

**Decizie de indexare a faptei de plagiat la poziția
00059 / 07.07.2013
și pentru admitere la publicare în volum tipărit**

care se bazează pe:

A. Nota de constatare și confirmare a indiciilor de plagiat prin fișa suspiciunii inclusă în decizie.

Fișa suspiciunii de plagiat / Sheet of plagiarism's suspicion	
Opera suspicionată (OS) Suspicious work	Opera autentică (OA) Authentic work
OS	BOJIȚĂ, M.; ROMAN, L.; SÂNDULESCU, R. și OPREAN, R. <i>Analiza și controlul medicamentelor, vol.2: Metode instrumentale în analiza și controlul medicamentelor</i> . Deva: Intelcredo. 2003.
OA	BEEBE, K.R.; PELL, R.J. and SEASHOLTZ, M.B. <i>Chemometrics: A Practical Guide</i> . John Willy & Sons, 1998.
Incidența minimă a suspiciunii / Minimum incidence of suspicion	
p.677:20 – p.694:21	p.27:01 – p.44:27
p.678: Figura 16.1	p.28: Figure 3.1
p.679: Figura 16.2	p.29: Figure 3.2
p.681: Figura 16.3	p.31: Figure 3.3
p.682: Figura 16.4	p.32: Figure 3.4
p.683: Figura 16.5	p.33: Figure 3.5
p.683: Figura 16.6	p.34: Figure 3.6
p.684: Figura 16.7	p.34: Figure 3.7
p.685: Figura 16.8	p.35: Figure 3.8
p.686: Tabelul 16.1	p.36: Table 3.2
p.687: Figura 16.9	p.37: Figure 3.9
p.687: Figura 16.10	p.38: Figure 3.10
p.688: Figura 16.11	p.39: Figure 3.11
p.688: Tabelul 16.2	p.40: Table 3.3
p.690: Figura 16.12	p.40: Figure 3.12
p.690: Figura 16.13	p.41: Figure 3.13
p.692: Figura 16.14	p.42: Figure 3.14
p.693: Figura 16.15	p.43: Figure 3.15
p.694: Figura 16.16	p.45: Figure 3.16
Fișa întocmită pentru includerea suspiciunii în Indexul Operelor Plagiate în România de la Sheet drawn up for including the suspicion in the Index of Plagiarized Works in Romania at www.plagiate.ro	

Notă: Prin „p.72:00” se înțelege paragraful care se termină la finele pag.72. Notăția „p.00:00” semnifică până la ultima pagină a capitolului curent, în întregime de la punctul inițial al preluării.

Note: By „p.72:00” one understands the text ending with the end of the page 72. By „p.00:00” one understands the taking over from the initial point till the last page of the current chapter, entirely.

B. Fișa de argumentare a calificării de plagiat alăturată, fișă care la rândul său este parte a deciziei.

Echipa Indexului Operelor Plagiate în România

Fișa de argumentare a calificării

Nr. crt.	Descrierea situației care este încadrată drept plagiat	Se confirmă
1.	Preluarea identică a unor pasaje (piese de creație de tip text) dintr-o operă autentică publicată, fără precizarea întinderii și menționarea provenienței și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
2.	Preluarea a unor pasaje (piese de creație de tip text) dintr-o operă autentică publicată, care sunt rezumate ale unor opere anterioare operei autentice, fără precizarea întinderii și menționarea provenienței și însușirea acestora într-o lucrare ulterioară celei autentice.	
3.	Preluarea identică a unor figuri (piese de creație de tip grafic) dintr-o operă autentică publicată, fără menționarea provenienței și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
4.	Preluarea identică a unor tabele (piese de creație de tip structură de informație) dintr-o operă autentică publicată, fără menționarea provenienței și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
5.	Republicarea unei opere anterioare publicate, prin includerea unui nou autor sau de noi autori fără contribuție explicită în lista de autori	
6.	Republicarea unei opere anterioare publicate, prin excluderea unui autor sau a unor autori din lista inițială de autori.	
7.	Preluarea identică de pasaje (piese de creație) dintr-o operă autentică publicată, fără precizarea întinderii și menționarea provenienței, fără nici o intervenție personală care să justifice exemplificarea sau critica prin aportul creator al autorului care preia și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
8.	Preluarea identică de figuri sau reprezentări grafice (piese de creație de tip grafic) dintr-o operă autentică publicată, fără menționarea provenienței, fără nici o intervenție care să justifice exemplificarea sau critica prin aportul creator al autorului care preia și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
9.	Preluarea identică de tabele (piese de creație de tip structură de informație) dintr-o operă autentică publicată, fără menționarea provenienței, fără nici o intervenție care să justifice exemplificarea sau critica prin aportul creator al autorului care preia și însușirea acestora într-o lucrare ulterioară celei autentice.	✓
10.	Preluarea identică a unor fragmente de demonstrație sau de deducere a unor relații matematice care nu se justifică în regăsirea unei relații matematice finale necesare aplicării efective dintr-o operă autentică publicată, fără menționarea provenienței, fără nici o intervenție care să justifice exemplificarea sau critica prin aportul creator al autorului care preia și însușirea acestora într-o lucrare ulterioară celei autentice.	
11.	Preluarea identică a textului (piese de creație de tip text) unei lucrări publicate anterior sau simultan, cu același titlu sau cu titlu similar, de un același autor / un același grup de autori în publicații sau edituri diferite.	
12.	Preluarea identică de pasaje (piese de creație de tip text) ale unui cuvânt înainte sau ale unei prefețe care se referă la două opere, diferite, publicate în două momente diferite de timp.	

Notă:

a) Prin „proveniență” se înțelege informația din care se pot identifica cel puțin numele autorului / autorilor, titlul operei, anul apariției.

b) Plagiatul este definit prin textul legii¹.

„...plagiul – expunerea într-o operă scrisă sau o comunicare orală, inclusiv în format electronic, a unor texte, idei, demonstrații, date, ipoteze, teorii, rezultate ori metode științifice extrase din opere scrise, inclusiv în format electronic, ale altor autori, fără a menționa acest lucru și fără a face trimitere la operele originale...”.

Tehnic, plagiul are la bază conceptul de **piesă de creație** care²:

„...este un element de comunicare prezentat în formă scrisă, ca text, imagine sau combinat, care posedă un subiect, o organizare sau o construcție logică și de argumentare care presupune niște premise, un raționament și o concluzie. Piesa de creație presupune în mod necesar o formă de exprimare specifică unei persoane. Piesa de creație se poate asocia cu întreaga operă autentică sau cu o parte a acesteia...”

cu care se poate face identificarea operei plagiata sau suspicioate de plagiat³:

„...O operă de creație se găsește în poziția de operă plagiată sau operă suspicioată de plagiat în raport cu o altă operă considerată autentică dacă:

- Cele două opere tratează același subiect sau subiecte înrudite.
- Opera autentică a fost făcută publică anterior operei suspicioate.
- Cele două opere conțin piese de creație identificabile comune care posedă, fiecare în parte, un subiect și o formă de prezentare bine definită.
- Pentru piesele de creație comune, adică prezente în opera autentică și în opera suspicioată, nu există o menționare explicită a provenienței. Menționarea provenienței se face printr-o citare care permite identificarea piesei de creație preluate din opera autentică.
- Simpla menționare a titlului unei opere autentice într-un capitol de bibliografie sau similar acestuia fără delimitarea întinderii preluării nu este de natură să evite punerea în discuție a suspiciunii de plagiat.
- Piesele de creație preluate din opera autentică se utilizează la construcții realizate prin juxtapunere fără ca acestea să fie tratate de autorul operei suspicioate prin poziția sa explicită.
- În opera suspicioată se identifică un fir sau mai multe fire logice de argumentare și tratare care leagă aceleași premise cu aceleași concluzii ca în opera autentică...”

¹ Legea nr. 206/2004 privind buna conduită în cercetarea științifică, dezvoltarea tehnologică și inovare, publicată în Monitorul Oficial al României, Partea I, nr. 505 din 4 iunie 2004

² ISOC, D. Ghid de acțiune împotriva plagiatului: bună-conduită, prevenire, combatere. Cluj-Napoca: Ecou Transilvan, 2012.

³ ISOC, D. Prevenitor de plagiat. Cluj-Napoca: Ecou Transilvan, 2014.

Chemometrics: A Practical Guide

KENNETH R. BEEBE
RANDY J. PELL
MARY BETH SEASHOLTZ
The Dow Chemical Company

149412
16604

BIBLIOTECA CENTRALĂ U.N.F. CLUJ



062117F 0159



A Wiley-Interscience Publication

JOHN WILEY & SONS, INC.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. (∞)

Copyright © 1998 by John Wiley & Sons, Inc. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted, under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, EMail: PERMREQ@WILEY.COM.

Library of Congress Cataloging-in-Publication Data:

Beebe, Kenneth R.

Chemometrics : a practical guide / Kenneth R. Beebe, Randy J. Pell, Mary Beth Seasholtz.

p. cm. — (Wiley-Interscience series on laboratory automation)

"A Wiley-Interscience publication."

Includes index.

ISBN 0-471-12451-6 (alk. paper)

1. Chemistry, Analytic—Statistical methods. 2. Chemistry, Analytic—Mathematics. I. Pell, Randy J. II. Seasholtz, Mary Beth. III. Title. IV. Series.

QD-5.4.S8B44 1998

543'.0015195—dc21

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CHAPTER

3

Preprocessing

Be prepared.

—Motto of The Boy Scouts of America

Preprocessing is a very important part of any chemometrics data analysis project. It is so important that it is delineated as one of the "Six Habits of an Effective Chemometrician" (see Chapter 1) and is defined as any mathematical manipulation of the data prior to the primary analysis. It is used to remove or reduce irrelevant sources of variation (either random or systematic) for which the primary modeling tool may not account. Keep in mind that preprocessing changes the data which will either positively or negatively influence the results. "Being prepared" by applying the appropriate preprocessing tool(s) is critical in order for the overall data analysis to be successful.

Selecting the optimal preprocessing may require some iteration between the primary analysis and the preprocessing step. Although this empirical approach is a common practice, it is best if the preprocessing tool is chosen because of a known characteristic of the data. For example, percent transmission spectra are often linearized with respect to concentration by converting them to absorbance units.

In this chapter a number of preprocessing tools are discussed. They are divided into two basic types depending on whether they operate on samples or variables. Sample preprocessing tools operate on one sample at a time over all variables. Variable preprocessing tools operate on one variable at a time over all samples. Therefore, if a sample is deleted from a data set, variable preprocessing calculations must be repeated, while the sample preprocessing calculations will not be affected.

3.1 PREPROCESSING THE SAMPLES

The first set of preprocessing tools discussed are those that operate on each sample. Table 3.1 lists the four methods discussed: normalizing, weighting, smoothing, and baseline corrections. Normalization can be used to remove

TABLE 3.1. Sample Preprocessing Tools

Method	Use
Normalizing	Puts all the samples on the same scale by dividing by a constant (e.g., removing variable injection volume in chromatography).
Weighting	Sample weighting gives some samples more influence on the analysis than others (e.g., a weight of zero eliminates a sample).
Smoothing	Reduces the amount of random variation (noise).
Baseline corrections	Reduces systematic variation.

sample to sample absolute variability (e.g., variable injection volumes in chromatography) while weighting emphasizes selected samples over others. Smoothing is primarily used to reduce random noise whereas the other sample preprocessing methods are used to remove systematic variations. Baseline features can be removed using explicit models, derivatives, or multiplicative scatter correction.

3.1.1 Normalization

Normalization of a sample vector is accomplished by dividing each variable by a constant. Different constants can be used and three are described here.

Normalizing to unit area is accomplished by dividing each element in the vector by the "1-norm." The 1-norm of a vector is the sum of the absolute value of all of the j entries in the vector x , as shown in Equation 3.1.

$$1\text{-norm} = \sum_{j=1}^{n\text{vars}} |x_j| \quad (3.1)$$

Normalizing to unit length is accomplished by dividing each element in the vector by the "2-norm." The 2-norm is calculated by taking the square root of the sum of all the squared values in the vector, as shown in Equation 3.2.

$$2\text{-norm} = \sqrt{\sum_{j=1}^{n\text{vars}} x_j^2} \quad (3.2)$$

Normalizing so that the maximum intensity is equal to 1 is accomplished by dividing each element in the vector by the infinity norm, defined as the maximum (in absolute value) of the vector.

Normalization is performed in order to remove systematic variation, usually associated with the total amount of sample. A common example of this is normalizing to the largest m/e peak in mass spectrometry (Howe et al., 1981, p. 19). In chromatography, normalization of the entire chromatogram to unit area is used to remove the effect of variable injection volume. Normalizing to

unit area is also used in library searching in mass spectrometry (Howe et al., 1981, p. 229), and in two-component curve resolution. (Lawton and Sylvestre, 1971).

Consider an example of a chromatographic application with two components of interest. The raw chromatograms from two injections of two samples are shown in Figure 3.1a. The chromatograms normalized to unit area (1-norm) displayed in Figure 3.1b demonstrate the elimination of injection volume variations (i.e., the chromatograms from the same sample overlay).

A second example of normalization comes from an application of near-infrared reflectance spectroscopy for sorting recycled plastic containers. Spectra of discarded containers were measured and pattern-recognition tools were applied in order to facilitate the sorting (see Sections 4.2.1.2 and 4.3.1.2). Prior to applying the pattern-recognition tools, extensive preprocessing was performed. Shown in Figure 3.2a are spectra of the polyethylene samples (the second derivative of the data has already been taken for reasons discussed later). Due to variations in the pathlength, the spectra vary in intensity. Normalization to unit area (dividing by the 1-norm) reduces this pathlength variation (see Fig. 3.2b). The spectra overlay more closely, especially in the 1750-

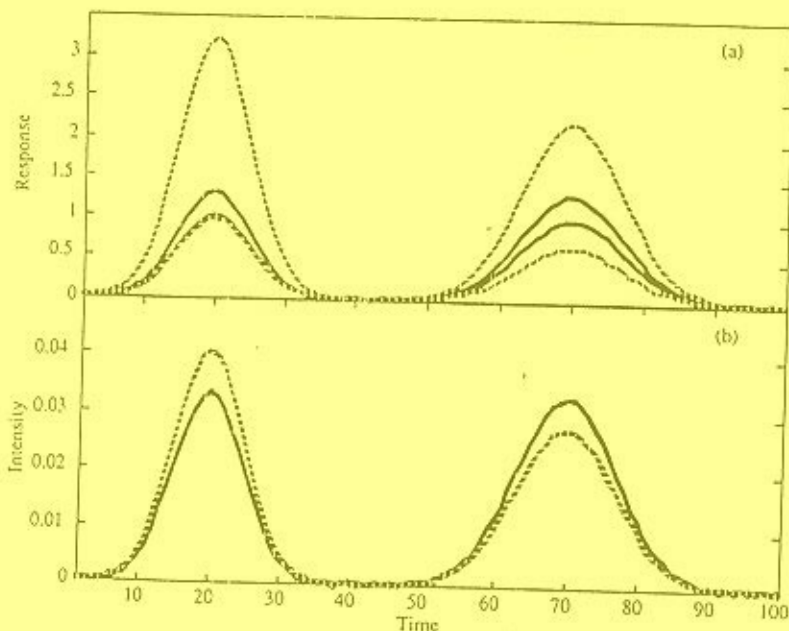


Figure 3.1. Chromatograms with the same relative concentrations of two components but with varying injection volumes. Results before (a) and after (b) normalization to unit area.

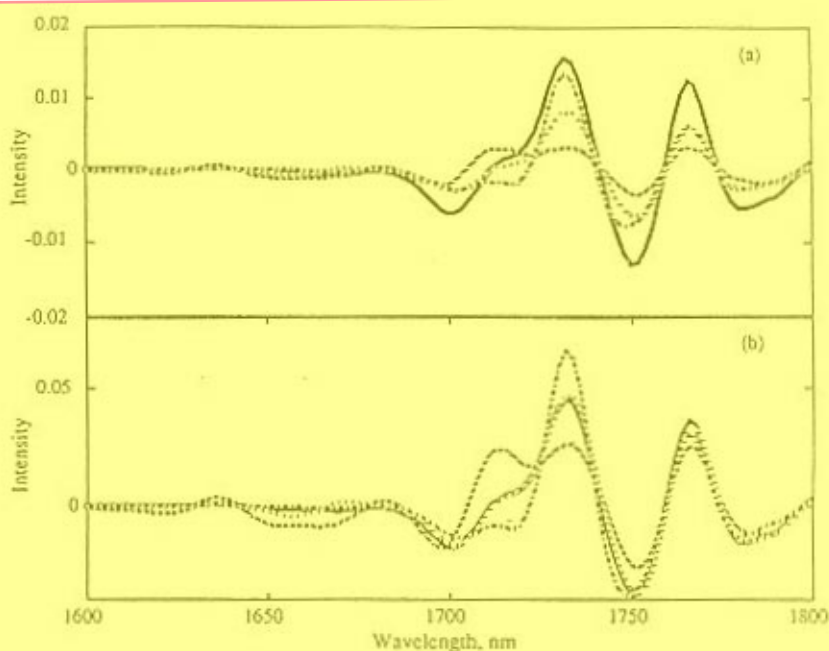


Figure 3.2. Second derivative near-infrared reflectance spectra of recycle polyethylene containers before (a) and after (b) normalization to unit area.

to 1800-nm region. The remaining spectral differences are due to chemical variations in the samples.

In summary, normalization reduces systematic variation in the data by dividing by a constant. Depending on what variation is to be removed, one normalization constant may be more appropriate than another. Beware that normalization may remove important concentration information.

3.1.2 Sample Weighting

Sample weighting is accomplished by multiplying each element in a sample vector by a constant. In this way, the influence a sample has on a mathematical model can be manipulated. Sample weighting is similar to normalization, but the criteria for defining the constants differ. The weights can be any values, although weighting should only be applied when reliable information is available about the relative importance of some samples over others. For example, the data from a highly experienced analyst can be given more weight than those of a trainee. Another use of weighting is to satisfy inherent assumptions of the primary method of analysis (e.g., the assumption of homoscedastic errors in linear regression, Draper and Smith, 1981).

3.1.3 Smoothing

In analytical chemistry, it is assumed that a measured signal consists of the true signal plus random noise. The amount and structure of the noise depends on the experiment. Smoothing tools (smoothers) are used to mathematically reduce the random noise with the goal of increasing the signal-to-noise ratio. A basic assumption made with these tools is that the noise is of higher frequency relative to the signal of interest. It is the redundant information contained in adjacent variables that enables smoothers to separate the "true" signal from the noise. Some sources of noise contribute to the low-frequency signal, but they are often difficult to mathematically remove without removing some of the chemical information of interest. Baseline correction methods are used to remove these low-frequency signals.

Smoothing methods typically use a window which can be thought of as a region of influence. All the points in the window are used to determine the value at the center of the window, and therefore the window width directly affects the resulting smooth. Five methods for smoothing are discussed below. Four of them use a window, but differ in how the points in the window "vote." The fifth method, Fourier smoothing, does not use a window.

3.1.3.1 MEAN SMOOTHER As defined here, a mean smoother is used to decrease the number of variables in a sample vector. This may be needed if, for example, the calculation speed must be increased. To begin, a window width (n) is chosen and the mean of the first n points in the sample vector is calculated. This defines the first entry in the mean smoothed vector. The second entry is calculated as the mean of the $n + 1$ to $2n$ points in the original sample vector. This process is repeated for all elements in the original vector. The resulting smoothed vector has a factor of n fewer elements. The mean smoother with a reasonable window width is better than contracting the vector by taking every n th point because the mean calculation results in signal averaging. Figure 3.3a displays a spectrum containing 800 variables. In Figure 3.3b, a mean smoother with a window width of 20 has been applied which has reduced the number of variables to 39. The mean smoother always reduces the resolution; this is evidenced in Figure 3.3b by the elimination of the sharper features. Therefore, choosing an inappropriate window width may eliminate important features in the data.

3.1.3.2 RUNNING MEAN SMOOTHER Running smoothers operate by moving the window across the data vector one element rather than one window width at a time as with the mean smoother described above. This results in a smoothed vector that is the same (or almost the same) length as the original sample vector. Specifically for the running mean smoother, the j th element in the new vector is the mean of the original data located in the window centered around the j th element. These smoothers introduce features in the ends of the sample vector and are termed "end effects." For example, the first element in the vector is often deleted because it cannot be in the middle of a window. In fact (window size-1)/2 points on either end of the vector cannot be smoothed in the same manner as the remaining points.

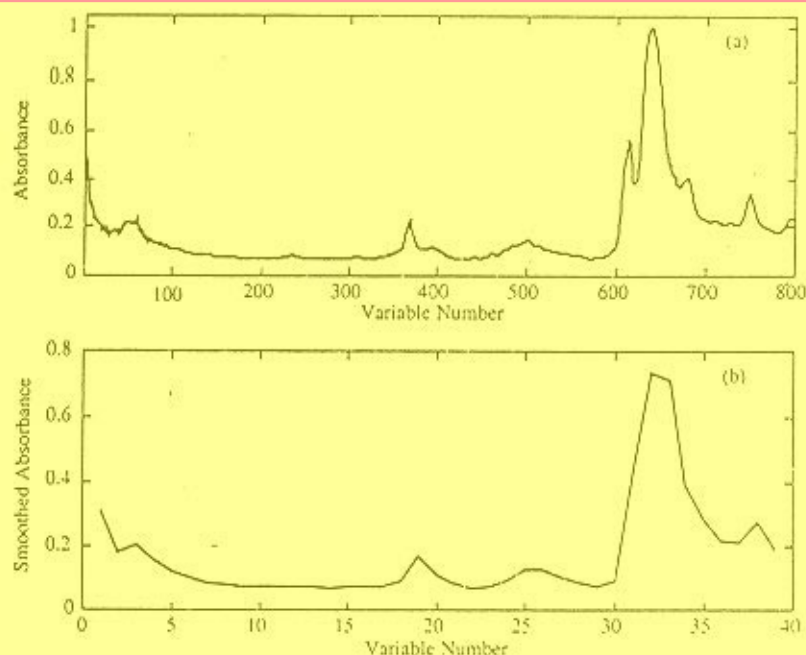


Figure 3.3. Spectrum before (a) and after (b) applying a mean smoother with a window width of 20.

This smoother is one way to improve the signal-to-noise ratio of the data, as shown in Figure 3.4. The original data has some visible random noise and a large spike. The mean smoother with a 3-point window reduces the noise significantly, but does not remove the spike. With the largest window size (21 point), the spike is removed but the shape of the peak has changed (broader and lower intensity). The apparent shift in the peak to lower variable number is due to the end effect (i.e., 10 points from each end of the sample vector have been removed).

3.1.3.3 RUNNING MEDIAN SMOOTHER The running median smoother is similar to the running mean smoother except the median is used instead of the mean. The median is not as sensitive to extreme points as the mean (Honglin et al., 1983) and, therefore, the median smoother is very effective at removing spikes from the data. However, it is not as efficient at filtering noise. The median smoother applied to the raw data presented in Figure 3.4 is shown in Figure 3.5. Compare the 3-point window results in Figures 3.4 and 3.5. The median smoother removed the spike better than the mean smoother, but the latter more effectively reduced the noise. Because of the complementary nature of the two approaches, a combination of running mean and running median smoothers may be appropriate for some data sets.

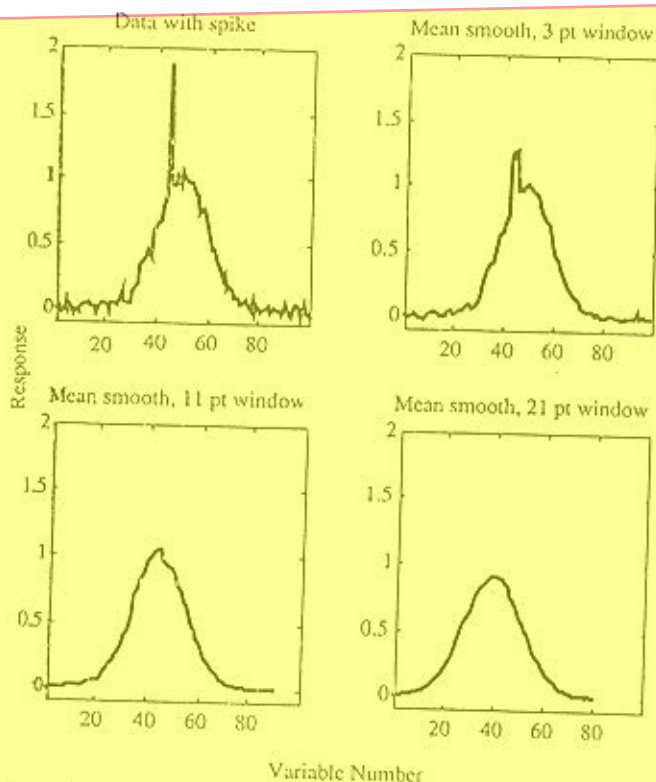


Figure 3.4. Results after applying a running mean smoother with varying window widths to a sample vector.

3.1.3.4 RUNNING POLYNOMIAL SMOOTHER The running polynomial smoother differs from the running mean and median smoothers in that a low-order polynomial is fit to the points in the window. The j th element in the smoothed data vector is equal to the polynomial prediction at element j . A convenient implementation of this approach is that of Savitzky and Golay (1964). An example of a polynomial fit over one window width (13 points) is shown in Figure 3.6. The data in this window are used to calculate the smoothed value in the middle of the window. The solid line shows the second-order polynomial fit to the data, and the "X" is the smoothed value for data point 7.

Using the Savitzky-Golay method results in the elimination of (window size - 1)/2 points on each end of the sample vector. If this is unacceptable, Gorry has developed a method that does not result in the elimination of points (Gorry, 1990, 1991). While this method preserves the original number of variables, it can introduce aberrant features to the ends of the sample vectors (Hui and Gratzl, 1996). Overall, the Gorry method is recommended and is used in all subsequent discussions.

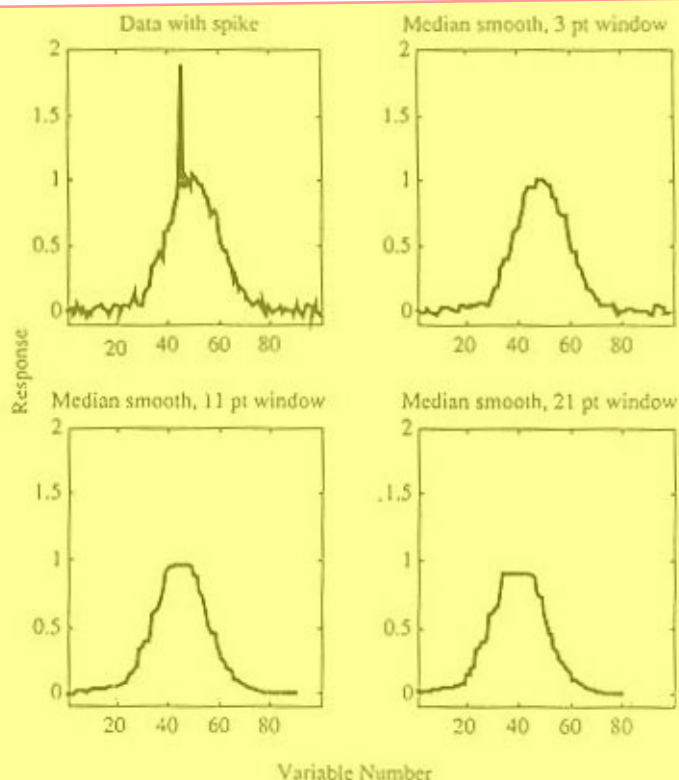


Figure 3.5. Results after applying a running median smoother with varying window widths to the same data found in Figure 3.4.

As with all window-based smoothers, the choice of the window size in polynomial smoothers is very important. Another decision to make for polynomial smoothers is the order of the polynomial to be fit (Barak, 1995). Typically, a second- or third-order polynomial is used. An example of applying a polynomial smoother is shown in Figure 3.7, where a second-order polynomial is fit with window sizes of 7, 13, and 25 points. As the window size increases, the noise is continually reduced. However, when the window is too large, sharp peaks may be removed and the remaining peaks distorted. This is demonstrated in Figure 3.8 where a spectrum is shown before (solid) and after (dashed) applying a 49-point second-order polynomial smoother.

For a given data set, the optimal window size and polynomial order depend on the nature of the data. Of primary importance is the width of the peaks relative to the window width (e.g., choosing a window width 10 times the width of a peak will most likely distort or eliminate it). An approach to selecting a reasonable window width and polynomial order is to apply several combinations and evaluate the resulting preprocessed data and final results.

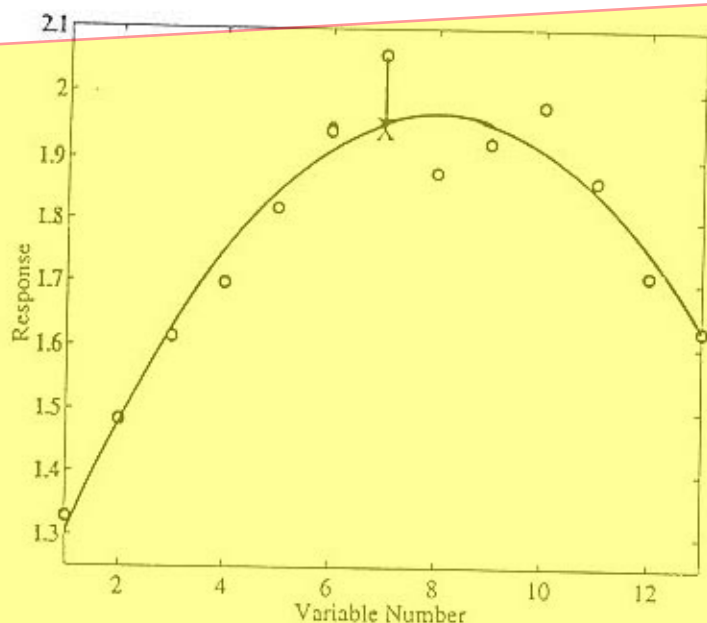


Figure 3.6. Polynomial fit with a window width of 13 points. The smoothed value of data point 7 is shown as X.

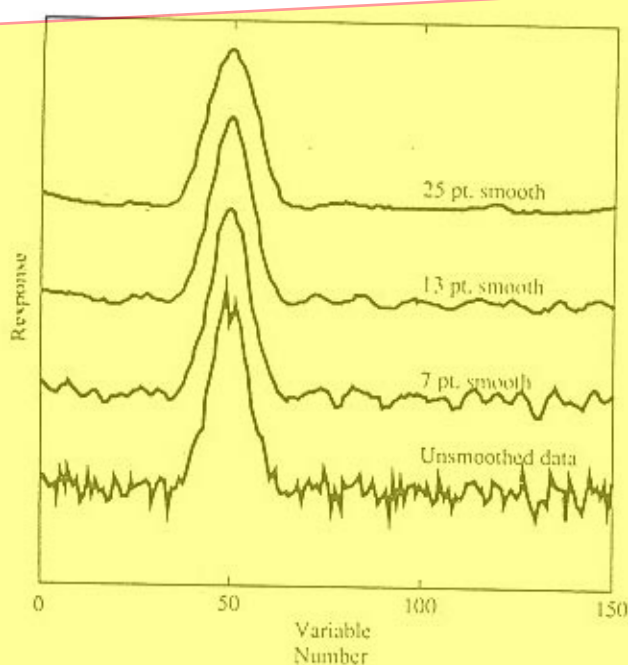


Figure 3.7. Results after applying a running polynomial smoother with varying window widths to a sample vector. (The offset was added for clarity.)

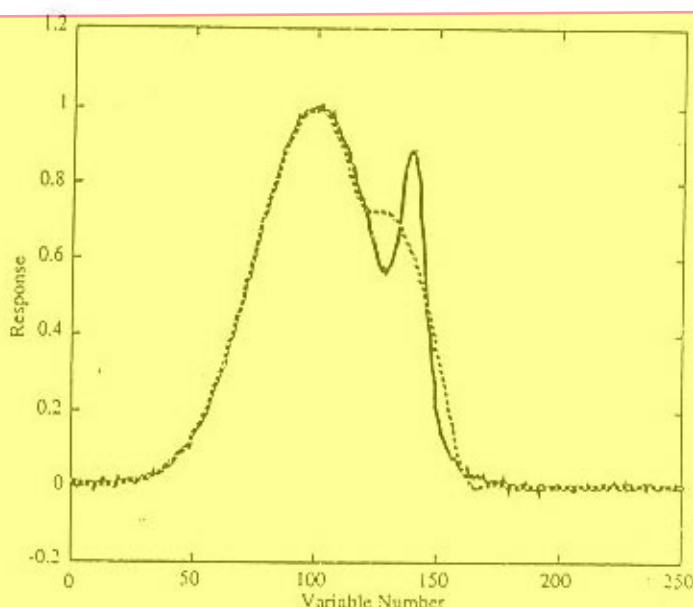


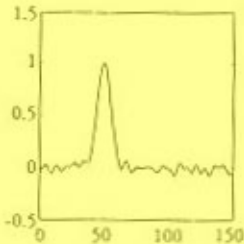
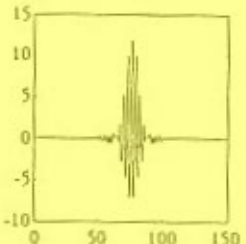
Figure 3.8. Illustration of peak distortion when using a smoother with a too-large window width (49 point). (Solid—raw data; dashed—smoothed data.)

3.1.3.5 FOURIER FILTER SMOOTHER Fourier filtering can be used for general smoothing or removal of specific frequency components. Both of these are accomplished by Fourier transforming the signal, weighting the transform with an apodization function, and back transforming to the original units.

With so many fields using Fourier analysis, the notation is varied and sometimes conflicting (Table 3.2 lists some of the notation). Interestingly, the infrared spectroscopy and mathematics notation are completely opposite. Because of this ambiguity, we will define column 3 in Table 3.2 as the apodization domain because this is where the apodization is always applied. For lack of a better term, we will refer to column 2 as the time domain.

General Fourier filter smoothing is accomplished by using an apodization function in the apodization domain. The interferogram is multiplied by the apodization function before transforming to the time domain. There are many types of apodization functions (Griffiths and de Haseth, 1986), perhaps the most simple being boxcar apodization. One example of boxcar apodization is zeroing high-frequency Fourier coefficients. Figure 3.9a displays a data vector that is transformed to the apodization domain in Figure 3.9b. The first and last 50 points of this interferogram are set to zero, as displayed in Figure 3.9c. This is transformed back to the original units, as shown in Figure 3.9d. This final sample vector is more smooth than the original vector because the high-frequency components are removed. Changing the number of Fourier coefficients that are zeroed in Figure 3.9b yields a different result. Figures 3.10a and

TABLE 3.2. Fourier Analysis Notation

	Time Domain	Apodization Domain
Typical Graph		
Chromatography	Time	Frequency
FT-NMR	Shift	Relaxation time FID
FTIR	Frequency (e.g., wavelength, wavenumber)	Position, time, interferogram
FTNIR		
Mathematics	Time	Frequency

b display the results of zeroing 70 coefficients rather than 50. This has produced a significantly broadened peak and has introduced artifacts to the baseline.

Fourier filtering can also be used to remove a specific frequency present in the data. Common examples of this include removing baseline offsets (low frequency) and 60 Hz line noise. Figure 3.11*a* displays a sample vector with a periodic feature. After the Fourier transform has been applied, this feature appears in the interferogram as a single point at variable 100 (Figure 3.11*b*). Figure 3.11*c* has this frequency zeroed, which when back transformed results in smoothed data without the periodic noise (Figure 3.11*d*).

This concludes the discussion of smoothers; a summary with recommendations is provided in Table 3.3.

3.1.4 Baseline Corrections

Besides high-frequency components (noise), the measured signal may also contain low-frequency sources of variation that are not related to the chemistry under investigation. In this book, these components are called baseline features. These systematic variations can be large relative to changes in the signal of interest and may dominate the analysis if not removed. They may also vary randomly in intensity and shape from sample to sample. A number of approaches for reducing baseline features are discussed below.

3.1.4.1 EXPLICIT MODELING APPROACH Any sample vector can be written as a function of variable number (*x*),

$$r = f(x) \quad (3.3)$$

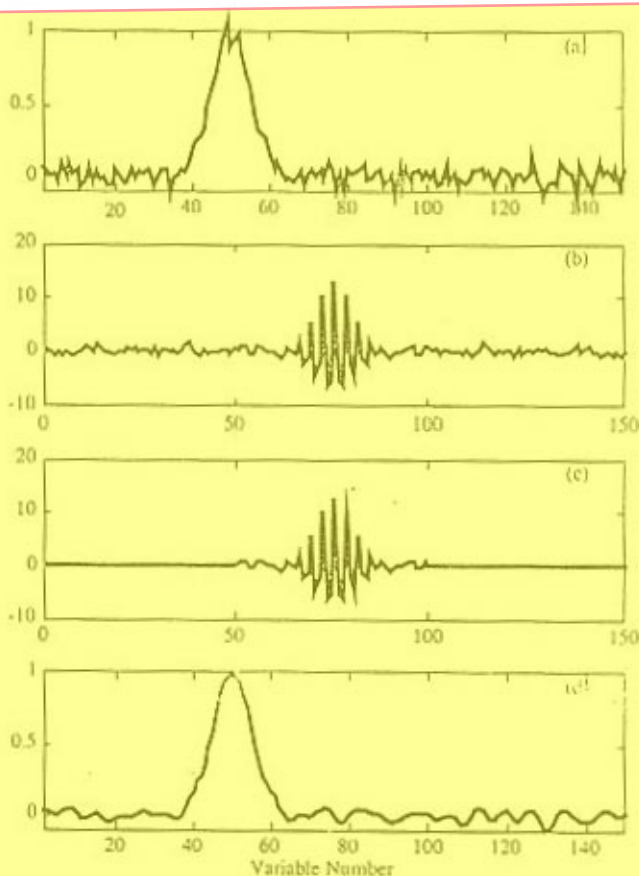


Figure 3.9. Example of a Fourier smoother with boxcar apodization (zeroing 50 points on each end). (a) Original spectrum; (b) spectrum in apodization domain; (c) first and last 50 points set to zero; (d) after transformation back to original units.

This function is equal to the sum of the signal of interest plus some baseline feature (if present). The baseline can be approximated using a polynomial, as shown in Equation 3.4

$$r = \bar{r} + \alpha + \beta x + \gamma x^2 + \delta x^3 + \dots \quad (3.4)$$

where \bar{r} is the signal of interest and the remainder of the equation approximates the baseline feature. By postulating a model for the baseline (e.g., offset, linear, polynomial), one can account for it directly by subtraction. For example, a sample vector with an offset baseline feature (i.e., a horizontal line) can be written as

$$r = \bar{r} + \alpha \quad (3.5)$$

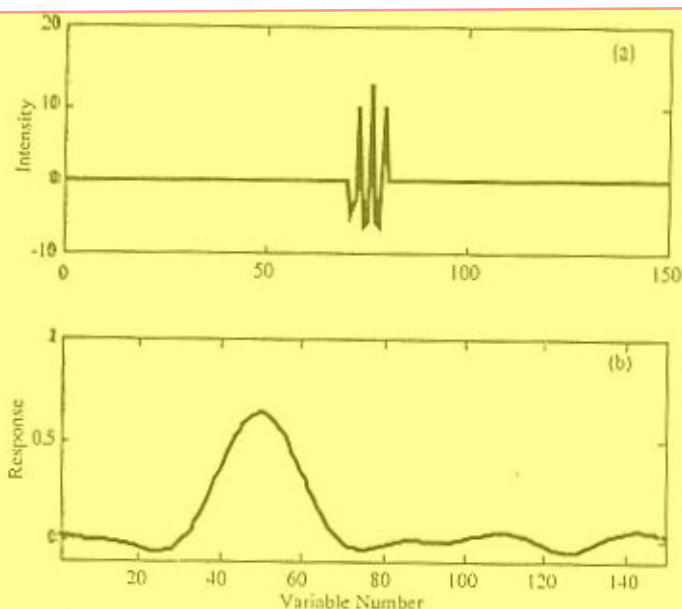


Figure 3.10. Example of a Fourier smoother with boxcar apodization (zeroing 70 points on each end). (a) Spectrum in Figure 3.9 in apodization domain with first and last 70 points set to zero; (b) after transformation back to original units.

In this case, the baseline can be removed by estimating α and subtracting it from the sample vector (r). This is illustrated in Figure 3.12a, where a series of sample vectors with offset baseline features is plotted. The offset can be removed by subtracting the intensity of a variable from all variables for each sample vector. The optimal variable is one that contains only baseline information (this is α in Equation 3.5). In this example, variable 60 is used to estimate α , which is subtracted from all elements in the sample vector. The preprocessed data shown in Figure 3.12b now reveal two groups of samples which were not apparent prior to preprocessing. Note also in Figure 3.12b that all sample vectors have zero intensity at variable 60. The average intensity of several baseline variables is often used to estimate α . This yields a better estimate of α and reduces the amount of noise introduced into the sample vectors by the baseline subtraction.

Another example of explicit baseline modeling is presented in Figure 3.13, where the sample vector contains a linearly sloping baseline. This type of baseline is encountered in chromatography as well as spectroscopy. It can be due, for example, to solvent gradients in chromatography or wavelength-dependent scattering in spectroscopy. Mathematically, this can be represented as

$$r = \bar{r} + \alpha + \beta x \quad (3.6)$$

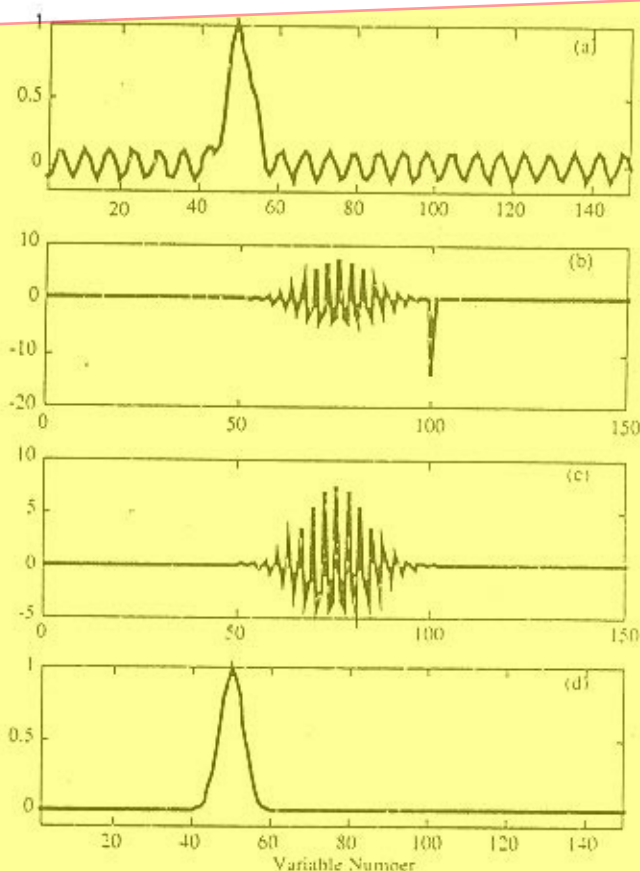


Figure 3.11. Example of eliminating a periodic noise component in a sample vector using a Fourier smoother. (a) Original spectrum with 60 Hz line noise; (b) spectrum in apodization domain; (c) spike at variable 100 set to zero; (d) after transformation back to original units.

To remove the baseline component, a line is estimated (α and β) using two or more points that are assumed to contain only baseline information. The estimated line for this example is shown as a dashed line in Figure 3.13a. To remove the baseline feature, this line is subtracted from the sample vector as shown in Figure 3.13b.

Other functions can be estimated if the baseline has a more complex shape. Regardless of the function used, the key is to choose points for estimating the coefficients in Equation 3.4 (α , β , γ , ...) that are only influenced by the baseline. If the points are chosen poorly, a portion of the chemical variation will be removed in addition to the baseline.

TABLE 3.3. Summary of Smoothing Tools

Method	Summary and Recommendations
Mean	Use this method to reduce the number of variables, but beware of the consequences of lower resolution. It is preferred over simply taking every m th data point because of the signal averaging that results from calculating the averages.
Running mean	This method works reasonably well for general smoothing. Use if no better smoothing methods are available.
Running median	Use for removal of high-frequency spikes. Not as efficient as the running mean smoother for noise reduction.
Running polynomial	Preferred method for noise reduction. The method of Gorry is recommended because no truncation of data occurs.
Fourier filtering	Good method for general smoothing, but must select an appropriate apodization function. Best method for removing specific periodic features in the raw data provided the corresponding frequency(ies) can be identified in the apodization domain.

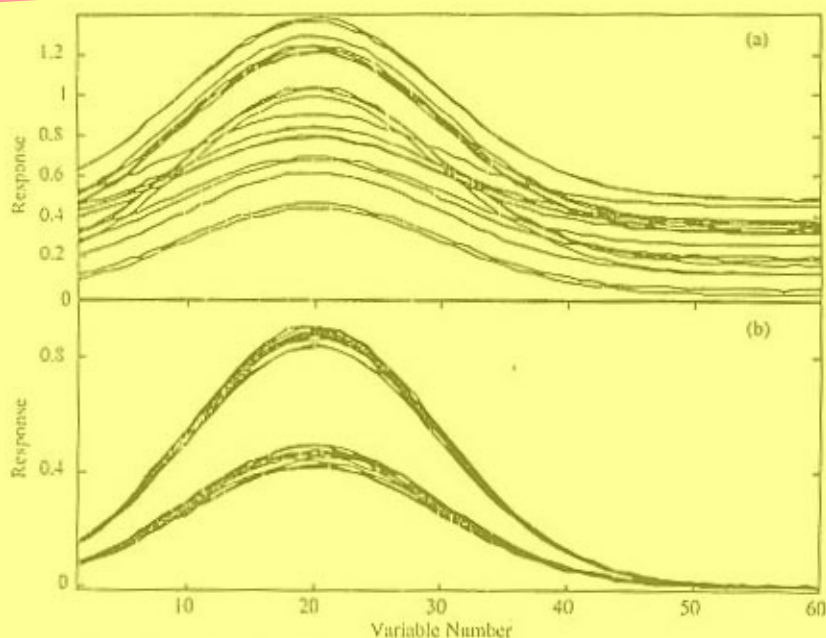


Figure 3.12. Data with a baseline offset before (a) and after (b) baseline correction using the explicit modeling approach.

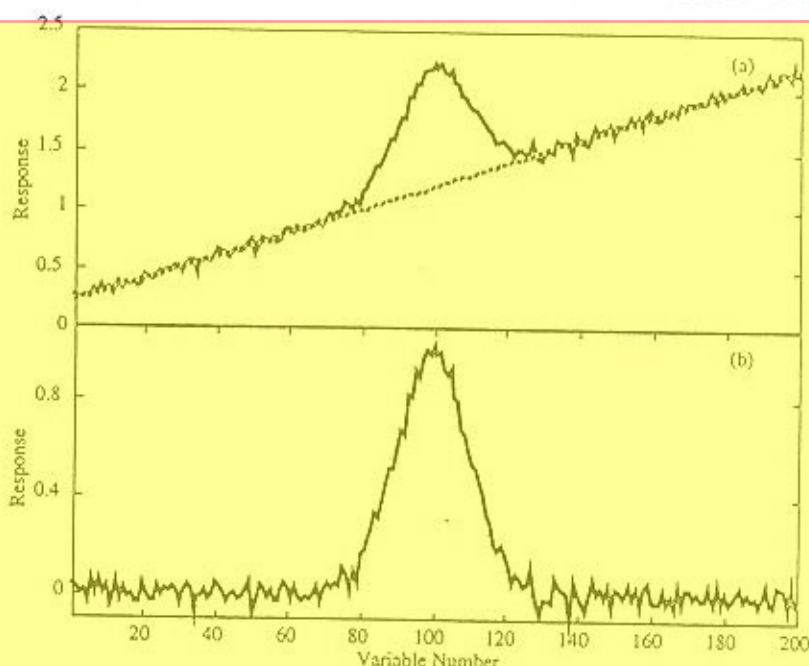


Figure 3.13. Data with a linear baseline feature before (a) and after (b) baseline correction using the explicit modeling approach.

3.1.4.2 DERIVATIVES Another way of removing baseline features is to take derivatives with respect to variable number. By using this approach, it is not necessary to select points that only contain baseline information. This is especially useful in cases where baseline points are difficult or impossible to identify.

To see how the derivatives are able to remove baseline features, refer back to Equation 3.4. Taking the first derivative of the sample vector with respect to variable number (x) yields r' ,

$$r' = \tilde{r}' + 0 + \beta + 2\gamma x + 3\delta x^2 + \dots \quad (3.7)$$

where \tilde{r}' is the derivative of the signal of interest. Equation 3.7 reveals that the first derivative has completely removed the offset feature (α). If the baseline is only comprised of an offset, the other coefficients in Equation 3.4 (and therefore Equation 3.7) would be zero.

If a more complex baseline is present ($\beta, \gamma, \delta, \dots \neq 0$), repeated application of the derivative will successively remove the higher-order terms. For example, taking the derivative of Equation 3.7 yields

$$r'' = \tilde{r}'' + 0 + 0 + 2\gamma + 6\delta x + \dots \quad (3.8)$$

Equation 3.8 is equivalent to the second derivative of Equation 3.4 and the linear feature (α and β) has been completely removed.

Running Simple Difference

The simple difference between adjacent data points can be used to estimate the first derivative. For a sample vector $\mathbf{r} = [r_1, r_2, \dots, r_n]$ the first derivative can be estimated as $\mathbf{r}' = [r_2 - r_1, r_3 - r_2, \dots, r_n - r_{(n-1)}]$. This procedure can be repeated to estimate the second and successive derivatives. As an example, Figure 3.14a displays a noise-free Gaussian peak with a constant offset of one unit and Figure 3.14b shows the derivative calculated by simple difference. The baseline has been removed (the region where there was no peak in the raw data now has zero intensity), and as expected, the peak has the shape of the derivative of a Gaussian.

When noise is present, the simple difference approach for calculating a derivative is not effective. The difference calculation propagates errors into the derivative which degrades the signal-to-noise. This is illustrated in Figures 3.15a and b where the same Gaussian peak with noise added and the simple difference result are shown, respectively. It is clear that the signal-to-noise has been decreased by the preprocessing.

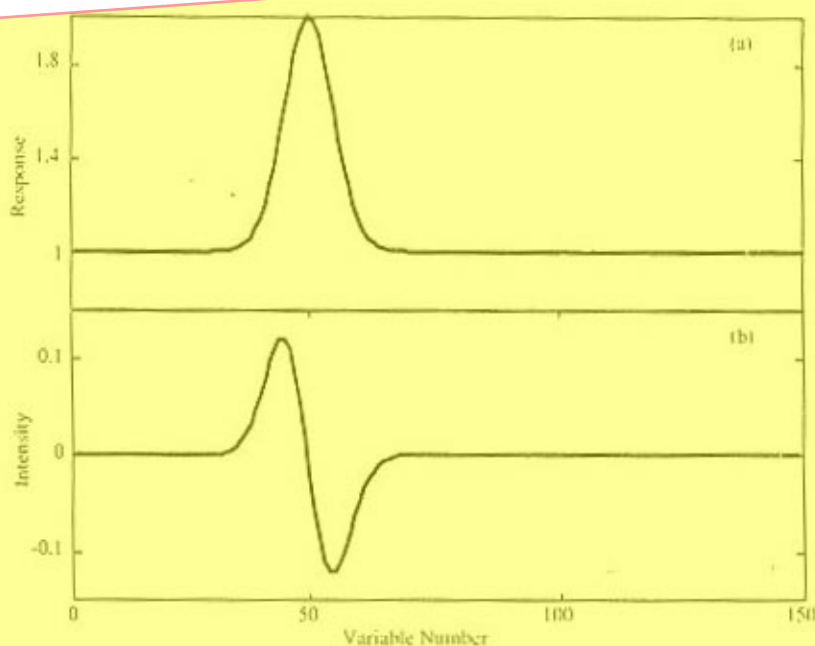


Figure 3.14. Noise-free data with a baseline offset before (a) and after (b) baseline correction using a simple difference derivative.

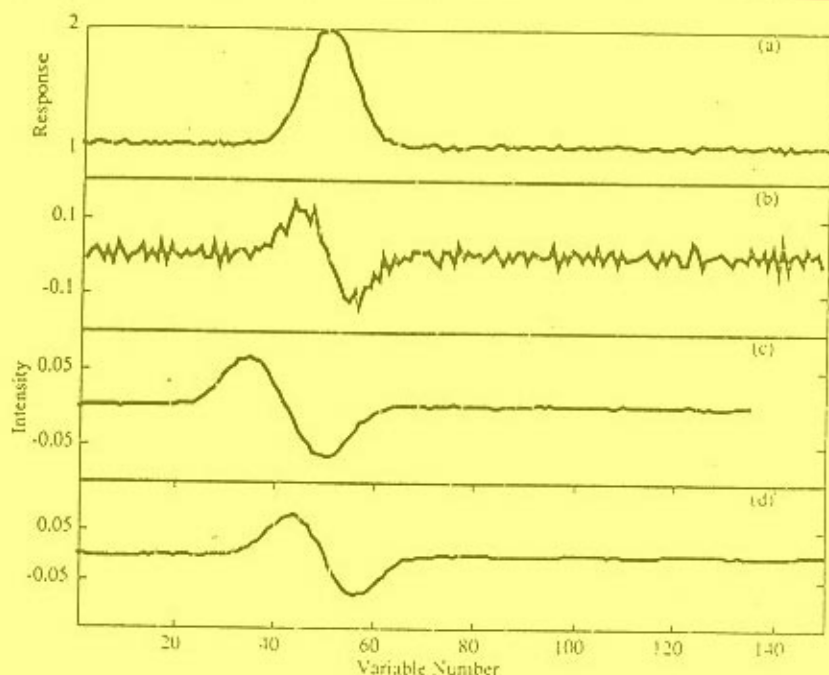


Figure 3.15. Results after applying three different derivative preprocessing tools to a sample vector. (a) A sample vector with noise and an offset of one unit. (b) The derivative calculated by simple difference. (c) The derivative calculated using a running mean difference with a window width of 15. (d) The derivative calculated using the Gorry method with a window width of 15.

Running Mean Difference

One way of improving the error-propagation properties of the simple difference method is to use the difference of means for the derivative calculation. With the running mean difference, a window width is selected and the difference between means is taken instead of using the difference between individual points. For example, using a window size of 3, the first derivative of the sample vector $r = [r_1, r_2, \dots, r_n]$ can be estimated as

$$r' = \left[\frac{r_2 + r_3 + r_4}{3} - \frac{r_1 + r_2 + r_3}{3}, \dots \right] \quad (3.9)$$

This results in a smoothed derivative calculation (see Section 3.1.3 for discussion of smoothing). Figure 3.15c shows the result of applying a running mean difference derivative with a window width of 15 to the data in Figure 3.15a. The signal-to-noise of this derivative is much better than the simple difference derivative. (The apparent shift of the derivative is due to the end effects.)

The Methods of Savitzky-Golay and Gorry

Another approach to calculating derivatives is based on the Savitzky-Golay and Gorry smoothing methods (see Section 3.1.3.4; Savitzky and Golay, 1964; and Gorry, 1990, 1991). Recall that these methods fit a simple polynomial to a running local region of the sample vector. A window width is selected and the point in the center of the window is replaced with the polynomial estimate of that point. With the derivative, this point is instead replaced with the *derivative* of the polynomial at that point. Because polynomials are used, it is a simple mathematical step to determine the derivative. As with smoothing, the contribution from Gorry was developing a method for treating the ends of the vector so as not to eliminate points. The improvement in signal-to-noise achieved using the method of Gorry with a window width of 15 is demonstrated in Figure 3.15d for the first derivative of the data in Figure 3.15a. While the signal-to-noise of the running mean and the Gorry methods are similar, we recommend using the Gorry method because fitting a polynomial preserves the peak shape better than the running mean.

To show the importance of smoothing, especially when calculating higher-order derivatives, Figures 3.16a-c show a sample vector with a linear baseline, a simple difference second derivative, and a second derivative by the Gorry method using an 11-point window, respectively. Using the simple difference method, propagation of error is more problematic when calculating the second derivative compared to the first derivative (compare Figures 3.15b and 3.16b). Therefore, using a smoothed derivative method, such as Gorry's, is even more important for this and higher-order derivatives. (In Figure 3.16c the Gorry method has introduced aberrant features to the ends of the second derivative. See also Section 3.1.3.4.)

A critical consideration when taking derivatives is the window width for the polynomial fit. If the window size is too small, too little smoothing takes place, resulting in derivatives with poor signal-to-noise. If the window size is too large, features will be smoothed out. The optimal window size depends on the data, because smoothing away features may or may not be detrimental to the primary analysis. The noise level, the number of data points, and the sharpness of the features should all be considered when selecting a window width. The sample vector shown in Figure 3.17 is used to demonstrate the effect of differing window widths. The first derivative results are shown in Figure 3.18. Using a window width of three results in a derivative with poorer signal-to-noise than the original data. As the window width is increased to 21, the signal-to-noise improves, but the peak at variable 55 has been smoothed away. The effect of the window width on the second derivative results in Figure 3.19 is even more dramatic. It is not possible to make a general statement as to which of the window widths is best. In practice, the primary analysis is repeated with several window widths to determine which yields the best results.