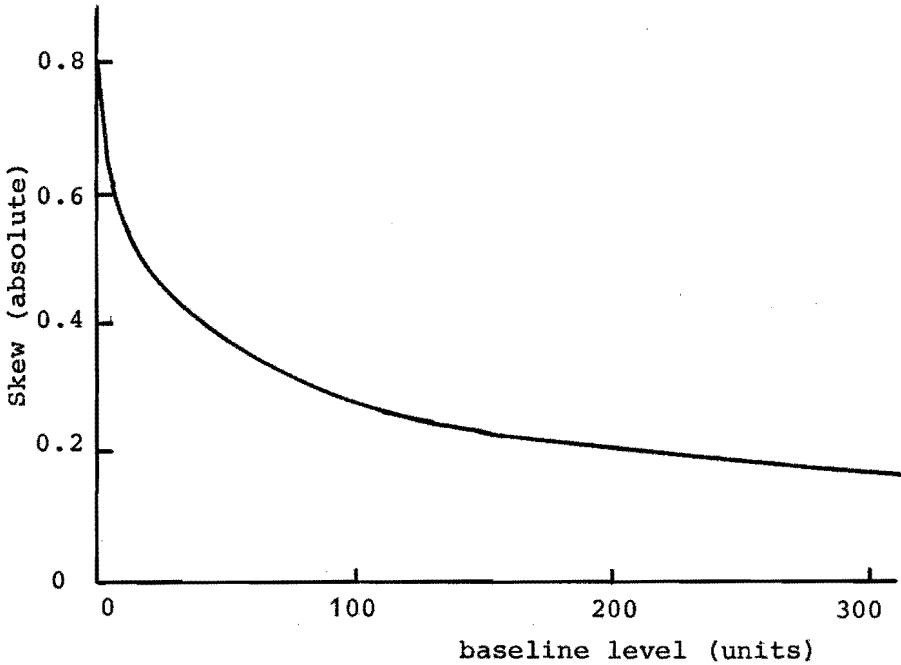The peakheight is 3246 units.

Figure 7.5. Skew as a function of the baseline level.
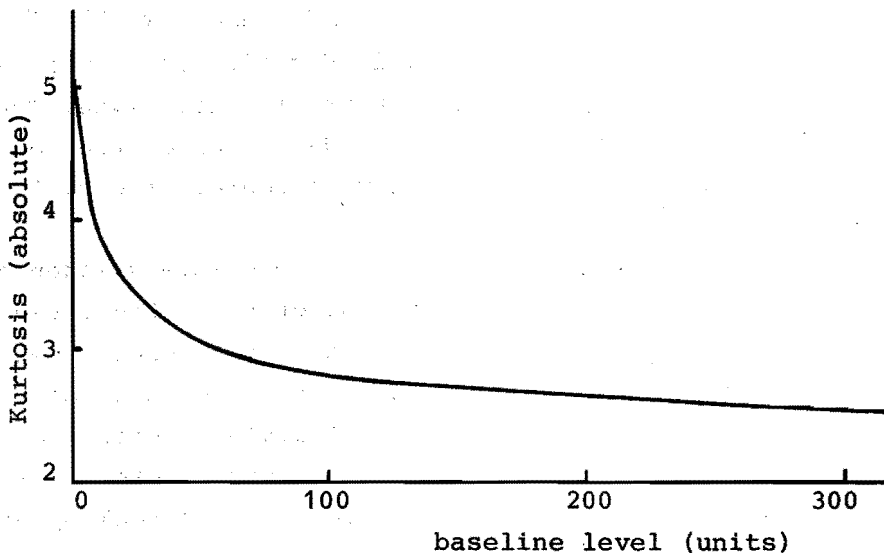The peakheight is 3246 units.

Figure 7.6. Kurtosis as a function of the baseline level. The peakheight is 3246 units.

7.3 RETENTION TIME ACCURACY

The retention time can be obtained from the peaktop or from the center of gravity. The influence of noise on the peaktop is shown extensively in Chapter 4. The influence of inaccurate peak boundaries on the center of gravity is shown in Section 2.5. Section 7.2 showed how the center of gravity depends on the baseline.

Practical measurements were made with gas chromatograph G1 and column C2. Data acquisition system D1 was used with a 0.5 sec data interval rate. With this equipment the reproducibility of the peaktop and the center of gravity was measured. This was done with a sample that consisted of a vapour mixture of the normal alkanes methane through hexane. This sample was run ten times.

The gas chromatograph was an older model with short term temperature fluctuations of $\pm 0.2^{\circ}C$. This made it necessary to use relative retentions, rather than retention times.

The results are shown in Tables 7.1 and 7.2. The first table shows the average and standard deviation of the peaktop. In addition some other typical quantities are given. Table 7.2 shows the average and standard deviation of the center of gravity and the difference between center of gravity and peaktop.

The calculation of the peaktop is completely independent of the calculations of the center of gravity. Consequently the errors are unrelated and hence the difference between center of gravity and peaktop contains the errors combined. Since the errors are unrelated the individual errors must be smaller than the combined error. In other words: the standard deviation in the calculations of the peaktop and center of gravity must be smaller than the standard deviation of the difference between the two. Table 7.1 and 7.2 however show the opposite. This is due to the fact that the gas chromatograph itself is not sufficiently constant. This is indicated best from the net retention time of the last peak.

Table 7.1. Reproducibility of the peaktop, calculated over ten chromatograms.

peak nr.	compound	relative retention			average			
		average	std. dev.		height	ret. time	area	sigma
			abs.	rel. (%)	(units)	(sec)	(units)	(sec)
1	$n-C_1$	0	–	–	16290	91.12	14100	1.76
2	$n-C_2$	1.620	0.001	0.07	1010	100.84	1140	2.46
3	$n-C_3$	5.131	0.001	0.02	3015	121.91	5180	3.97
4	$n-C_4$	15.117	0.004	0.03	4525	181.83	12770	6.61
5	$n-C_5$	39.780	0.008	0.02	10620	238.70	52940	11.74
6	imp.	75.67	0.03	0.04	130	454.06	1010	14.9
7	$n-C_6$	100	–	–	9110	691.17	81190	19.9

Table 7.2. Reproducibility of the center of gravity and the difference between peaktop and center of gravity.

peak nr.	compound	relative retention			difference between peaktop and center of gravity		
		average	std. dev.		average	std. dev.	
			abs.	rel.		abs.	rel.
1	$n-C_1$	0	-	-	0.162	0.008	4
2	$n-C_2$	1.661	0.01	0.6	0.472	0.06	13
3	$n-C_3$	5.186	0.001	0.02	0.68	0.01	1.6
4	$n-C_4$	15.265	0.005	0.04	1.60	0.02	1.5
5	$n-C_5$	39.952	0.007	0.02	2.63	0.04	1.5
6	imp.	75.53	0.04	0.05	2.1	0.2	11
7	$n-C_6$	100	-	-	3.76	0.05	1.3

7.4 INFLUENCE OF THE PEAKSIZE

From the point of view of data handling the peaksize does not affect any of the calculated properties. However in actual gas chromatography it is well known that the peaksize has its influence on the results (Ref. 7.1). The influence of the peaksize has been evaluated with equipment as described in Section 7.3. The samples consisted of methane, ethane, butane, hexane, and varying amounts of pentane. The pentane concentration ran from 1 to 95%.

The results are plotted in the Figures 7.7 through 7.12. The peaktop and center of gravity are given relative to hexane. The actual net retention time of pentane was approximately 209 sec.

Through the large number of experiments it is also possible here to get an impression of the reproducibility of the experiments.
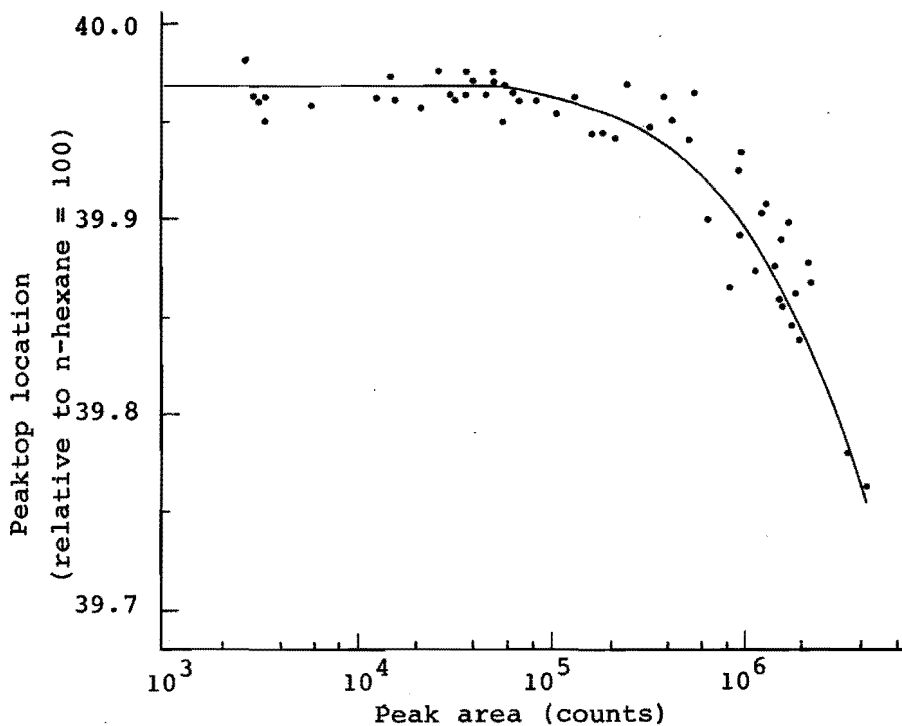
Figure 7.7.   Peaktop location as a function of the peak size for pentane.
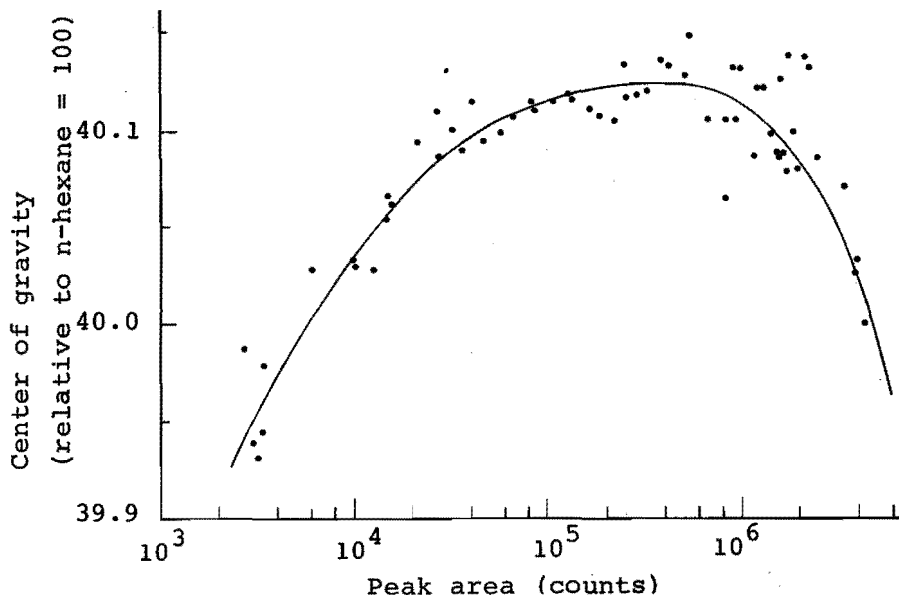

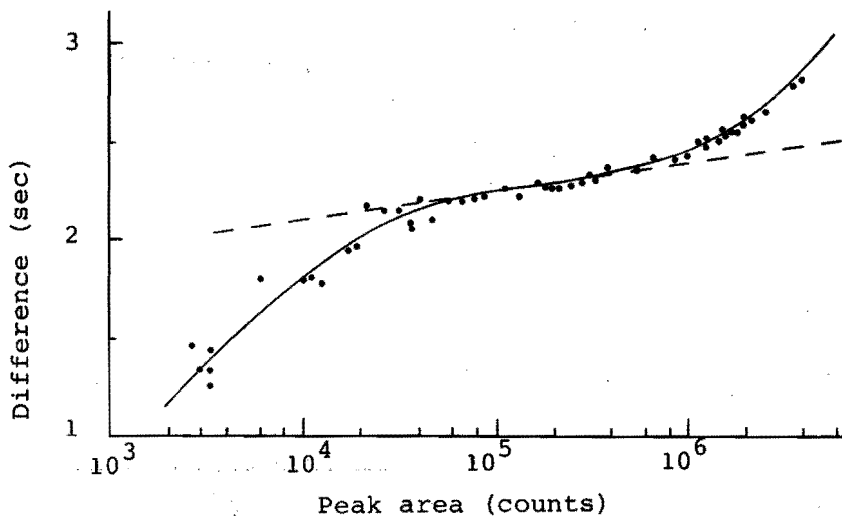Figure 7.8.   Center of gravity as a function of the peak size for pentane.

Figure 7.9. The difference between the peaktop location and the center of gravity as a function of the peak size. for pentane.
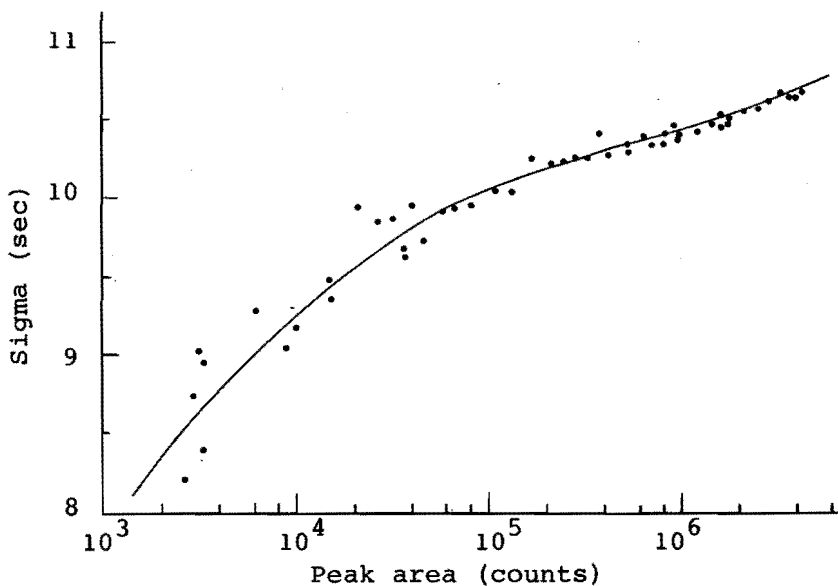

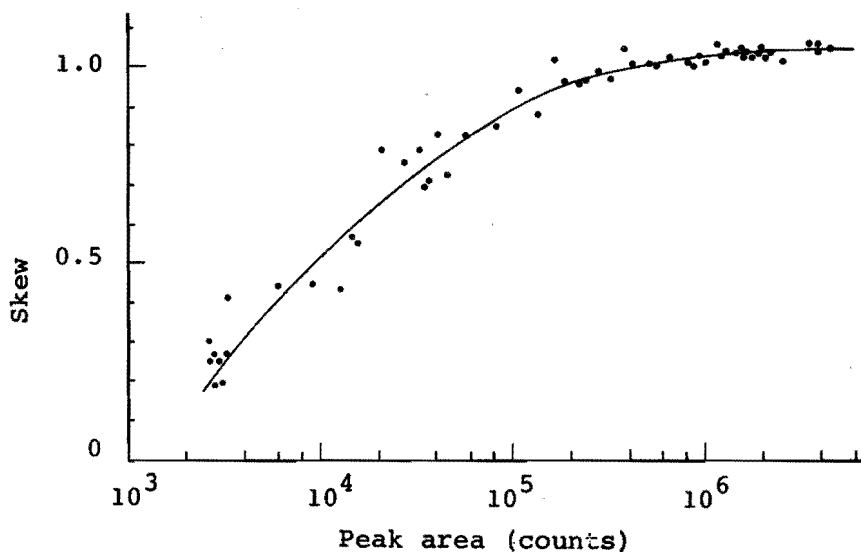Figure 7.10. The peaksigma as a function of the peak size for pentane.

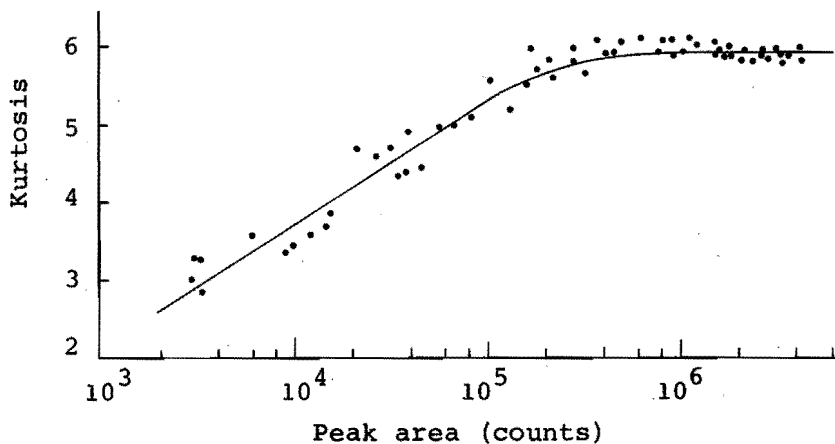Figure 7.11. The skew as a function of the peak size
for pentane.

Figure 7.12. The kurtosis as a function of the peak size
for pentane.

7.5 GENERAL APPLICATIONS

The computer program GCX has been briefly tested on its possible applicability in other fields of data handling. Notably NMR and mass spectrometry have been tried here. The results of these tests are quite good. The average below average technique for determining baseline and threshold level worked excellently. Also smoothing, peaktop allocation, and other calculations worked well without any alterations.

7.5.1 NMR data handling

Program GCX was called upon in order to improve the accuracy of peak integration. The actual experiment was carried out on a Varian A-60 NMR. Data acquisition system D2 was used. A part of a typically processed spectrum is shown in Figure 7.13. The accuracy of the area calculation was better than one percent.
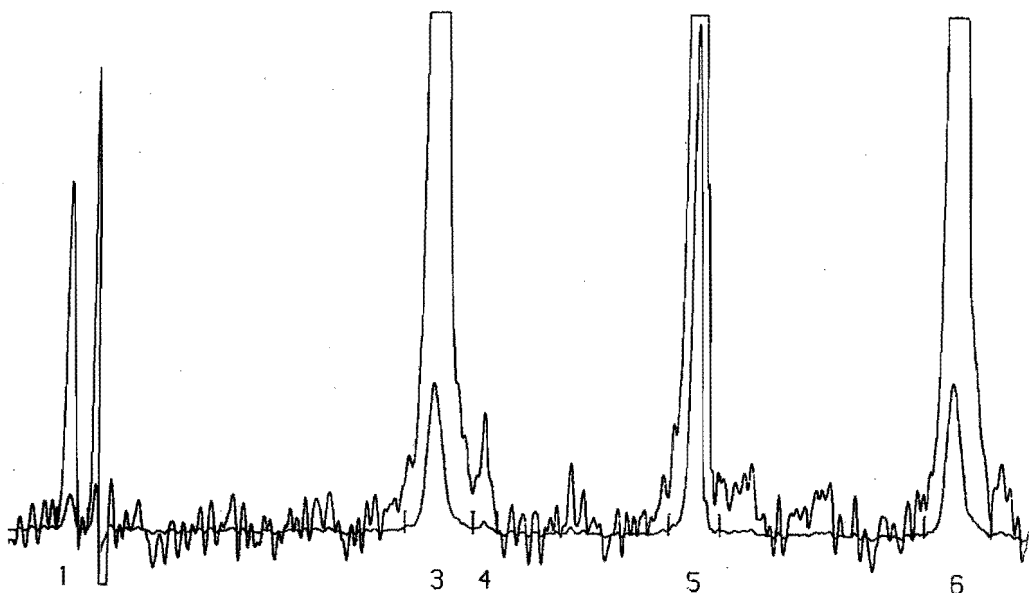

Figure 7.13. Part of a NMR spectrum after processing by program GCX.

7.5.2 Mass spectrometry data handling

The hardware for this experiment consisted of a A.E.I.
MS 12 mass spectrometer with data acquisition system D2. A
partial spectrum as obtained with this setup is shown in
Figure 7.14. The mass range is from 80 m/e through 100 m/e.
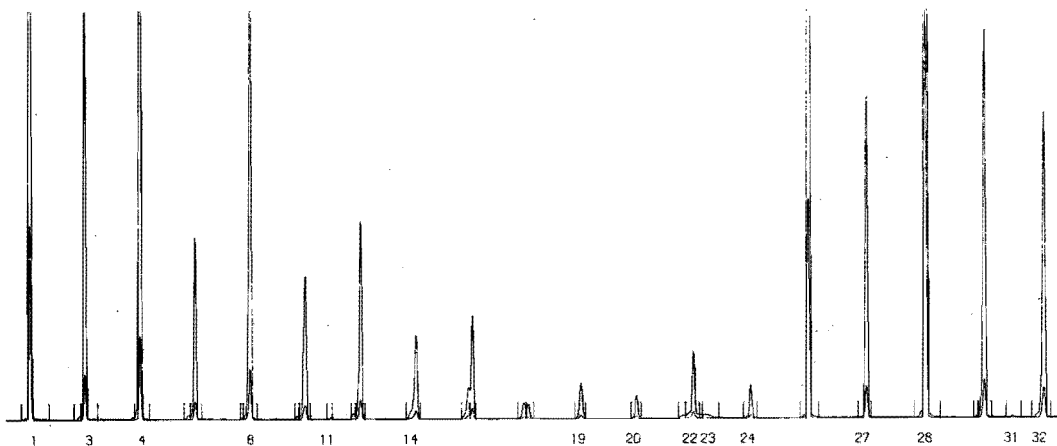The digitizing speed was five samples per second.

Figure 7.14. Partial mass spectrum processed by program GCX.

7.6 COMPUTER RUN TIMES

The total run time of program GCX consists of compi-
lation and execution time. The compilation takes about
four minutes. The execution time depends on the number of
datapoints, the number of peaks, the type of chromatogram,
the requested options, and the encountered irregularities.
Since there are so many factors involved it is not possible
to give very exact times. The average results are shown in
Table 7.3.

Table 7.3. Execution times of various stages of program
GCX (in seconds).

	fixed	per 1000 points	per 10 peaks
smoothing	-	1.4	-
spike removing	-	0.8	-
skimming	0.5	1.6	-
level calculation	0.7	0.4	-
procedure ESTIMATE	1	1.6	1
procedure ACCURATE	-	-	0.4
noise calculations	0.2	1	-0.3
conditions checking	1	-	-
print results	0.8	-	2.5
plot results	8	7	3

7.7 CONCLUSIONS

Some of the conclusions are already mentioned in this
chapter. The main conclusion being that program GCX works
well. This may be illustrated by the Figures 7.15 and 7.16,
which show a noisy steroid chromatogram without and with
peak skimming. The conclusion that GCX works well is sub-
stantiated by the successful processing of large numbers
of different chromatograms. For running these different
chromatograms, all the equipment mentioned in Chapter 6
has been used. In addition chromatograms have been run on
other instruments. Thereby using katharometers, flame
detectors, and electron capture detectors. The samples
that were analysed comprised air, hydrocarbons from C_1
through C_{32}, steroids, pesticides, and many others. The
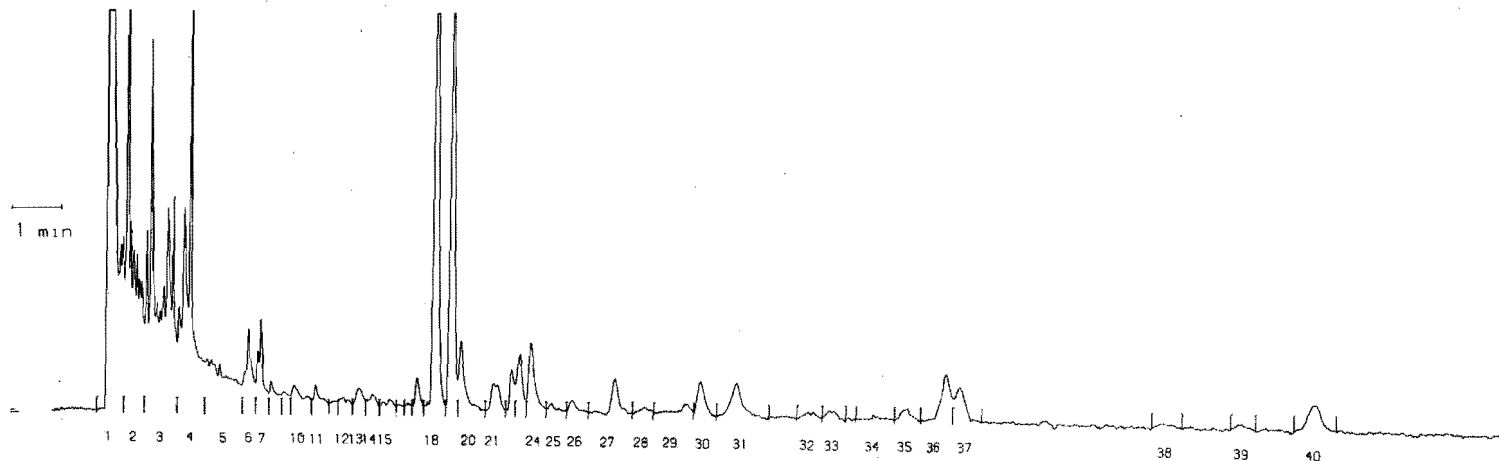columns were packed, capillary, or micro-packed and the

Figure 7.15. Typical steroid chromatogram processed by program GCX, without solvent peak skimming and without smoothing.
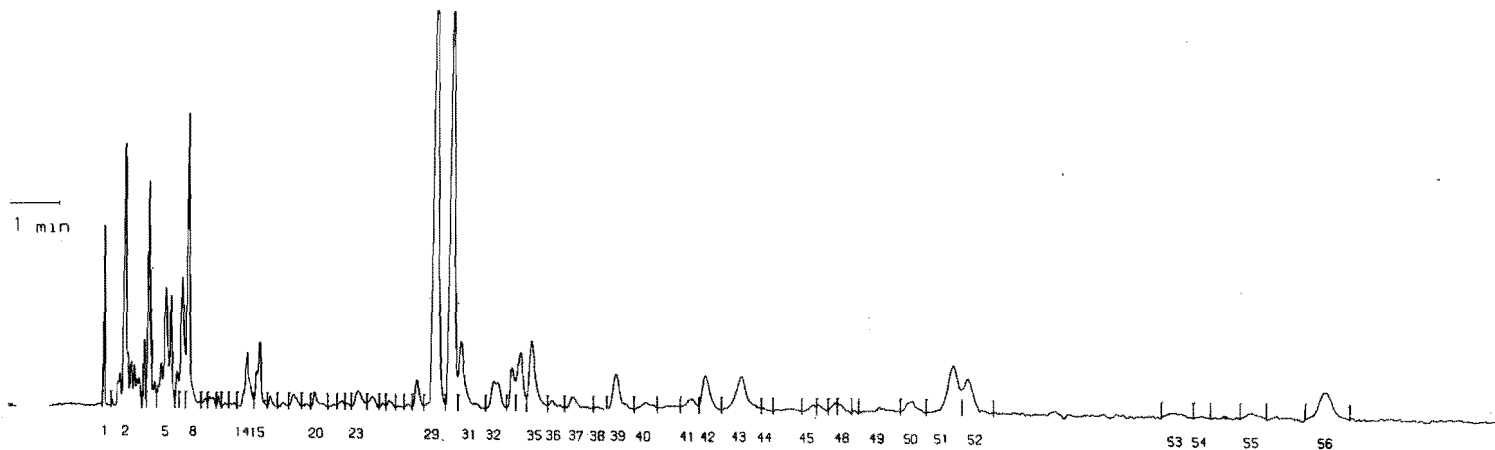

Figure 7.16. The same chromatogram as in Figure 7.15, however after solvent peak skimming and 15 point smoothing.

stationary phases were polar or non-polar liquids or molecular sieves.

This work was particularly aimed at obtaining high accuracy retention times. This goal was fully met with reproducibilities down to the 0.02% level for simple packed columns. The overall results are also in good agreement with the work from Goedert and Guiochon (Ref. 7.1, 7.2).

Program GCX has only a few shortcomings left. One is the limited compatibility with programmed gas chromatography. Another one is the separation of unresolved peaks by simple dropping a perpendicular. With some expansions of GCX the first problem can be overcome. A good general solution for the second problem does not exist yet.

The fact that program GCX easily handles NMR and mass spectrometry data shows that its scope is much wider than gas chromatography alone. The program or parts of it may be applied to curve evaluation in other fields as well. These applications then might include.
- Liquid chromatography
- Electrophoresis
- X-ray fluorescence
- Spectrometric methods, e.g. UV, IR, ESCA, etc.

7.8 LITERATURE REFERENCES

7.1 M. Goedert and G. Guiochon, Chromatographia, in press.
7.2 M. Goedert and G. Guiochon, J. Chromatog. Sci., in press.

SUMMARY

Since its first uses, gas chromatography has expanded tremendously. In 20 years it has grown from virtual non existence to the single most important analytical method. With today's advanced instrumentation data handling has increasingly become the main bottle neck for futher expansions. Fortunately there are now suitable computers available to solve this problem.

Data handling by means of computers has two aspects: hardware and software. For the scientist the hardware is largely a matter of money, since there is a large range of choice equipment available. The software determines what the hardware can do. The software is not so easily obtained and is certainly not always adequate.

Chromatographic data handling can be separated into a series of functions to be performed. The first function is baseline determination. The baseline can be represented by a straight or curved line, either over the whole chromatogram or part there of. The most simple solution is a straight horizontal line and this method works well if the gas chromatograph operates properly.

Another function is the peak detection. This may be done using a threshold level or using derivates. Both methods have their own advantages. The threshold level is somewhat easier to use and less sensitive to noise.

Accurate peaktop location is an interesting feature of automatic data handling. Fitting a parabola through the top data points of a Gaussian curve gives a systematic error that is small. This error can be compensated by applying a simple correction factor. Noise and asymmetry both impair the accuracy. For the best results noise requires a large number, and asymmetry a small number of data points. Therefore an optimum has to be found.

Noise limits the accuracy of all calculations. Noise may be reduced by applying smoothing routines. Spikes however should be removed by a special spike removing routine. Otherwise spikes still appear as small peaks.

When writing a computer program it is necessary to first specify all the design goals. Primarily the program should be easy to use, fail safe, and fool proof. In addition to performing the functions that are mentioned sofar, the program should be capable of the following:

- Removing a solvent peak.
- Consisely reporting all of the useful information that has been abstracted from the chromatogram.
- Issuing warnings for abnormal conditions.

Practice has shown that a correctly written program is indeed fail safe. Retention times may be obtained with an error of only a few milliseconds. Under these circumstances it is interesting to study the effects of sample size on retention time, peak sigma etc. Also the influence of the baseline level can be studied. The results indicate that particularly the higher central moments are very sensitive to changes in sample size or baseline level.

A data handling program as described in this thesis is not limited to gas chromatography alone. Virtually any method that yields a curve superimposed on a defined baseline can be processed. This has been shown in practice for NMR and mass spectrometry.

ACKNOWLEDGMENT

BIOGRAPHICAL NOTICE

1941   Born in The Hague.

1959   Finished high school (HBS-b) in Heerlen.
        Started studying chemistry at the Eindhoven
        University of Technology.

1962   Accepted a part-time job as a research
        assistant at the Eindhoven University of
        Technology.

1966   Obtained the masters degree in chemical
        engineering at the Eindhoven University of
        Technology.

1967   Obtained a one year sabbatical leave to
        become a research associate at Baylor
        College of Medicine in Houston, Texas.

1968   Extended the sabbatical leave for a brief
        assignment as a research associate at MIT.
        Returned to the Eindhoven University of
        Technology.

1972   Accepted a new job as systems manager for
        Techmation NV at Amsterdam.

STELLINGEN

1- Piekdeconvolutie in de gaschromatografie kan pas tot
een optimaal resultaat leiden indien een correct piek-
model is opgesteld, in plaats van de gebruikelijke
Gauss curves.

2- Voor biomedische toepassingen heeft vloeistofchroma-
tografie op theoretische gronden voordelen boven de
gaschromatografie. Er zal echter eerst een equivalent
aan de combinatie GC-MS-computer gevonden moeten wor-
den, alvorens de vloeistofchromatografie de gaschro-
matografie kan overvleugelen.

Lit.: E. Jellum, O. Stokke en L. Eldjarn, Clin. Chem.,
18, 800 (1972).

3- Een ieder die een uitgebreid computerprogramma schrijft
teneinde een technisch probleem op te lossen is daar-
mee nog geen mathematicus, maar noodzakelijkerwijs wel
een experimentator.

4- De voorgestelde kop-staart configuratie van de $(2M+1)^+$
ionen in de chemische ionisatie massaspectra van
aminozuren (1) is in tegenspraak met de gemeten hoge
entropie van die ionen (2).

(1) P.A. Leclercq en D.M. Desiderio, Org. Mass
Spectrom., in druk.
(2) F.H. Field, J. Amer. Chem. Soc. 91, 2827 (1969).

5- Na 20 jaar ontwikkeling in de gaschromatografie is de
thermostaat nog steeds een stiefkind. Dit verandert
pas indien men afstapt van het idee dat de thermische
massa van de oven persé klein moet zijn. Het is daar-
entegen veel belangrijker dat het intern warmtetrans-
port zo homogeen en groot mogelijk is.

6- Na het verschijnen van het vaak geciteerde artikel van
   Savitsky en Golay (1) heeft het acht jaren geduurd tot
   de uitgebreide correcties gepubliceerd werden (2).
   Hieruit blijkt dat zelfs bij artikelen van gerenommeer-
   de auteurs de nodige prudentie in acht genomen moet
   worden met betrekking tot de uitkomsten.

   (1) A. Savitsky en M.J.E. Golay, Anal. Chem., $\underline{36}$, 1627
       (1964).
   (2) J. Steinier, T. Termonia, J. Deltour, Anal. Chem.,
       $\underline{44}$, 1906 (1972).

7- Wetenschapsbeoefenaars blijken niet alleen in staat te
   zijn de statistiek, doch ook de economie te misbruiken
   om datgene aan te tonen dat ze graag zouden willen aan-
   tonen.

8- Een extrapolatie van de huidige trend in de landelijke
   uitgaven voor sociale voorzieningen zal eerder dan die
   voor het onderwijs het niveau van 100% van het natio-
   naal inkomen bereiken.

9- Een overheid die niet weet te verhinderen dat gevaar-
   lijke hoeveelheden lood en cadmium voorkomen in aarde-
   werk serviesgoed kan moeilijk volhouden de problemen
   van de volksgezondheid integraal te overzien.

   Lit.: Consumentengids $\underline{20}$, 382 (1972).

Eindhoven, 1 december 1972.   J.J.M. Wijtvliet.