JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 231–241
Published online 16 July 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.859

Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data

Giorgio Tomasi*, Frans van den Berg and Claus Andersson

The Royal Veterinary and Agricultural University (KVL), Department of Food Science, Food Technology, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

Received 16 May 2003; Revised 23 March 2004; Accepted 16 April 2004

Two different algorithms for time-alignment as a preprocessing step in linear factor models are studied. Correlation optimized warping and dynamic time warping are both presented in the literature as methods that can eliminate shift-related artifacts from measurements by correcting a sample vector towards a reference. In this study both the theoretical properties and the practical implications of using signal warping as preprocessing for chromatographic data are investigated. The connection between the two algorithms is also discussed. The findings are illustrated by means of a case study of principal component analysis on a real data set, including manifest retention time artifacts, of extracts from coffee samples stored under different packaging conditions for varying storage times. We concluded that for the data presented here dynamic time warping with rigid slope constraints and correlation optimized warping are superior to unconstrained dynamic time warping; both considerably simplify interpretation of the factor model results. Unconstrained dynamic time warping was found to be too flexible for this chromatographic data set, resulting in an over-compensation of the observed shifts and suggesting the unsuitability of this preprocessing method for this type of signals. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: DTW; COW; warping; retention time shift; PCA

1. INTRODUCTION

'Shift' is a common occurrence in chemistry. Many analytical techniques yield data where the same phenomena may yield variations at different positions (e.g. retention times in a chromatogram, wavelengths in NIR spectroscopy due to temperature influences) or may have different 'durations' depending on the specific analytical conditions. Analogously, the measurements for the single samples can have different time scales or axes, or the sample vectors may have different lengths (e.g. different batch lengths in industrial processes).

Warping is one of the numerous pretreatment methods that have been proposed to correct for shifts, conditioning data for multilinear models like PCA, PLS or PARAFAC for exploratory purposes as well as quantitative determination by alignment of the shifted variables [1]. As will be discussed later on, if data are not brought to a form where the observed variables of the samples under analysis express similar attributes, the required assumption for using bi- and multilinear modeling, namely that like variables represent similar phenomena in all samples, is violated.

In this paper, a general framework will be given to employ the two warping algorithms on chromatographic data, and their connection is illustrated using a data set from a food research experiment as a case study. The subject of this research was the effect of different packaging conditions on changes in ground coffee composition during storage over several weeks. The time-shift problem in the chromatograms is depicted in Figure 1. The problem results from a clear deterioration in the columns separation performance over time. An additional difficulty is the confounding of the column performance with storage time, the factor of primary interest in this experiment.

2. THEORY

Two different warping algorithms have received much attention in recent years for the alignment of time trajectories, chromatographic profiles and spectra [2–4]. The first method, termed dynamic time warping (DTW), was initially devised for aligning frequency spectra of words pronounced by different speakers for recognition purposes [5,6]. The more recent approach for aligning signals, termed correlation

*Correspondence to: G. Tomasi, The Royal Veterinary and Agricultural University (KVL), Department of Food Science, Food Technology, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
E-mail: gt@kvl.dk

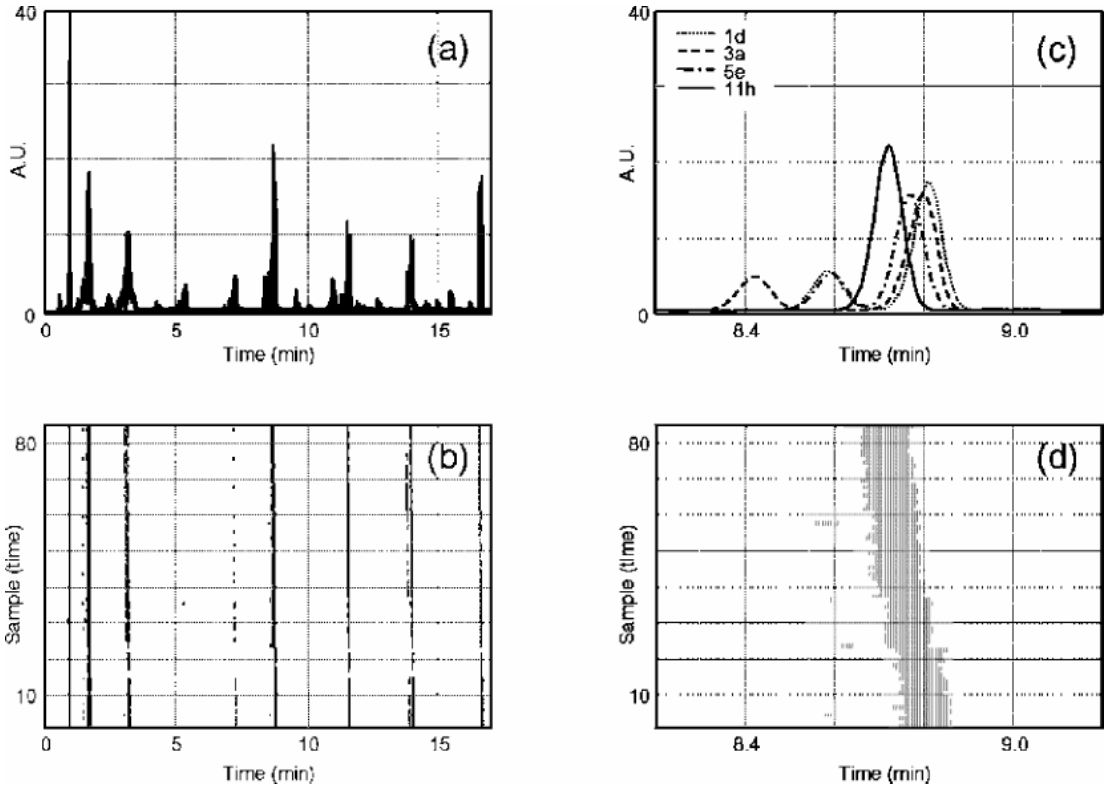


Figure 1. The shift problem as observed in the chromatograms of coffee extracts: (a) raw chromatograms for four selected samples 4(1d),15(3a),33(5e) and 84(11h); (b) top-plot of the whole data set (cut-off 5 a.u.; note sample order equals analysis order); (c) detail plot for one peak area in raw data; (d) top-plot for selected peak area (notice the systematic retention-time reduction for the main peak original found at 8.8 minutes observed in all samples). This trend is observed for all the common components in the data set.

optimized warping (COW), was proposed in 1998 as a means to correct chromatograms for shifts in the time axis prior to multivariate modeling [7].

The features of both warping algorithms will be studied in relation to PCA bilinear modeling for exploratory data analysis [8–10]. In PCA, the number of retained components expresses the complexity of the observed variations in data and may be regarded as the number of independent phenomena that cause the observed variations among objects and variables, e.g. the chemical rank of the data set in case of NIR spectroscopy. The first principal components are the ones that capture the boldest patterns; henceforth, more and more refined information may be captured by successive principal components. However, aside from the mathematical aspects of rank there are more practical implications [10]. Data sets often require preprocessing before the actual modeling step. This preprocessing effectively reduces the rank, leading to more parsimonious, more robust and better interpretable models. In NIR spectroscopy removing, for example, baseline offsets or using derivative spectra is common practice, thereby eliminating what are considered instrumental artifacts, usually of no interest to the experimenter. If these artifacts are not removed beforehand, they may form a relevant pattern in the data, frequently obscuring the important information. Time-shifts as discussed in this paper form another category of artifacts and the warping algorithms proposed can be seen as preprocessing steps applicable to factor models such as PCA.

2.1. Nomenclature and terminology

In the present work, the focus is on correcting non-analytical changes in the time-axis of samples by optimizing similarity with a reference sample. Hence, two measurements of similar nature are involved in each alignment operation in the warping procedures: the 'reference' and the 'sample' (designated by the letters 'r' and 's', respectively). The direction along which the warping is performed is simply referred to as *time*.

Throughout this work, italics are used for scalars (e.g. m) and bold for vectors (e.g. $\mathbf{m}$). N and M indicate the vector lengths for reference and sample, respectively. The mth element ($m = 1, 2, \ldots, M$) in the *time* mode of the sample is designated by s(m). Vectors $\mathbf{n}$ and $\mathbf{m}$ will be used to denote element indexing in reference and sample. For example, if $\mathbf{n} = [2\ 3\ 4]$, $\mathrm{r}(\mathbf{n}) = \mathrm{r}([2\ 3\ 4])$ selects those elements from the reference indexed by the entries in $\mathbf{n}$. A special reservation is made for $\mathrm{s}\{\mathbf{n}\}$ where the braces indicate entries in the sample vector estimated by interpolation, corresponding to matching reference points with index $\mathbf{n}$.

2.2. Correlation optimized warping

To correct for misalignments or shifts in discrete data signals, a procedure called COW was introduced by Nielsen *et al.* [7]. It is a piecewise or segmented data preprocessing method (operating on one sample record at a time) aimed at aligning a sample data vector towards a reference vector by allowing limited changes in segments lengths on the sample vector. The ratio between the number of points in the

reference vector N and the selected segment length I determines the number of segments, or rather the number of segment borders. An equal number of segments (borders) is specified on the sample vector. The maximum length increase or decrease in a sample segment is controlled by the so-called slack parameter t. When the number of time-points in a corresponding sample and reference segment differs, the former is linearly interpolated in order to create a segment of equal length.

In COW, the different segment lengths on the sample vector are selected (or the borders are shifted; 'warped') so as to optimize the overall correlation between sample and reference. The problem is solved by breaking down the global problem in a segment-wise correlation optimization by means of a dynamic programming algorithm (DP) [7,11]. The solution space of this optimization is defined by two parameters: the number of segment borders $I+1$ and the length of the slack area t. It is conventional to fix the initial and final boundaries so that the first and last points in the sample and reference vectors are forced to match. The algorithm and various changes to the original one are described in detail in the original publications [1,4,7] and will be inserted in a DTW framework in Section 2.4. In the subsequent paragraphs the implementation of the COW algorithm employed in this study is explained by means of a simple example.

Figure 2 shows the general setup of the COW algorithm. In this example, both the reference and sample vector ($N=18$ and $M=17$) are divided into $I=4$ segments of length i. Since the total vector length is not a multiple of the target segment length, the last segment is expanded to cover the entire vector. Alternatively, the remainders could be spread over all the segments [4,7]. The value of the slack parameter $t=1$ is normally fixed for all segments, but, as for the segment length i, one could define a specific value for each border [1].

The current implementation starts from the last entry in the data vector and progresses towards the first. For slack $t=1$, there are three possible boundaries in Step 1. For two out of the three possibilities an interpolation is performed to yield a number of data-points in the sample segment equal to that in the corresponding reference segment. The measure of correlation is then computed:

$$\rho(\mathbf{n}) = \frac{\mathrm{cov}[r(\mathbf{n}), s\{\mathbf{n}\}]}{\sqrt{\mathrm{var}[r(\mathbf{n})]\mathrm{var}(s\{\mathbf{n}\})}} \quad (1)$$

In step 2 (Figure 2) the segment borders are allowed for by the slack parameter from the penultimate segment. For each of these three possible positions in the last segment there are three new candidate positions in the penultimate (total = 9). Again, segments of equal length are constructed by linear interpolation. Note that, for example, pathways 2c–1a, 2b–1b and 2a–1c lead to the same boundary position for the

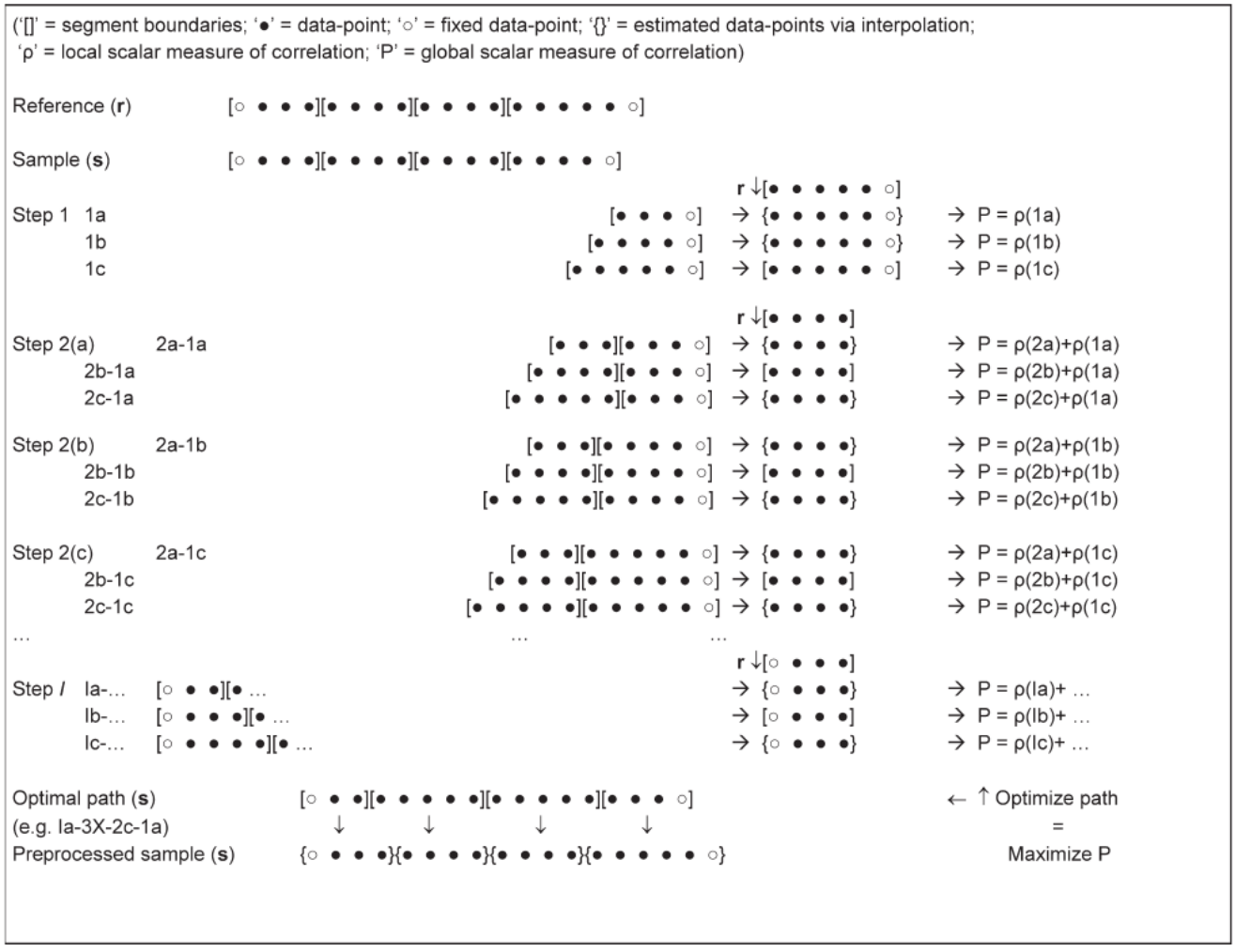


Figure 2. The concept of COW explained by means of a small example.

penultimate segment. Since not restricted by the optimization space for the algorithm as defined here, the user could select segment length and slack so that border points can pass over each other (make loops or knots in the time line). This undesirable situation has to be prevented by putting additional restrictions into the algorithm, effectively eliminating these potential knot-points during the optimization search.

In accordance with the DP principles, the optimum for the last border positioning in step I equals the global optimum and the optimal path is found by determining the global optimum from the last local optimum and all its predecessors. In this implementation a simple summation of the local measures of correlation is used, but alternatives are again possible (e.g. the product or a weighted sum). Once known, tracking back the optimal path, positioning all the borders at the right position and finding warped/aligned sample segments by linear interpolation will reconstruct the best-matching preprocessed sample signal for the predetermined set of parameters I and t.

Notice that the degree of flexibility increases for segment border points in the middle of the data vector compared to the edges. The total flexibility of the border positions in the example of Figure 2 shows the binomial-like structure 1–3–9–3–1 (1 for the two boundary points, 9 for the center border). This observation, together with algorithm parameters segment length i and slack parameter t, determines the corrective power of the COW preprocessing.

2.3. Dynamic time warping

Dynamic time warping 'nonlinearly warps the two trajectories in such a way that similar events are aligned and a minimum distance between them is obtained' [7]. The algorithm was first presented by Sakoe and Chiba [12] and further developed in numerous papers [5,6,13–15]. In recent years it has found application in chromatography [2–4], in batch process monitoring [16–20] and in gene expression studies [21,22]. The general algorithm is described in great detail in these publications and will be only briefly outlined in this section. Much focus will be put on the constraints and the synchronization step, which are essential to illustrate the connection between DTW and COW and critical to yielding a meaningful alignment of chromatographic profiles.

2.3.1. *The algorithm*

The warping path $\mathbf{F}$ is a mapping of the sample and the reference time axes on a common time axis

$$\mathbf{F} = \langle [m(k), n(k)] | k = 1, \ldots, K \rangle \tag{2}$$

where K is the length of the common time axis. The kth element of $\mathbf{F}$, $[m(k), n(k)]$, contains the indexes for the sample and the reference at the kth point on the common (warped) time axis. Figure 3 shows an example mapping grid and illustrates the concept of the warping path.

The global optimization problem in DTW can be written as follows [6]:

$$\underset{\mathrm{F}}{\operatorname{argmin}} \mathrm{D}(\mathbf{F}) = \frac{\sum_{k=1}^{k} d_{\mathrm{rs}}[m(k), n(k)] w(k)}{\sum_{k=1}^{k} w(k)} \tag{3}$$

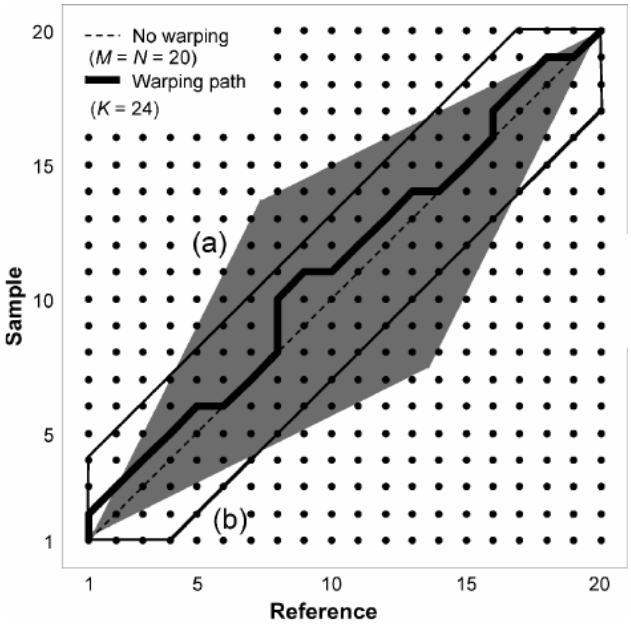


Figure 3. Mapping grid with nomenclature, the abscissa is the time axis for the reference, the ordinate is the time axis for the sample. The solid line is the warping path, i.e. the common time axis k. Feasible areas related to local continuity constraints for table $T^{(3,1)}$ (a) and band constraint with $A = 4$ (b).

where $d_{\mathrm{rs}}[m(k), n(k)]$ is a dissimilarity measure, e.g. the squared Euclidean distance between r[$n(k)$] and s[$m(k)$], and $w(k)$ are suitable weights.

In order for problem (3) to be solved correctly, $m(k)$ and $n(k)$ must not decrease along the common time axis to avoid creating knots in the time axis and trivial solutions as $\mathbf{F} = \langle (1,1), (M,N) \rangle$ must prevented. Therefore (3) is subjected to the so-called 'local continuity constraints' of the form:

$$\begin{aligned} 0 \leq m(k+1) - m(k) \leq a \\ 0 \leq n(k+1) - n(k) \leq b \end{aligned} \tag{4}$$

where a and b are integers and positive. For example, for $a = b = 1$, there are only three feasible predecessors for $[m(k), n(k)] : [m(k)-1, n(k)]$, $[m(k)-1, n(k)-1]$ and $[m(k), n(k)-1]$.

Typically, the end points for $\mathbf{F}$ are fixed and equal to (1,1) and (M,N), i.e. the initial (and the final) entries in reference and sample are constrained to be the same on the common time axis as well.

Moreover, if no weights are used, the solution to problem (3) is biased towards shorter paths, which involve the summation of fewer terms [13,22]. The main purpose of weights is thus to remove such a bias, rendering the optimal distance (and consequently the reconstructed $\mathbf{F}$) independent of the length of the warping path [20]. Not all weighting schemes fulfill this purpose [13], and although some bias may be desirable to decrease the occurrence of extreme warps [17,20], it also raises problems in establishing which optimization criterion is actually employed. Furthermore, stricter local continuity constraints (described in the next section) can be used to avoid excessive corrections in a more straightforward fashion.

The adopted weighting scheme was initially proposed in Sakoe and Chiba [6] and uses as weight for point $[m(k+1), n(k+1)]$:

$$w(k+1) = m(k+1) - m(k) + n(k+1) - n(k) \qquad (5)$$

The sum of such weights over any warping path going from (1,1) to (*M*,*N*) is equal to $M+N$ and is thus independent of *K*.

If reference and sample are treated equally, with respect to both weights and local continuity constraints, the warping is defined as symmetric [13]. The asymmetric option (e.g. if $a \neq b$) will not be described, as the principals are identical [13].

Under these restrictions, problem (3) can be solved using dynamic programming. Assuming that the global distances up to the allowed predecessors to (*m*,*n*) have been computed, the optimal global distance to point, (*m*,*n*) is the sum of $d_{\mathrm{rs}}(m,n)$, multiplied by a suitable weight depending on the local path, and the minimum among the global distances to any of the allowed predecessors.

Identifying the global optimum for the warping path is thus transformed in an efficient iterative procedure divided into a forward step and a backward step:

(1) Starting from point (1,1) and according to (4) construct the mapping grid $\mathbf{D}(M \times N)$, in which element d_{mn} is the optimal accumulated distance up to point (*m*,*n*) (the forward step).
(2) Find the optimal warping path by tracing backwards [i.e. from $(m(K),n(K)) = (M,N)$], once again in accordance with (4), the minimal accumulated distance up to point (1,1) (the backward step).

2.3.2. *Constraints*

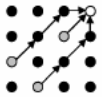
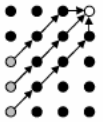
More elaborate constraints have been devised for the DTW problem than the essential ones mentioned thus far. Their use may yield warping paths more in agreement with *a priori* knowledge that may be available for the problem at hand, avoiding unfeasible compressions or expansions of reference or sample signal. As will be shown in the experimental section, such constraints are necessary to successfully apply dynamic time warping to the chromatographic data used for this application.

For a generalization of the local continuity constraints, it is necessary to introduce the concept of transition: an elementary transition describes the single advancement in the common time axis (corresponding to the single arrow in the graphs of Table 1).

The warping path can be described by a sequence of elementary transitions going from one end point to the other and its local behavior can be restricted by allowing only certain sequences of elementary transitions. Such a local series is referred to as 'rule'. The collection of rules used is named 'lookup table' and denoted by $T^{(x,y)}$, *x* being the largest block distance covered by any of the rules in the table and *y* the maximum number of horizontal/vertical consecutive transitions allowed for by the table. For example, the local continuity constraints of equation (4) with $a = b = 1$ can be translated into the lookup table:

$$T^{(2,\infty)} \equiv \begin{cases} t_1 = (1,1) \\ t_2 = (1,0) \\ t_3 = (0,1) \end{cases} \qquad (6)$$

Table I. Examples of different lookup tables with complex transition rules. The arrows represent the single elementary transition. They point in the direction of the common time axis and start from a legal predecessor to the element in the mapping grid toward which they point

Constraint [A]		Graph
Name	Slope: Min – Max	
$T^{(2,\infty)}$	0 - ∞	
$T^{(3,1)}$	½ - 2	
$T^{(5,1)}$	⅔ - 3/2	
$T^{(7,1)}_{\mathrm{COW}}$	⅔ - 4/3	

[A]The first index is the largest block distance covered by any step in the table, the second is the maximum number of horizontal/vertical transitions allowed by any step.

where the infinity indicates that this particular lookup table does not impose any restriction on the number of consecutive horizontal/vertical transitions and the two integers in parentheses are, respectively, the advancements in the sample and in the reference time axes corresponding to the elementary transition t_x.

Apart from $T^{(2,\infty)}$ (which is referred to in the remaining part of this work as the 'unconstrained DTW'), all lookup tables limit the grid points that can be reached given the end points [13]. The feasible part of the grid is typically delimited by four lines passing through either (1,1) or (*M*,*N*) and having slope equal to the minimum or the maximum slope allowed by the lookup table [13]. An example of the resulting lozenge is depicted in Figure 3.

The corrective power of the single lookup table is a function of the minimum and maximum slopes it allows for the warping path and the number of points spanned by rules in the lookup table. The closer the former are to M/N (hence 1 when sample and reference are of equal length), the more rigid are the constraints.

A second relevant type of constraint limits the feasible area to a band delimited by two lines of slope M/N. These lines pass at $|M-N|+A$ points from (*M*,*N*), where *A* is an arbitrary integer, defining the maximum compression/expansion in time-points of the sample and reference with respect to their original lengths [6]. Figure 3 shows the feasible area around the diagonal of the mapping grid. Although they prevent extreme behaviors of the warping path, band-constraints alone are not adequate for this purpose and additional local restrictions to the number of consecutive vertical or horizontal transitions are still required to prevent the optimal path from moving from the 'top' line to the 'bottom' line delimiting the search space [22].

2.3.3. *Synchronization*

The symmetric DTW algorithm yields a warped reference and a warped sample of identical length K (see Figure 3). If a warping correction takes place, K will be larger than either M or N. The extent of elongation is unpredictable until the warping process is finished and may vary from sample to sample. Therefore, an additional synchronization step rendering vectors of length N is required, if sample vectors are to be stacked for bilinear modeling. This synchronization step is not part of the original DTW algorithm, which merely used the optimal distance for classification purposes [6,22], and is necessarily asymmetric since horizontal vs vertical transitions (or reference vs sample time-points) are treated differently. To synchronize, one can take the average of the measurements at the different sample points forming a sequence of vertical transitions (e.g. reference point 8 in Figure 3). The rationale in this approach is that, by using the average, all the information in the sample time-points is taken into account. Alternatively, one could use an asymmetric warping algorithm that maps the sample time axis on the reference time axis. In this case, $n(k) = k$ and K would be equal to N. This choice may, however, lead to information loss and discontinuities in the warped sample because some points are ignored [20].

Furthermore, if slope constraints are imposed and the distinct points spanned by the rules comprising vertical or horizontal transitions are deemed as forming a segment in the sample, it is possible to use interpolation, analogously to COW. In this case, the series of distinct sample time-points determined by the optimal transition is interpolated to estimate a new series of points of length equal to the corresponding one in the reference. For example, if the series includes a horizontal transition followed by three diagonal ones (e.g. reference points 5 to 8 in Figure 3), four distinct points are involved for the reference, but only three for the sample. In order to yield the same length after the synchronization, the three points must be interpolated to four.

2.4. The connection between DTW and COW

Although COW and DTW are treated in the literature as two distinct solutions to the warping problem, there is a connection between the two that helps to shed some light on the success of COW applied to chromatographic data [1,7] opposed to the very poor results yielded by DTW on these data (see the Experimental section).

This link can be established from a combination of DTW-constraints and interpolation.

First of all, COW, expressed in a DTW framework, requires imposing the condition that rules including horizontal or vertical transitions (i.e. those correcting the shift) can be applied starting only at fixed points of the reference. Stated differently, given a lookup table of the form $T_{\mathrm{COW}}^{(7,1)}$ in Table 1, only a limited set of points on the reference are candidate end points of a rule (namely, the gray and white dots in the graph).

Lookup tables of the form $T_{\mathrm{COW}}^{(2i+y,y)}$ represent a further restriction to 'slope constraints' and each rule spans the same number of points i in the reference, whereas the number of distinct (i.e. not repeated) points in the sample can vary from $i–y$ to $i+y$. Hence, i is equivalent to the segment length as was previously defined for the COW algorithm and y corresponds to the slack parameter.

Note that tables like $T_{\mathrm{COW}}^{(2i+y,y)}$ alone are not sufficient to guarantee equivalence between COW and DTW: local continuity constraints only require that any subsection of the warping path complies with one of the rules in the lookup table and a sequence of, for example, 10 consecutive diagonal transitions would not violate any local constraint in $T_{\mathrm{COW}}^{(7,1)}$. Thus, the constraint on the initial/final points for the rules is necessary. When such restrictions are applied, all points in the lozenge of Figure 3 remain feasible, but the allowed end points form $I+1$ vertical (one-point-wide) bands within the feasible area.

Under the above constraints the local distance at the end points for each rule in the DTW algorithm may be the correlation coefficient as in COW, although problem (3) should be changed from minimization to maximization.

Moreover, if interpolation is applied to the distinct sample points of each rule prior to the computation of the distance when their number differs from i, and afterwards in the synchronization step, one yields an algorithm that is almost identical to COW. Note that, because of this interpolation, the position of the vertical or horizontal transitions in the single rules of tables $T_{\mathrm{COW}}^{(2i+y,y)}$ need not be uniquely defined to yield equivalence and in Table 1 it is set as the 'last' in the series only for simplicity.

To yield complete equivalence, one further constraint needs to be applied to the warping path because of the condition set in COW that the first point of one segment is adjacent to the last of the previous one. In the DTW context, this means that a diagonal transition with zero weight (to remove the influence of this transition from the optimal distance) connects the boundaries for the two segments.

Hence, COW may be regarded as a special case of DTW where additional constraints are added to reduce the search space for the optimal warping and to employ correlation coefficient as optimization criterion. The shape-preserving features of COW and the quality of the warping [1,7] thus appear linked to the relative rigidity of the slope constraints of the corresponding warping paths rather than the focus on correlation instead of the Euclidean distance.

Nielsen *et al.* [7] suggest that the segment length should be at least equal to the width of the smallest feature one wants to align (e.g. peaks in chromatography) and that lower values may result in the alignment of noise or other non-chemical information and alterations in peak shapes. Depending on the chromatographic analysis, this value may vary, but in the case study presented below it is approximately 30 points (same order as the largest shift observed). The corresponding slope constraints are significantly stricter than in standard DTW, where tables $T^{(2,\infty)}$ and $T^{(3,1)}$ are most often employed [4,16–18,20]. From this view point, unconstrained (or loosely constrained) DTW is expected to be too flexible and to deform peaks and features. Using rigid slope constraints may, however, prevent adequate correction at the two ends of the chromatogram because of the fixed end point assumption. Even though such restrictions on the path can be relaxed by modifying the algorithm [1,20], similar results may be obtained by appending to the sample (and/or

to the reference) a segment of suitable length containing only white noise that is removed again after warping.

DTW with rigid slope constraints (DTWc in the rest of the paper) and Euclidean distance presents a clear advantage over COW, as one can avoid setting restrictions on the rules' end points (i.e. fixed boundaries for the reference segments can be dropped) and the correction for shift is not bound to any specific position in the reference. However, this flexibility comes at the price of a considerably larger number of feasible end points (all those within the gray lozenge of Figure 3) and, consequently, computational cost. In this respect the COW algorithm has the advantage of being simple to implement, in general faster and less memory-demanding [7,20]. Furthermore, the additional flexibility granted by the DTWc may not be necessary, in which case COW would yield a perfectly acceptable solution.

3. DATA AND EXPERIMENTAL CONDITIONS

The data set is obtained by gas chromatographic (GC) analysis of extracts from ground coffee. The sample set is built up from material stored under different packaging conditions (combinations of gas headspace, vacuum and temperature) for different time durations. The instrumentation used was a GC Shimadzu 14A chromatograph (Shimadzu, Tokyo, Japan) with flame ionization detection (FID) and a Supelcowax 10 column: 30 m, 0.53 mm i.d., 1.0 μm film (Supelco, Bellefonte, PA, USA). Temperatures in the system were 280°C at the injector, 260°C in the oven, and 260°C at the detector. Internal standard was Caprylic acid (C8). Each sample was extracted by ether, and the extract was injected into a GC carrier gas-flow for in-line thermally assisted methylation and the output from a FID was recorded for 17 min. The first 2000 points were used for warping and PCA ($N = M = 2000$ data points).

The separation of the gas sample into fatty acids of different sizes results in the lightest and most volatile molecules (<C5) to be detected within the first 4 min, whereas the larger and less volatile molecules (>C25) leave the system after 16 min. All chromatograms were individually baseline-corrected by subtracting the average signal for the first 120 s from the full chromatogram prior to modeling. The typical time for a chromatographic column to wear out of course depends on the application and in particular on the number of analyses applied on the column material. However, a clear continuous drop in performance can be observed by visual inspection (see Figure 1), the incentive to the warping study described in this paper. Figure 1(a,c) shows four typical samples, Figure 1(b,d) shows an overview of the systematic variation of the entire set. The figure clearly shows the column material deteriorating over time, giving rise to shorter retention times. As mentioned previously, this column deterioration is confounded with the order of collecting samples. Because coffee extracts of this nature cannot be stored for a long time, randomization of GC analysis over the experimental condition storage time is not feasible (see below). Based on visual inspection of the 88 different samples, a total of four samples were considered outliers and were removed from the data set. From *a priori* knowledge about the sample set, chromatogram no. 9 (hence, a sample from the beginning of the measurement series) was chosen as reference for warping, since this sample contained the highest number of the chemical constituents compared with all the other samples collected at the start of the experiment. Objects in the sample series are labeled by two digits: numbers '1'–'11' indicating the different sampling times during the storage experiments, 'a'–'h' indicating the eight different packaging conditions. In this notation, the reference sample is labeled '2a'.

As the scope of this study is limited to the time warping as preprocessing, the finer details of the experiments are not given here.

4. RESULTS AND DISCUSSION

In this section the different effects of various DTW and COW implementations will be illustrated. A PCA model was fitted on the coffee extract GC data set without any shift-correcting pretreatment, meaning no warping or data mean centering. A detail of the loadings is shown in Figure 4(a), and a score plot for the first two PCs in Figure 5(a). The first principal component is identified as representing the average chromatogram. A moderate grouping according to packaging can be observed in the tendency of PC1 where the most expensive and best quality treatment (type 'g') is isolated

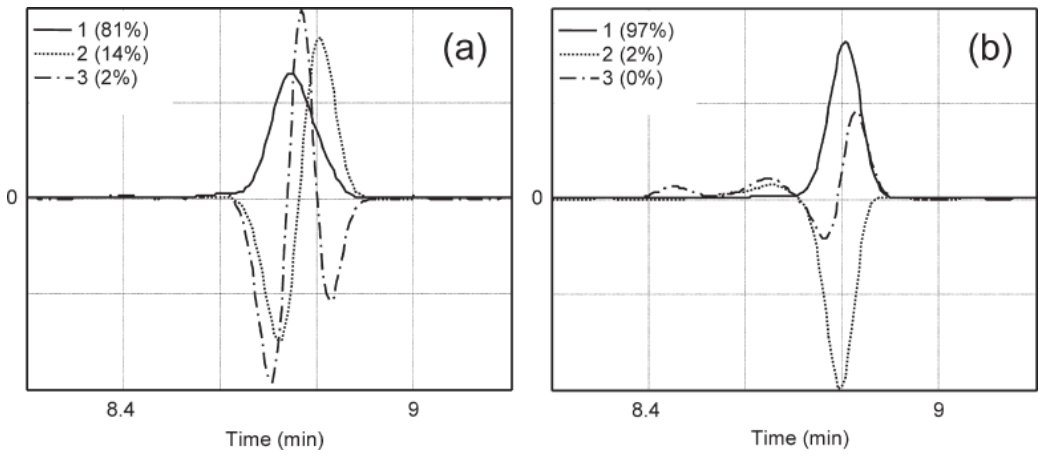


Figure 4. (a) PCA loading details raw coffee data; (b) PCA loading details after COW correction.

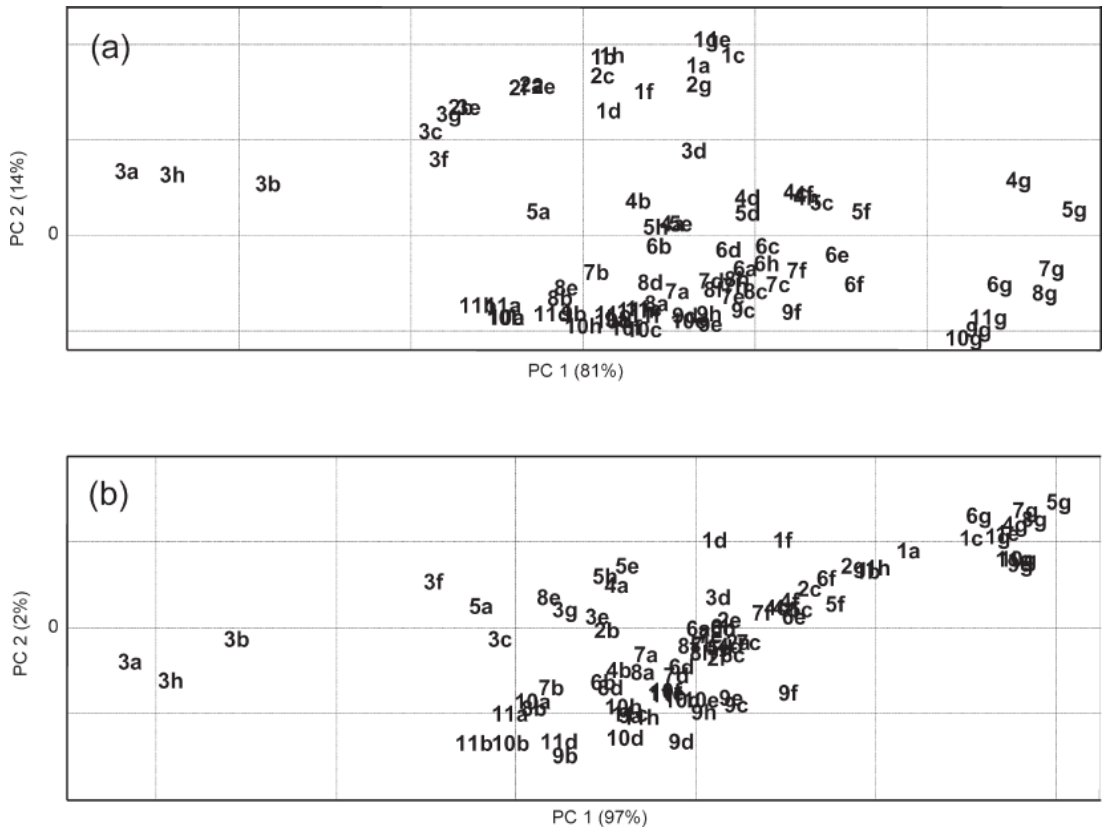


Figure 5. Score plot (a) raw coffee data (b) after COW correction.

from the others. It was anticipated during the setup of the experiment that this packaging type would show little to no effect over storage time. The second PC has a loading vector that resembles the first derivative of the loading vector for PC1. This 'Taylor series' development is characteristic for PCA modeling of data sets with shift phenomena [23]. As a consequence of the time-shifts in the chromatograms (see Figure 1), the PC1 vs PC2 score plot shows a typical 'horseshoe' configuration, visible in Figure 5(a). This shape is not related to the storage time and packaging 'g' is particularly suited to show this. Visual inspection of the set of 11 chromatograms from series 'g' is consistent with the observation that the curved shape in the score plot is almost exclusively caused by retention time shifts in the chromatograms, not by chemical changes. Moreover, what may appear as two large groupings (i.e. '1', '2' and '3' on one side and '4' to '11' on the other) is the primary consequence of the retention time shifts, following the curved pattern just described.

In conclusion, the PCA components for untreated data describe both chemical information and 'artifacts' due to the shift confounded with the experimental factor storage time (e.g. the location of the '3' samples in the score plot, in particular '3a', '3b' and '3h', is due to an altered composition of the extracts probably related to the fact that the samples in this series were stored in a freezer for several days up to analysis). Exploratory PCA modeling of the untreated (unwarped) data set is then clearly unfeasible.

In order to verify the change in the quality of the warping for DTW as the slope constraints become more and more strict, several lookup tables have been applied [Figure 6(a, b, d)]. A locally unconstrained DTW algorithm using a $T^{(2,\infty)}$ lookup table was applied with band constraints. Band limitation A was set equal to 200, corresponding to a maximum $\pm 10\%$ allowed correction for the chromatograms, significantly larger than the observed 20–30 points maximum shift. For the synchronization step, averaging was used [20]. The results of this preprocessing were very unsatisfactory [Figure 6(a)], e.g. small peaks like those found in the region 8.4–8.6 min in Figure 1(c) disappear. They are completely merged with the bigger one found in every chromatogram in the range 8.7–8.8 min. The disappearance of the small peaks is not caused by the warping itself, but by the averaging in the synchronization step, which compresses all information occurring during long sequences of vertical transitions [4]. Synchronization is also responsible for a second type of artifact related to the peak height: when the sample peak is larger than the matching one in the reference, the former is cut in height. Conversely, when the sample peak is smaller, its top element is repeated until the two sides of the synchronized peak match those on the reference, resulting in a plateau. This behavior cannot be the consequence of the described warping procedure, which is symmetric. When the reference peak is smaller, its top is repeated until the sides overlap the larger sample peak. In the warping path this appears as a sequence of vertical transitions, for which corresponding sample points are successively averaged. Thus, after the synchronization the sample peak is still taller than the reference peak, but the computed average may still be significantly lower than the original peak height. Note that normalization of reference and sample vector to length one, as suggested in some

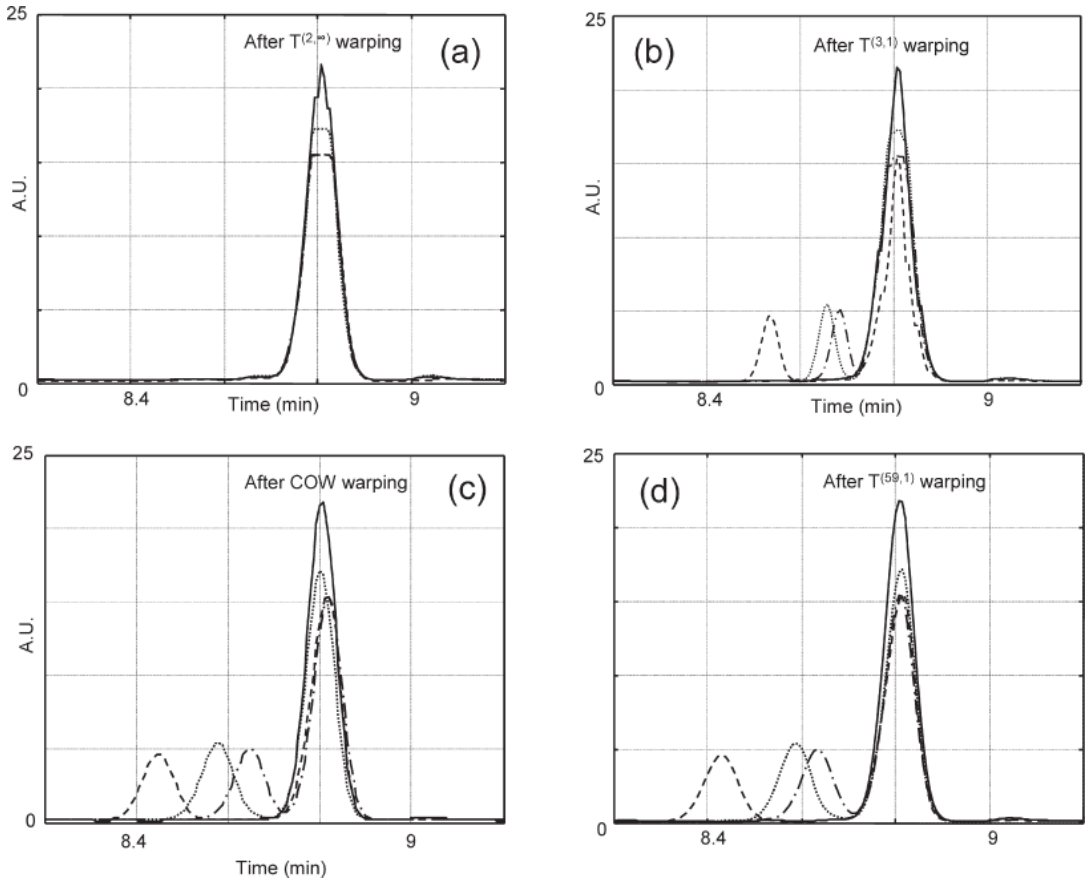


Figure 6. Signal detail reconstruction after (a) $T^{(2,\infty)}$ warping; (b) $T^{(3,1)}$ warping; (c) COW; (d) $T^{(59,1)}$ warping.

publications, does not remedy these errors. These artifacts were observed at various retention times for different peaks and substantially altered the chemical diversity present in the data, confirming the inadequacy of unconstrained DTW for the correct realignment of chromatograms. It should be emphasized that a different choice of reference (e.g. one that contains the extra peaks) is somewhat helpful in preserving the peaks like those around 8.3–8.6 min, but cannot be considered as a solution to make unconstrained DTW viable. In fact, an analysis of the shifts in the raw data indicates that the small peaks around 8.3–8.6 min are distinct and a $T^{(2,\infty)}$ warping aligns them as if they were one, which is evidently incorrect. It would be very difficult and time-consuming to find the references that contain all the peaks or even most of them. Moreover, using alternative references would not avoid the artifacts, but only show them at different positions in the chromatogram.

Figure 6(b) shows some details of the warped samples using lookup table $T^{(3,1)}$ with 10% band constraints and averaging. A clear reduction in the occurrence of artifacts, both the misalignments and peak-top kind, can be observed from this figure. In particular, the peaks at 8.4–8.6 are no longer merged with the larger one at 8.7–8.8, and are maintained separate. Nonetheless, their widths are modified beyond need as a result of the several vertical transitions (albeit alternated with diagonal ones because of the constraints) in the warping path. Repeated points (due to horizontal transitions) also modify the shape of the peaks, but the effect can be partly reduced using interpolation in the synchronization step (not shown).

In any case, the quality of the alignment cannot be compared with the best one obtained with COW, using a segment length $i = 100$ time-points ($I = 20$ segments) and a slack of $t = 3$ time-points [Figure 6(c)]. The limited correcting flexibility and the interpolation still allowed the correct alignment of the peaks around 8.7–8.8 min while maintaining their original shape and the series of smaller peaks before this retention time are visible as individual entities and not deformed.

The results of DTW are comparable to those of COW only when very rigid constraints are used [Figure 6(d)]. As anticipated, the rules in the lookup table should be larger than the peaks and the best results were obtained with a $T^{(59,1)}$ warping (i.e. with rules spanning 30 points) and using interpolation for the synchronization step. The results do not completely overlap those of COW, as the optimization criterion is different (the squared Euclidean distance was used in DTW), but the discrepancies are marginal and the warped chromatograms are considerably better than those obtained with shorter rules. Note that the corrective power of $T_{\mathrm{COW}}^{(203,3)}$ and $T^{(59,1)}$ is in practice identical. Shorter segments are required for DTWc because in this algorithm interpolation is not concurrent to the computation of the optimal local distances.

When rigid slope constraints were applied, a residual retention time shift, albeit limited to one single point (before or after the position of the matching peak in the reference) at

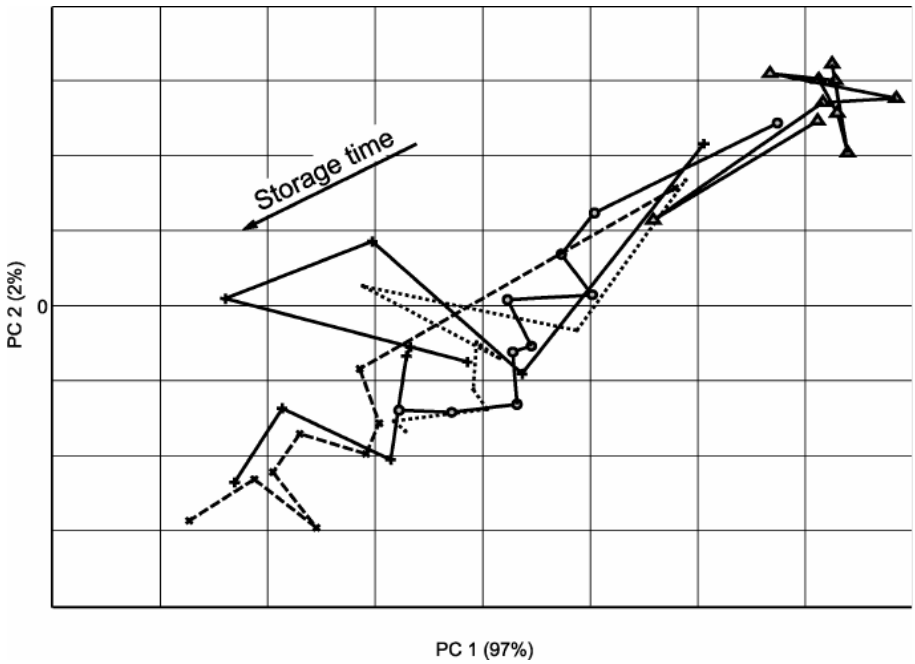


Figure 7. PCA score plots for coffee data set after $T^{(59,1)}$ warping and removal of storage time '3'. Clustering of packaging 'g' (▲), and trends of packagings 'a' (+), 'b' (x), 'c' (●), and 'h' (. . .). The scores relative to the other packagings were removed from the plot for clarity.

the peak's top, could be observed and could not be removed with either COW or DTWc, regardless of the choice of segment length and slack. This residual shift seems to be related to the slightly different peak shape due to noise and the widening associated with column deterioration.

Although the correction was satisfactory in the central part of the chromatograms with both COW and DTWc, as expected, the quality of the realignment degraded at the end of the chromatogram (starting approximately at 11–12 min). Appending a segment of 500 points with low intensity white noise to both sample and reference prior to warping and removing it afterwards effectively dealt with the problem without any modification to the warping algorithm.

From a thorough visual inspection of the warped chromatograms, none of the two algorithms distinctively emerged as better than the other.

The score plot in Figure 5(b) shows a storage-time trend in the first PC, with '1' and 'g' (the most expensive packaging) on the right and '11' on the left. The time-'3' packaging can be considered an outlier, probably the consequence of freezing the extracts before analysis. If this last category is removed, an even clearer storage time trend (particularly packagings 'a', 'b', 'c' 'd' and 'h') and clustering in packaging conditions (namely the already mentioned high-quality 'g') is found, revealing the different capabilities in preserving coffee volatiles (Figure 7). In particular, the second component is related to the ratio between low and high molecular weight fatty acids and the third component describes the 'extra' peaks like those found at 8.3 minutes. Such peaks are present in a relatively small number of samples and are likely indexes of chemical degradation of some of the fatty acids in the ground coffee. They precede all the 'standard' peaks by 0.2–0.5 min and are present only in a few samples of four packaging types ('a', 'd', 'e' and 'h'), but show no evident trend over the entire set. DTWc and COW gave nearly identical results in the PCA analysis (not shown), confirming again the substantial equivalence of the two methods as preprocessing methods for bilinear modeling.

5. CONCLUSIONS

In this paper two different algorithms—correlation optimized warping and dynamic time warping—were studied as a preprocessing step in (bilinear) factor modeling. Both the theoretical properties and some practical implications were investigated on the basis of a case study of chromatographic data vectors with retention time artifacts and a connection between the two methods has been established. The most relevant conclusion is that time alignment corrections should be handled with great care. Simple correction schemes can lead to severe distortion of the signal, and unconstrained (or loosely constrained) DTW is clearly too flexible for the coffee case study and most likely for chromatographic data in general. More rigid settings for both methods where found to be successful as alignment operation, making an exploratory PCA analysis of the coffee data set much more interpretable.

More generally, it should be kept in mind that the original DTW was constructed for pattern recognition, thus with the aim of minimizing the distance between the profile and the possible match. In this context, it has been shown repeatedly that rigid slope constraints increase the error rate [6]. Conversely, when applied to chromatographic data (or for alignment of other time trajectories or even spectra), great attention should be paid to the original shape retaining as much as possible. Thus, slope and segment lengths (rules)

should be chosen so that the warping is as rigid as possible within the limits given by the problems in question.

This observation conflicts to some extent with earlier findings of other researchers on chromatographic data vectors [4]. However, in these earlier works an artificial sinusoidal baseline was introduced [7]. We believe that experience has shown that signal processing without proper de-trending can lead to unrealistic conclusions [24]. Therefore, we opted for studying and showing local alignment capabilities instead of more global behaviors.

Another non-trivial step is finding the correct reference vector. If no prior knowledge is available on the data set or experimental conditions, any vector should potentially be able to serve as such, possibly leading to very different solutions. It is also possible to generate or simulate a reference vector (e.g. the first PCA scaled loading-vector from untreated data), again leading to a distinct preprocessing.

For practical purposes, the COW algorithm and DTWc are relatively insensitive towards the parameters values. Any reasonable choice for segment length and slack will give an indication of the anticipated synchronization performance. Furthermore, due to the relatively small search space, a trial and error approach for finding the best settings is feasible even on a modest computer system. In this respect, COW may be more suitable than DTW, as the number of possible paths is much smaller and, consequently, the memory requirements are significantly reduced.

REFERENCES

1. Bylund D, Danielsson R, Malmquist G, Markides KE. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 2002; **961**: 237–244.
2. Reiner E, Abbey LE, Moran TF, Papamichalis P, Shafer RW. Characterization of normal human cells by pyrolysis–gas-chromatography mass spectrometry. *Biomed. Mass Spectrom.* 1979; **6**: 491–498.
3. Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal. Chem.* 1987; **59**: 649–654.
4. Pravdova V, Walczak B, Massart DL. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* 2002; **456**: 77–92.
5. Itakura F. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. ASSP* 1975; **AS23**: 67–72.
6. Sakoe H, Chiba S. Dynamic-programming algorithm optimization for spoken word recognition. *IEEE Trans. ASSP* 1978; **26**: 43–49.
7. Nielsen NPV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998; **805**: 17–35.
8. Jolliffe IT. *Principal Components Analysis* (2nd edn). Springer: New York, 2002.
9. Jackson JEA. *User's Guide to Principal Components*. Wiley-Interscience: New York, 1991.
10. Munck L, Norgaard L, Engelsen SB, Bro R, Andersson CA. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometr. Intell. Lab. Syst.* 1998; **44**: 31–60.
11. Hillier FS, Liebernan GJ. *Introduction to Operations Research* (7th edn). McGraw-Hill: New York, 2001.
12. Sakoe H, Chiba S. *Proceedings of the International Congress of Acoustics*, Budapest, 1971, paper 20 C13.
13. Myers C, Rabiner LR, Rosenberg AE. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Trans. ASSP* 1980; **28**: 623–635.
14. Sakoe H. 2-Level Dp-matching—dynamic programming-based pattern-matching algorithm for connected word recognition. *IEEE Trans. ASSP* 1979; **27**: 588–595.
15. Myers CS, Rabiner LR. A level building dynamic time warping algorithm for connected word recognition. *IEEE Trans. ASSP* 1981; **29**: 284–297.
16. Kassidas A, MacGregor JF, Taylor PA. Synchronization of batch trajectories using dynamic time warping. *AICHE J.* 1998; **44**: 864–875.
17. Ramaker HJ, van Sprang ENM, Westerhuis JA, Smilde A. Dynamic Time Warping of spectroscopic BATCH data. *Anal. Chim. Acta* 2003; **498**: 133–153.
18. Kassidas A, Taylor PA, MacGregor JF. Off-line diagnosis of deterministic faults in continuous dynamic multivariable processes using speech recognition methods. *J. Process Control* 1998; **8**: 381–393.
19. Gollmer K, Posten C. Supervision of bioprocesses using a dynamic time warping algorithm. *Control Eng. Pract.* 1996; **4**: 1287–1295.
20. Kassidas A. Fault detection and diagnosis in dynamic multivariable chemical processes using speech recognition methods. Ph.D., Mc Master University, Hamilton, Ontario, Canada, 1997.
21. Aach J, Church GM. Aligning gene expression time series with time warping algorithms. *Bioinformatics* 2001; **17**: 495–508.
22. Kruskal JB, Liberman M. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (Reissue). Center for the Study of Language and Information Publications, Leland Stanford University, Stanford, CA 1999; 125–161.
23. Wulfert F, Kok WT, Smilde AK. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Anal. Chem.* 1998; **70**: 1761–1767.
24. Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis, Forecasting and Control* (3rd edn). Prentice-Hall: Upper Saddle River, NJ, 1994.