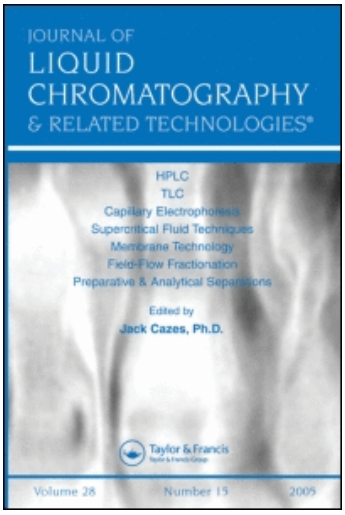

Journal of Liquid Chromatography & Related Technologies
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597273

Non-Linear Data Bunch

Darko Kantoci[a]
[a] Laboratory of Chemical Endocrinology, Loma Linda University School of Medicine, Loma Linda, CA

To cite this Article Kantoci, Darko(1997) 'Non-Linear Data Bunch', Journal of Liquid Chromatography & Related Technologies, 20: 7, 1049 — 1055
To link to this Article: DOI: 10.1080/10826079708010957
URL: http://dx.doi.org/10.1080/10826079708010957

J. LIQ. CHROM. & REL. TECHNOL., 20(7), 1049–1055 (1997)

NON-LINEAR DATA BUNCH

Darko Kantoci

Loma Linda University
School of Medicine
Laboratory of Chemical Endocrinology
Loma Linda, CA 92350

ABSTRACT

Application of data bunching to improve chromatographic quality can generally be performed in two ways. One approach is to apply linear data bunching which results in averaging data points and, consequently, risking the loss of information. By appropriately varying data bunch rates at opportune areas in a data series, the original chromatographic information will be preserved. This paper discusses the theory behind the development of a computer algorithm that automatically applies non-linear data bunching to chromatograms. Results show the reduction of data file size while improving chromatogram quality when compared to linear data bunch techniques.

INTRODUCTION

Data bunching is a widely used technique in analytical chemistry to improve chromatographic or spectral peak slope and reduce random noise.[1] If during data bunching there is no loss of significant information, the technique can be used for data reduction. This will not only improve the quality of the chromatogram (or spectrum), but also increase data access by reducing data file size.

The "linear" data bunching method averages a fixed number of data points throughout a curve. For example, a 3 point linear data bunch means that groups of three data points are averaged. The application of data bunching is user selectable and can be applied to a whole or a part of a curve. Data bunching at an inappropriate rate can degrade the quality of a curve. The solution is to bunch data at different rates depending upon the importance of data at any one area.

Our goal is to bunch data automatically at a non-linear rate in order to preserve information, reduce data set size, improve integration accuracy, and eliminate user errors in the selection of a data bunch rate.

EXPERIMENTAL

The example HPLC chromatogram was obtained on a Beckman System Gold (126 pump, 168 diode array detector) using software v 5.1, controlled by a Compaq ProLinea 4/50 computer.

The algorithm was developed with a Macintosh computer in HyperTalk, running under HyperCard.

DISCUSSION

A linear data bunch averages a consecutive group of data points, regardless of their position in a chromatogram or spectra. A common problem is in a 4 point linear data bunch where the first point is at the baseline and the remaining three points are up-slope. The linear data bunch will average these points and part of the information will be lost. The worst scenario involves a very sharp peak (as the 12 minute peak in Figure 1) where one of the former points is in the up-slope, the second at the apex and the third on the down-slope. Averaging will reduce peak intensity.

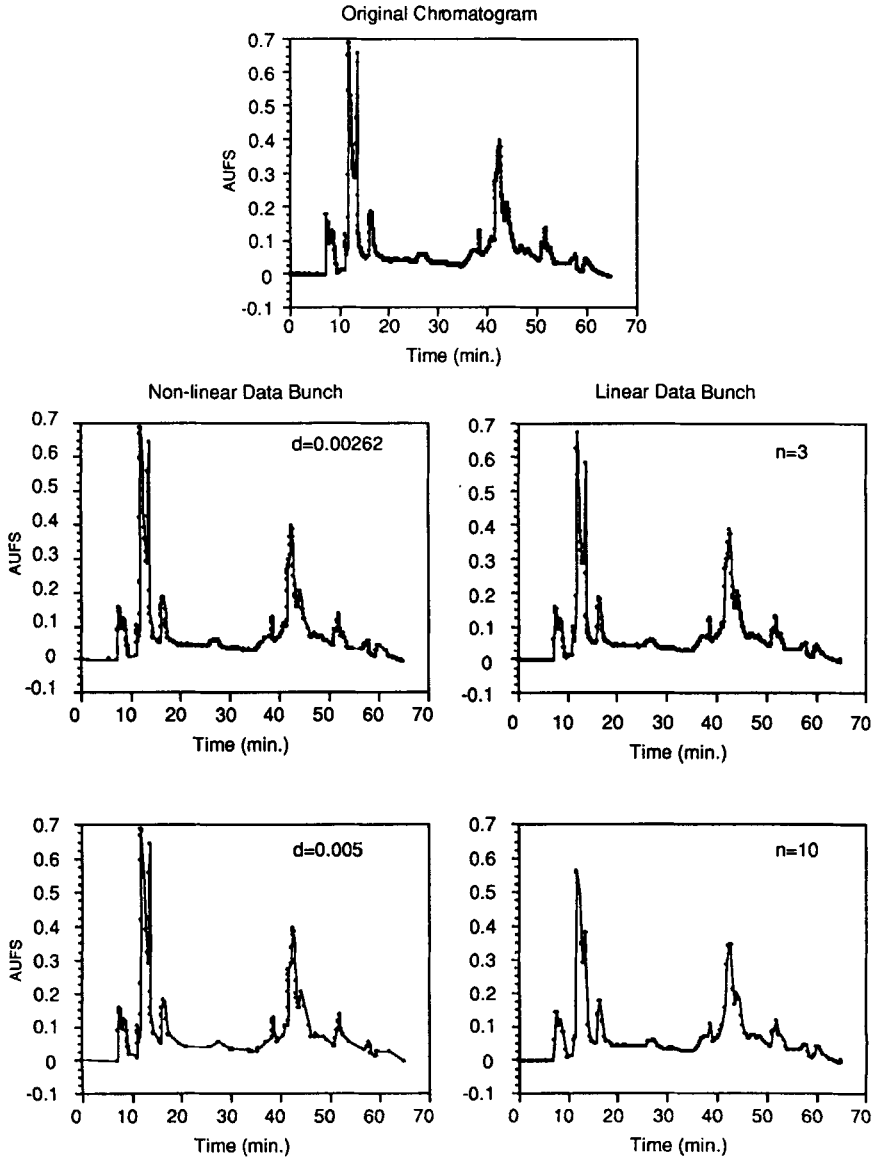

Figure 1. Non-linear and linear data bunch.

Table 1

Data Sizes of Linear and Non-Linear Data Bunch Sets

Linear Bunch			Non-Linear Bunch		
Bunch By	**No. of Data Points**	**% Red'n. in Size**	**d**	**No. of Data Points**	**% Red'n. in Size**
0 pt	1561	0.0	0	1561	0.0
			0.001	994	36.3
2 pt	780	50.0	0.002	647	58.6
3 pt	520	66.6	0.00262	475	69.6
4 pt	390	75.0	0.003	408	73.9
			0.004	251	83.9
10 pt	156	90.0	0.005	176	88.7

In the algorithm presented for non-linear data bunching, only data points that do not significantly change the slope are averaged. Thus, data points that carry information will be left intact. In this approach the data bunch rate for various parts of a curve is a function of the data itself. Such an approach can shrink data files up to 95% without significant loss of information. In contrast, a 10 point linear data bunch produces a file of the same size as a d = 0.005, but with significant information loss.

The non-linear data bunch algorithm was tested on a randomly chosen HPLC chromatogram and compared with linear data bunch results. The reduction of size is represented in Table 1. The "d" value represents the "tightness" of data compression. A lower value means that more data points will be left intact, and vice versa. The "d" value can be entered manually or calculated (estimated) using software and is constant throughout the chromatogram.

The d value of 0.00262 was estimated using our software. The original chromatogram had 1561 data points. The number of data points after non-linear data bunching depends on the information in the chromatogram. This is in contrast to linear data bunching which has a fixed number of data points regardless of the information the chromatogram carries.

Figure 1 represents the results of the calculation of linear and non-linear data bunch approaches with different "compression" ratios. Note that when the baseline is linear, it is represented with only four data points in a non-linear bunch (d= 0.00262) and the peaks contain most of the data points. The quality of the chromatogram is better than a comparable chromatogram with similar size applying linear data bunch (n= 3). With high data compression ratios (d = 0.005) there is also a noticeable baseline smoothing, without peak smoothing. However, in a linear data bunch chromatogram of similar size (n= 10) the data points are evenly distributed throughout the chromatogram. The loss of information is evident in linear data bunching with higher average rates. Note the loss of the fine structure of big peaks in a 10 point linear bunch and the loss of peak intensity (Figure 1 and Figure 2). The greatest data reduction is obtained in "clean" chromatograms, i.e., a straight baseline with a few sharp peaks; the data reduction approaches 98% with no significant loss of information (data not shown).

ALGORITHM

The algorithm[2] is basically a least squares fit. The set of equations describing a least squares fit[3] is:

$$y = ax + b$$

$$d = a + bx_i + y_i$$

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

As a first step, the algorithm calculates a and b parameters in the equation y = ax+b through the first two points. Then, the line is extrapolated to the next data point and the d_{n+1} value calculated. The d (d_{n+1}) value is defined by the equation that arises from a linear least squares fit and represents the distance between an extrapolated line and a data point (Figure 3). If the d_{n+1} value is equal to or less than the estimated or entered d value, then the next point is added to the temporary list of data points and the calculation repeated. The d value determines how close the data points must be to the propagating vector.

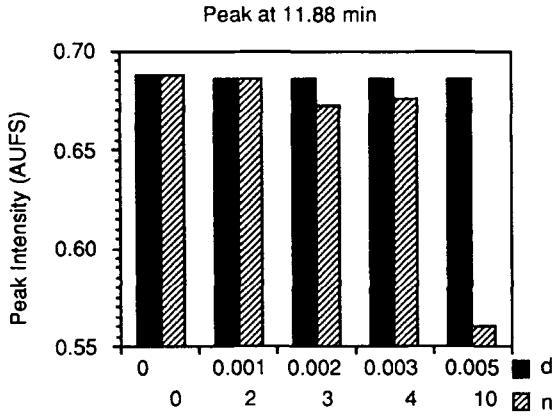

Figure 2. Reduction of peak intensity in non-linear and linear data bunch.

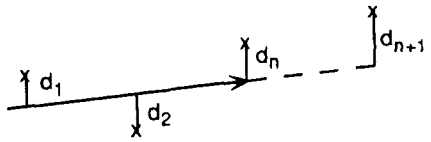

Figure 3. Least squares fit.

If the d_{n+1} value exceeded the defined d value, then the data points in the temporary list of data points are averaged and stored in another data list that contains bunched data points. The pointer is then automatically shifted to the next data point and the process is repeated throughout the chromatogram or spectrum.

Since the algorithm is sensitive to the distance of the next point from the extrapolated line, it acts as a slope detector for the integrator routine. The difference is that the non-linear data bunch routine operates with a small slope detection limit.

CONCLUSION

The algorithm for non-linear data bunching results in improved chromatogram quality. It also reduces the data file an average of 80% without

significant quality degradation. This algorithm can be applied to any type of spectra, e.g. NMR, UV, IR, etc. Such results can be obtained with a 13C NMR spectrum that contains a noisy baseline and very sharp peaks. Data compression in those instances is greater then 90% without data loss. This feature could be very useful with NMR spectra which have, in general, very large data sets.

ACKNOWLEDGMENT

The author wishes to express appreciation to Dr. William J. Wechter for assisting in the preparation of the manuscript and comments.

REFERENCES

1. G. I. Ouchi, LC GC, **13(9)**, 714-719 (1995).

2. Copyright 1995, D. Kantoci.

3. M. R. Spiegel, **Probability and Statistics**, McGraw Hill, 1975, pp. 258.

Received March 18, 1996
Accepted October 1, 1996
Manuscript 4117