Automated Alignment of Mass Spectrometry Data Using Functional Geometry

Anuj Srivastava

1 Introduction

The use of mass spectrometry data in profiling metabolites present in a specimen is important in biomarker discovery, enzyme substrate assignment, drug development, and many other applications. Liquid chromatography-mass spectrometry (LC-MS) is a common data collection technique in this area and it provides information about retention times of different metabolites. These metabolites are identified by peaks (high y values) in observed chromatograms at the corresponding retention times (x axis). As stated in [17], it is difficult to reproduce run-to-run retention times across experiments/observations due to instrument instability. Different measurements can exhibit variability in peak locations when observing the exact same specimen. In liquid chromatography, the common causes of such shifts in peak locations are changes in column separation temperature, mobile phase composition, mobile phase flow rate, stationary phase age, etc. We illustrate this issue using an example taken from [7] involving proteomics data collected for patients having therapeutic treatments for Acute Myeloid Leukemia. This example studied earlier in [16] and shown here in Fig. 1 displays two chromatograms in blue and red, representing the same chemical specimen. Despite having the same chemical contents, the chromatograms show differences in peaks locations throughout the domain. Due to these shifts, a simple comparison of chromatograms using standard norms will result in unusually high differences despite representing the same material. Thus, any analysis of such data is faced with the challenge of random nonlinear shifts in the peaks, that needs to be reconciled before drawing statistical inferences.

A. Srivastava (✉)
Department of Statistics, Florida State University, Tallahassee, FL 32306, USA
e-mail: anuj@stat.fsu.edu

S. Datta, B.J.A. Mertens (eds.), *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-45809-0_2

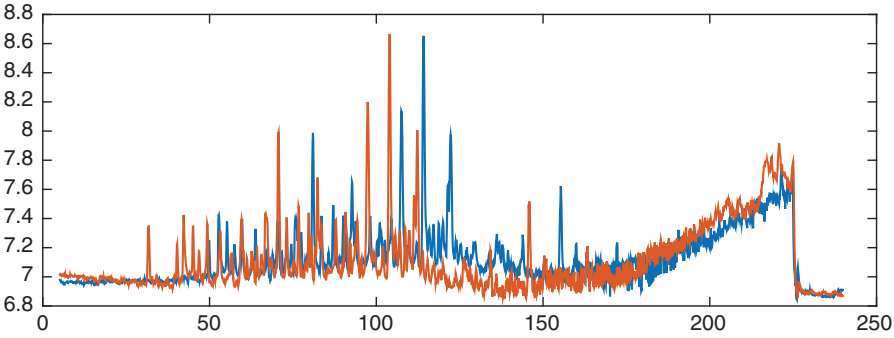


Fig. 1 Example of LC-MS data samples from a proteomics study. Taken from [16]

Consequently, an important goal in LC-MS data analysis is to align peaks in a principled way using some form of nonlinear time-warping.

The literature contains several ideas for handling the nonlinear alignment problem although they are neither fully automatic or nor completely successful in the alignment task. One prominent approach is to find dominant (taller) peaks across chromatograms, match them in a certain way, and then estimate warpings that facilitate that matching. However, the problems of detecting and matching prominent peaks are not straightforward themselves. They are often subjective and require human inputs. Wong et al. [18] developed a peak detection and alignment method that is based exactly on this idea. Bloemberg et al. [2] provide a survey of some current techniques in warping- based alignment of chromatograms, including some computer programs associated with these techniques. Several authors have also developed online tools for spectral alignment, see, e.g., [18]. A statistical approach for spectral alignment and modeling, using a Gaussian mixture model, is presented in [3].

In this paper, we present a fully automated method for alignment of chromatograms, by considering them as functional data and using the nonlinear time-warping of their domains for matching peak locations. The problem of functional data alignment has been studied by several authors, including [4, 6, 9, 14]. In our approach, the actual alignment becomes an optimization problem with a novel objective function which, in turn, is derived using ideas from functional information geometry. Any two chromatograms are aligned by minimizing a distance between them; this distance can be seen as an extension of the classical nonparametric *Fisher-Rao distance*, derived originally for comparing probability density functions, to more general class of functions. Although this distance has nice theoretical properties (invariance to time-warping, etc.), it is too complex to be useful in practical algorithms, especially for processing high-throughput data. This issue is resolved using a change of variable, i.e., replacing the original functions by their square-root slope functions (SRSFs) so that the required distance becomes simply the $\mathbb{L}^2$ norm between their SRSFs. Now, returning to the alignment of

several chromatograms, the basic idea is to derive a mean of the given functions (SRSFs of chromatograms) and to align the individual SRSFs to this mean. This procedure is iterative as the mean itself is updated using the aligned SRSFs. Upon convergence, we obtain aligned SRSFs that can then be mapped back to aligned chromatograms using the inverse maps. This last step relies on the fact that the SRSF representation is invertible up to a constant. This framework, termed the *extended Fisher-Rao framework*, is a fully automated procedure and does not require any peak detection or matching. Instead, it utilizes the aforementioned metric (extended Fisher-Rao) and a corresponding representation (SRSF) to formulate alignment as an optimization problem.

The rest of this paper is as follows. We briefly describe the main mathematical challenges in solving the alignment problem and highlight the limitations of a commonly used approach based on the $\mathbb{L}^2$ norm. In Sect. 3, we lay out our mathematical framework, leading up to the presentation of the automated alignment algorithm. This is followed by experimental results on a number of simulated and real LC-MS datasets in Sect. 4, and the paper ends with a short conclusion in Sect. 5.

2 Fundamental Issues in Functional Alignment

As mentioned above, we view the observed chromatograms as real-valued functions on a fixed interval. Without any loss of generality, we will use $[0, 1]$ as domain of these functions. Let $\mathscr{F}$ momentarily denote the relevant set of real-valued functions on $[0, 1]$ although the precise definition of $\mathscr{F}$ comes later when we present more details. An important problem in statistical analysis of functional data (and specifically in LC-MS data analysis) is the alignment of functions using domain warping. The broad goal of an alignment process is to warp the retention-time (or parameter) axis in such a way that their peaks and valleys are better aligned. This alignment problem has also been referred to as the separation of *phase* and *amplitude* [10], or the *registration* [6], or the correspondence of functions in the given data.

Towards this goal, we need to specify the set of valid warping functions. We will use the set of positive diffeomorphisms of $[0, 1]$ as the set of allowed time-warping function; we will denote it by Γ. It is important to note that Γ is a *group* with composition as the binary operation. For any two elements $\gamma_1, \gamma_2 \in \Gamma$, their composition $\gamma_1 \circ \gamma_2 \in \Gamma$. The identity function $\gamma_{\text{id}}(t) = t$ forms the identity element of Γ and, finally for every $\gamma \in \Gamma$, there exists an inverse element $\gamma^{-1} \in \Gamma$ such that $\gamma \circ \gamma^{-1} = \gamma_{\text{id}}$.

In this context, we can pose two kinds of registration problems:

1. **Pairwise Alignment Problem**: Given any two functions f_1 and f_2 in $\mathscr{F}$, we define their pairwise alignment or registration to be the problem of finding a warping function γ such that a certain energy term $E[f_1, f_2 \circ \gamma]$ is minimized. Figure 2 shows an example of this idea where f_2 is time-warped to align it with f_1.

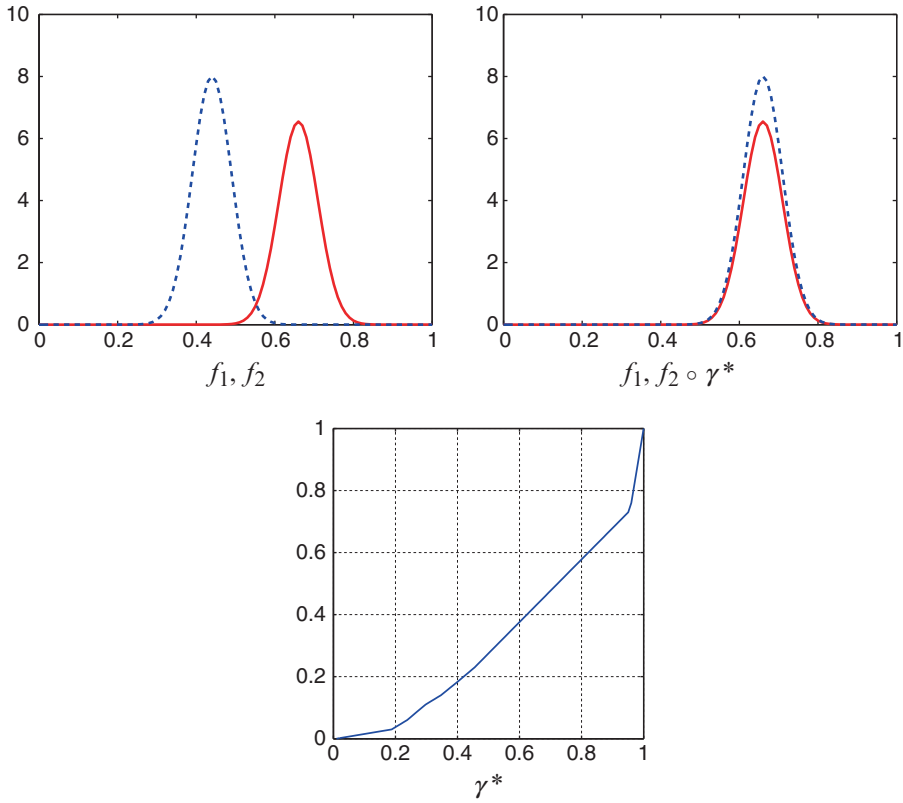


Fig. 2 Illustration of pairwise alignment

2. **Groupwise or Multiple Alignment**: In this case we are given a set of functions $\{f_i \in \mathcal{F} | i = 1, 2, \ldots, n\}$. The problem of finding a set of warping functions $\{\gamma_i | i = 1, 2, \ldots, n\}$ such that, for any $t \in [0, 1]$, the values $f_i(\gamma_i(t))$ are said to be registered with each other, is termed the problem of *joint* or *multiple alignment*. Figure 3 shows an example of this idea where the given functions $\{f_i\}$ (left panel) are aligned (middle panel) using the warping functions shown in the right panel.

We will start by considering the pairwise alignment problem and then later extend that solution to address multiple alignment.

The main question in solving pairwise registration is: What should be the optimization criterion E? In other words, what is a mathematical definition of a *good registration*? Visually one can evaluate an alignment by comparing the locations of peaks and valleys, but how should one do it in a formal, quantifiable and, most importantly, automated way. Before we present our solution, we look at one of the most popular ways of registering functional data in the current literature and highlight its limitations.

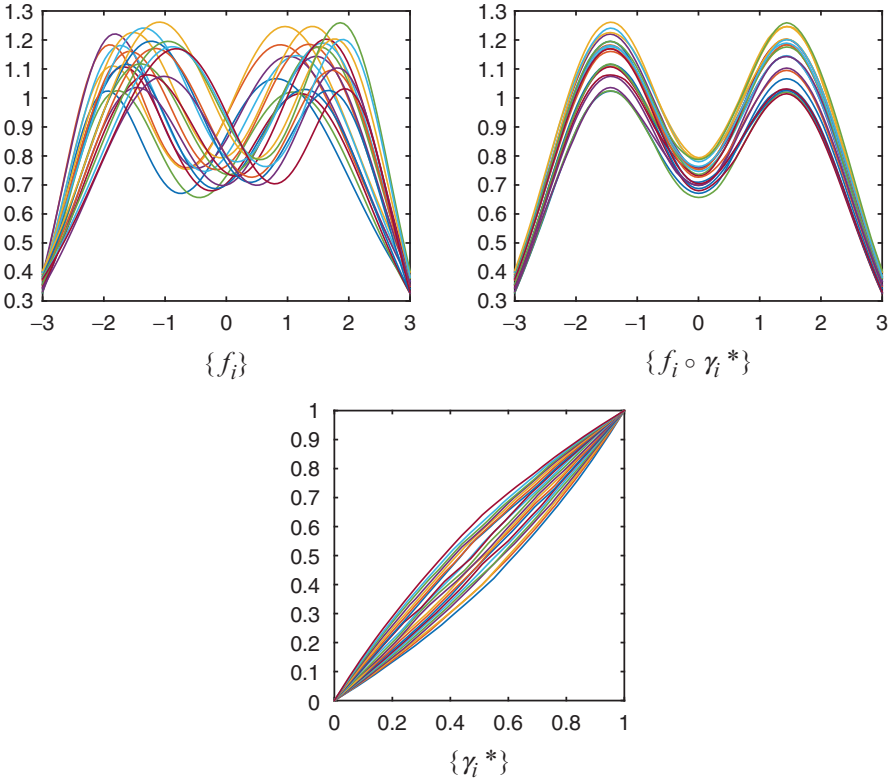


Fig. 3 Illustration of multiple function alignment

2.1 Problems in Using $\mathbb{L}^2$-Norm for Pairwise Registration

When searching for an objective function that measures alignment between f_1 and $f_2 \circ \gamma$, a natural quantity that comes to mind is the $\mathbb{L}^2$ norm of their difference. That is, we can define the optimal warping function to be:

$$\gamma^* = \arg\inf_{\gamma \in \Gamma} \|f_1 - f_2 \circ \gamma\|^2 . \tag{1}$$

It is well known that this formulation is problematic, as it leads to a phenomena called the *pinching effect* [10]. What happens is that in matching of f_1 and f_2 one can squeeze or pinch a large part of f_2 and make this cost function arbitrarily close to zero. An illustration of this problem is presented in Fig. 4 using a simple example. Here a part of f_2 is identical to f_1 over $[0, 0.6]$ and is completely different over the remaining domain $[0.6, 1]$. Since f_1 is essentially zero and f_2 is strictly positive in $[0.6, 1]$, there is no warping that can match f_1 with f_2 over that subinterval.

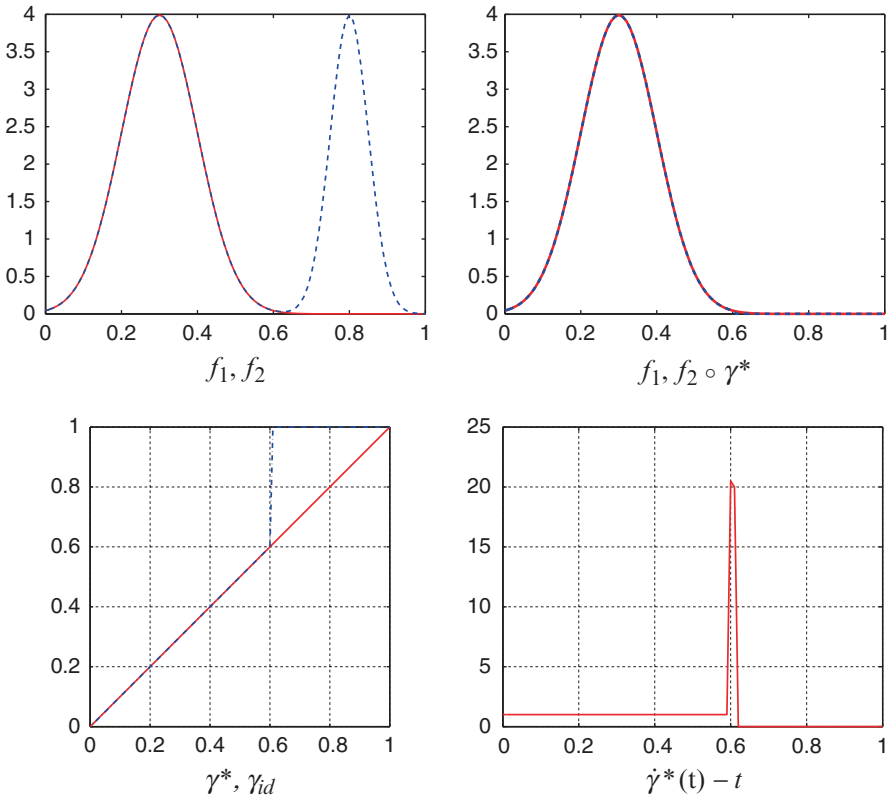


Fig. 4 Illustration of the pinching effect. The *top row* shows a degenerate analytical solution to matching under $\mathbb{L}^2$ norm for functions. The *bottom row* shows the corresponding warping function

The optimal solution is, therefore, to decimate that part of f_2 by using the following γ^*: it coincides with γ_{id} over $[0, 0.6]$, climbs rapidly to $1 - \epsilon$ around 0.6, and then goes slowly from $1 - \epsilon$ to 1 over the interval $[0.6, 1]$. It is easy to check that in the limit:

$$\lim_{\epsilon \to 0} \|f_1 - f_2 \circ \gamma_\epsilon\| = 0 \, .$$

The top-right panel of Fig. 4 shows the limiting case where f_1 and $f_2 \circ \gamma_0$ are identical. In the bottom row, we provide results from a numerical procedure for this optimization problem. Sometimes, in practice, we restrict γ to have positive bounded slope and do not obtain the same result as the theoretical limit. However, one can still see the pinching effect in numerical implementations.

To avoid the pinching problem, one frequently imposes an additional term to the optimization cost, a term that penalizes the roughness of γ. This term, also called a *regularization term*, results in the registration problem of the type:

$$\gamma^* = \arg\inf_{\gamma\in\Gamma} \left(\|f_1 - f_2 \circ \gamma\|^2 + \lambda \mathscr{R}(\gamma) \right) , \tag{2}$$

where $\mathscr{R}(\gamma)$ is the regularization term, e.g., $\mathscr{R}(\gamma) = \int \ddot{\gamma}(t)^2 dt$ and $\lambda > 0$ is a constant. While this solution avoids pinching, it has several other problems including the fact it does not satisfy inverse symmetry. That is, the registration of f_1 to f_2 may lead to a completely different result than that of f_2 to f_1. We illustrate this using an example.

Example 1. As a simple example, let $f_1(t) = t$ and $f_2(t) = 1 + (t - 0.5)^2$. In this case, for the minimization problem $\gamma_{12} = \min_{\gamma\in\Gamma} \|f_2 - f_1 \circ \gamma\|^2$, the optimal solution is as follows. Define a warping function that climbs quickly (linearly) from 0 to $1-\epsilon$ on the interval $[0, \epsilon]$ and then climbs slowly (also linearly) from $1 - \epsilon$ to 1 in the remaining interval $[\epsilon, 1]$. The limiting function, when $\epsilon \to 0$, results in the optimal γ for this case.

For the inverse problem, $\gamma_{21} = \min_{\gamma\in\Gamma} \|f_1 - f_2 \circ \gamma\|^2$, the optimal warping is the following. Define a warping function that rises quickly from 0 to $0.5 - \epsilon$ in the interval $[0, \epsilon]$, climbs slowly from $0.5 - \epsilon$ to $0.5 + \epsilon$ in the interval $[\epsilon, 1 - \epsilon]$, and finally climbs quickly from $0.5+\epsilon$ to 1 in the interval $[1-\epsilon, 1]$. The optimal solution is obtained when $\epsilon \to 0$. A numerical implementation of these two solutions, based on the dynamic programming algorithm [1], are shown in Fig. 5. Since this implementation allows only a limited number of possible slopes for optimal γ, the results are not as accurate as the analytical solution. Still, it is clear to see a large difference in the solutions for the two cases.

2.2 Desired Properties in Alignment Framework

In view of these limitations of the popular $\mathbb{L}^2$-norm-based framework, we first enumerate a set of basic properties that any (alignment) objective function E should satisfy. Actually, only the first one is a fundamental property, the remaining two are simple consequences of the first one (and some additional structure). Still, we list all three of them to highlight different aspects of the registration problem.

1. **Invariance to Simultaneous Warping**: We start by noting that an identical warping of any two functions preserves their registration. That is, for any $\gamma \in \Gamma$ and $f_1, f_2 \in \mathscr{F}$, the function pair (f_1, f_2) has the same registration as the pair $(f_1 \circ \gamma, f_2 \circ \gamma)$. What we mean by that is the application of γ has not disturbed their point-to-point correspondence. This is easy to see since γ is a diffeomorphism. The two height values across the functions that were matched, say $f_1(t_0)$ and $f_2(t_0)$, for a parameter value t_0, remain matched. They are now labeled $f_1(\gamma(t_0))$ and $f_2(\gamma(t_0))$, and the parameter value has changed to $\gamma(t_0)$, but they still have the same parameter and, thus, are still matched to each other. This is illustrated using an example in Fig. 6 where the left panel shows a pair f_1 and f_2. Note that the two functions are nicely registered since their peaks and valleys are perfectly

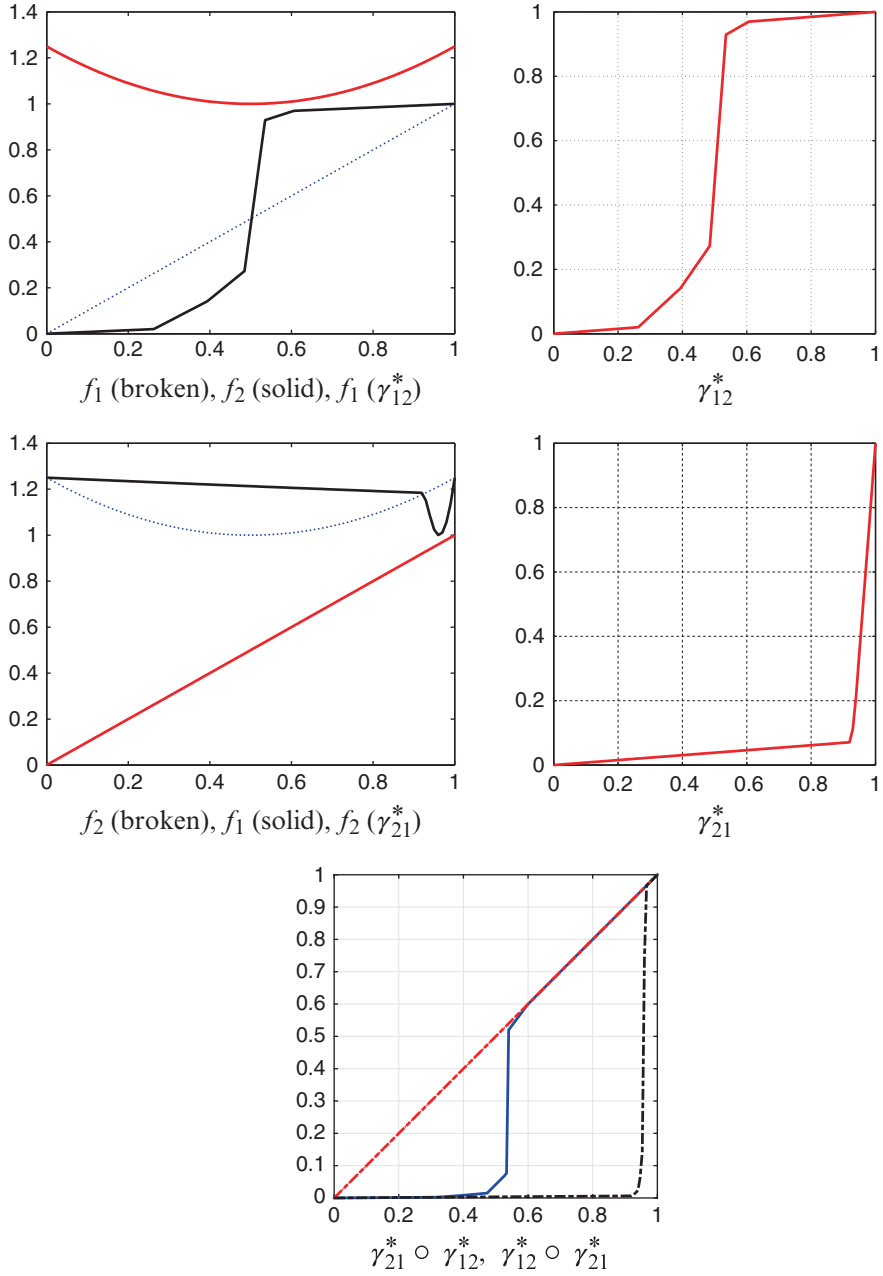


Fig. 5 Example of asymmetry in registration using penalized $\mathbb{L}^2$ norm. The *top row* shows solution of registering f_1 to f_2, while the *bottom two rows* shows the reverse case. *Bottom most* emphasizes that the two γ functions are not inverses of each other

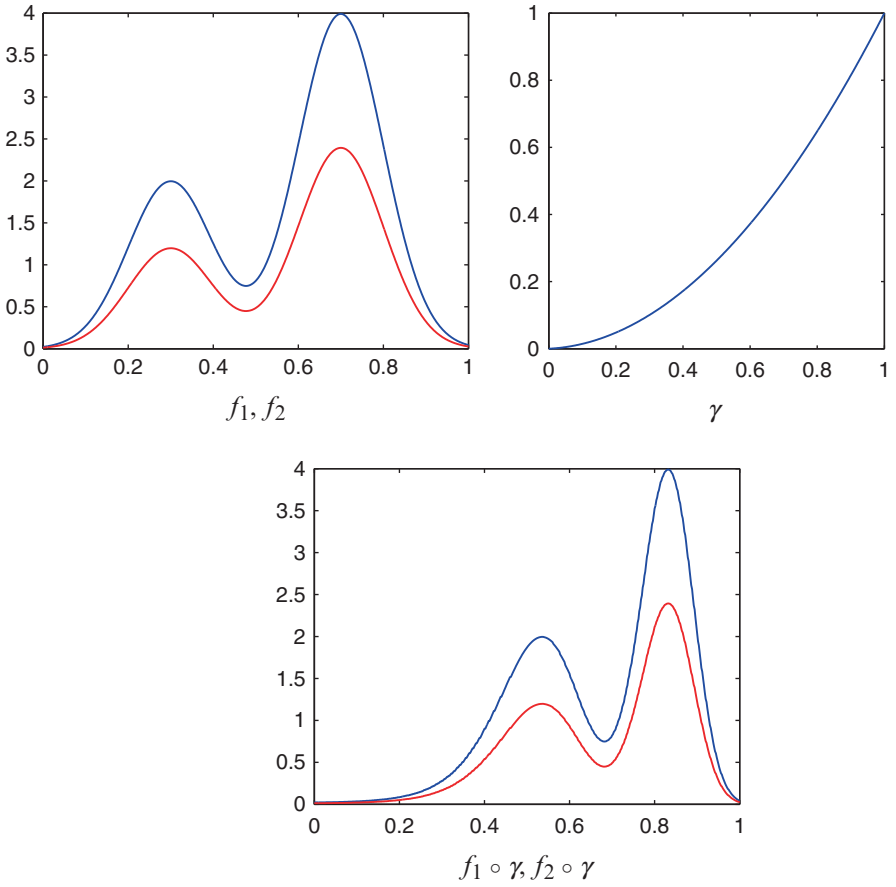


Fig. 6 Identical warping of f_1, f_2 by γ preserves their registration

aligned. If we apply the warping function shown in the top right panel to both their domains, we get the functions shown in the bottom. The peaks and valleys, and in fact all the points, have the same matching as before. This motivates the following invariance property of E. Since E is expected to be a measure of registration of two functions, it should remain unchanged if the two functions are warped identically. That is, for all $f_1, f_2 \in \mathscr{F}$:

$$\textbf{Invariance Property}: \quad E[f_1, f_2] = E[f_1 \circ \gamma, f_2 \circ \gamma], \quad \text{for all } \gamma \in \Gamma. \tag{3}$$

2. **Effect of Random Warpings**: Suppose we have found the optimal warping function $\gamma^* \in \Gamma$, defined by

$$\gamma^* = \underset{\gamma \in \Gamma}{\operatorname{argmin}} E[f_1, f_2 \circ \gamma] \, .$$

Once this γ^* is found, the resulting matched height values are $f_1(t)$ and $f_2(\gamma^*(t))$, for all t. Now, let us suppose that we warp f_1 and f_2 by random functions, say γ_1 and γ_2. What is the optimal correspondence between these new functions? It will be given by

$$\tilde{\gamma}^* = \underset{\gamma \in \Gamma}{\operatorname{argmin}} E[f_1 \circ \gamma_1, (f_2 \circ \gamma_2) \circ \gamma] = \underset{\gamma \in \Gamma}{\operatorname{argmin}} E[f_1, f_2 \circ (\gamma_2 \circ \gamma \circ \gamma_1^{-1})],$$

where the last equality follows from the above *invariance property*. The last two equations immediately imply that $\gamma^* = \gamma_2 \circ \tilde{\gamma}^* \circ \gamma_1^{-1}$, or $\tilde{\gamma}^* = \gamma_2^{-1} \circ \gamma^* \circ \gamma_1$. More interestingly, the optimal registration of functions and the minimum value of E remains unchanged despite the presence of random γ_1 and γ_2, i.e.,

$$\min_{\gamma \in \Gamma} E[f_1, f_2 \circ \gamma] = \min_{\gamma \in \Gamma} E[f_1 \circ \gamma_1, (f_2 \circ \gamma_2) \circ \gamma] \, .$$

This important equality intimately depends on the invariance property [Eq. (3)] and the group structure of Γ. Without the invariance property one can expect the results of registration to be highly dependent on γ_1 and γ_2. For instance, this undesirable situation will occur if we use the $\mathbb{L}^2$ metric between the functions to define E.

3. **Inverse Symmetry**: We know that registration is a symmetric property. That is, if f_1 is registered to f_2, then f_2 is also registered to f_1. Similarly, if f_1 is optimally registered to $f_2 \circ \gamma$, then f_2 is optimally registered to $f_1 \circ \gamma^{-1}$. Therefore, the choice of E should be such that this symmetry is preserved. That is,

$$\gamma^* = \underset{\gamma \in \Gamma}{\operatorname{argmin}} E[f_1, f_2 \circ \gamma] \Rightarrow \gamma^{*-1} = \underset{\gamma \in \Gamma}{\operatorname{argmin}} E[f_1 \circ \gamma, f_2] \, .$$

This symmetry property has also been termed as *inverse consistency*. If the invariance property holds, then this inverse symmetry follows immediately from the group structure of Γ.

In addition to registering any two functions, it is often important to compare them and to quantify their differences. For this, we need a proper distance function, to be able to compare f_1 and $f_2 \circ \gamma^*$, where γ^* is the optimal warping of f_2 that registers it with f_1. One can always choose an unrelated distance function on the space, e.g., the $\mathbb{L}^2$ norm, that measures differences in the registered functions f_1 and $f_2 \circ \gamma^*$. However, this makes the process of registration an unrelated pre-processing step for the eventual comparison. Ideally, we would like to jointly solve these two problems, under a unified metric. Therefore, it will be useful if the quantity $\inf_{\gamma \in \Gamma} E[f_1, f_2 \circ \gamma]$ is also a proper distance in some sense. That is, in addition to symmetry, it also satisfies non-negativity and the triangle inequality. The sense in which we want it to be a distance is that the result does not change if we randomly warp the individual functions in arbitrary ways.

3 Extended Fisher-Rao Approach

In order to reach a cost function E that avoids the pinching effect, satisfies the *invariance property* and, as a consequence, allows inverse symmetry and invariance of E to random warping, we introduce a new mathematical representation for functions.

3.1 Mathematical Background

Our general framework for alignment of chromatograms is adapted from ideas in shape analysis of curves [5, 12] and is described more comprehensively in [8, 13, 15]. For a broader introduction to this theory, including asymptotic results and identifiability results, we refer the reader to these papers.

Starting fresh, this time we are going to restrict to those f that are absolutely continuous on $[0, 1]$; let $\mathscr{F}$ denote the set of all such functions. The new representation of functions is based on the following transformation. Define a mapping: $Q : \mathbb{R} \to \mathbb{R}$ according to:

$$Q(x) \equiv \begin{cases} \text{sign}(x)\sqrt{|x|}, & x \neq 0 \\ 0, & x = 0 \end{cases} . \tag{4}$$

Note that Q is a continuous map. For the purpose of studying the function f, we will represent it using the SRSF defined as follows:

Definition 1 (SRSF Representation of Functions). Define the SRSF of f to be the function $q : [0, 1] \to \mathbb{R}$, where

$$q(t) \equiv Q(\dot{f}(t)) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|} .$$

This representation includes those functions whose parameterization can become singular in the analysis. In other words, if $\dot{f}(t) = 0$ at some point, it does not cause any problem in the definition of $q(t)$. It can be shown that if the function f is absolutely continuous, then the resulting SRSF is square integrable [11]. Thus, we will define $\mathbb{L}^2([0, 1], \mathbb{R})$ (or simply $\mathbb{L}^2$) to be the set of all SRSFs. For every $q \in \mathbb{L}^2$ there exists a function f (unique up to a constant) such that the given q is the SRSF of that f. In fact, this function can be obtained precisely using the equation: $f(t) = f(0) + \int_0^t q(s)|q(s)|ds$. Thus, the representation $f \Leftrightarrow (f(0), q)$ is invertible.

The next question is: If a function is warped, then how does its SRSF change? For an $f \in \mathscr{F}$ and $\gamma \in \Gamma$, let q be the SRSF of f. Then, what is the SRSF of $f \circ \gamma$? This can simply be derived as:

$$\tilde{q}(t) = Q\left(\frac{d}{dt}(f\circ\gamma)(t)\right) = \text{sign}\left(\frac{d}{dt}(f\circ\gamma)(t)\right)\sqrt{|\frac{d}{dt}(f\circ\gamma)(t)|} = q(\gamma(t))\sqrt{\dot{\gamma}(t)}\,.$$

In more mathematical terms, this denotes an action of group Γ on $\mathbb{L}^2$ from the right side: $\mathbb{L}^2 \times \Gamma \to \mathbb{L}^2$, given by $(q,\gamma) = (q\circ\gamma)\sqrt{\dot{\gamma}}$. One can show that this action of Γ on $\mathbb{L}^2$ is *compatible* with its action on $\mathscr{F}$ given earlier, in the following sense:

$$\begin{array}{ccc} & f & \stackrel{\text{SRSF}}{\longrightarrow} & q & \\ \text{action on } \mathscr{F} & \downarrow & & \downarrow & \text{action on } \mathbb{L}^2 \\ & f\circ\gamma & \stackrel{\text{SRSF}}{\longrightarrow} & (q\circ\gamma)\sqrt{\dot{\gamma}} & \end{array}$$

We can apply the group action and compute SRSF in any order, and the result remains the same. The most important advantage of using SRSFs in functional data analysis comes from the following result. Recall that the $\mathbb{L}^2$ inner-product is given by: $\langle v_1, v_2\rangle = \int_0^1 v_1(t)v_2(t)\,dt$.

Lemma 1. *The mapping $\mathbb{L}^2 \times \Gamma \to \mathbb{L}^2$ given by $(q,\gamma) = (q\circ\gamma)\sqrt{\dot{\gamma}}$ forms an action of Γ on $\mathbb{L}^2$ by isometries.*

Proof. That the mapping is a group action has been mentioned earlier. The proof of isometry is an easy application of integration by substitution. For any $v_1, v_2 \in \mathbb{L}^2$,

$$\begin{aligned}\langle (v_1,\gamma),(v_2,\gamma)\rangle &= \int_0^1 v_1(\gamma(t))\sqrt{\dot{\gamma}(t)}v_2(\gamma(t))\sqrt{\dot{\gamma}(t)}dt \\ &= \int_0^1 v_1(\gamma(t))v_2(\gamma(t))\dot{\gamma}(t)dt = \int_0^1 v_1(s)v_2(s)ds = \langle v_1, v_2\rangle\ .\end{aligned}$$

□

This lemma ensures that any framework based on this SRSF and $\mathbb{L}^2$ norm will satisfy the *invariance property* listed in the previous section.

It is well known that the geodesics in $\mathbb{L}^2$, under the $\mathbb{L}^2$ Riemannian metric, are straight lines and the geodesic distance between any two elements $q_1, q_2 \in \mathbb{L}^2$ is given by $\|q_1 - q_2\|$. Since the action of Γ on $\mathbb{L}^2$ is by isometries, the following result is automatic. Still, for the sake of completeness, we provide a short proof.

Lemma 2. *For any two SRSFs $q_1, q_2 \in \mathbb{L}^2$ and $\gamma \in \Gamma$, we have that $\|(q_1,\gamma) - (q_2,\gamma)\| = \|q_1 - q_2\|$.*

Proof. For an arbitrary element $\gamma \in \Gamma$, and $q_1,\ q_2 \in \mathbb{L}^2$, we have

$$\begin{aligned}\|(q_1,\gamma) - (q_2,\gamma)\|^2 &= \int_0^1 (q_1(\gamma(t))\sqrt{\dot{\gamma}(t)} - q_2(\gamma(t))\sqrt{\dot{\gamma}(t)})^2 dt \\ &= \int_0^1 (q_1(\gamma(t)) - q_2(\gamma(t)))^2\dot{\gamma}(t)dt = \|q_1 - q_2\|^2\ .\end{aligned}$$

□

An interesting corollary of this lemma is the following.

Corollary 1. *For any $q \in \mathbb{L}^2$ and $\gamma \in \Gamma$, we have $\|q\| = \|(q, \gamma)\|$.*

This implies that the action of Γ on $\mathbb{L}^2$ is actually a norm-preserving transformation. Conceptually, it can be equated with the rotation of vectors in Euclidean spaces. Due to the norm-preserving nature of warping in this representation, the pinching effect is completely avoided.

3.2 Pairwise Alignment Procedure

With this mathematical foundation, the pairwise alignment of chromatograms can be accomplished as follows. Let f_1, f_2 by functional forms of the two given spectra and let q_1, q_2 be the corresponding SRSFs. Then, the optimization problem is given by:

$$\inf_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|^2 = \inf_{\gamma \in \Gamma} \|q_2 - (q_1, \gamma)\|^2 \,. \tag{5}$$

This minimization is performed in practice using a numerical approach called *the dynamic programming algorithm* [1]. We have already mentioned (in Lemma 2) that the use of SRSFs and $\mathbb{L}^2$ norm satisfies the invariance property from Sect. 2.2. We now show that the remaining two properties—effect of random warpings and inverse symmetry—are also satisfied. Using Lemma 2 and the group structure of Γ one can show that if γ^* is a minimizer on the left side of Eq. (5), then $(\gamma^*)^{-1}$ is a minimizer on the right side! Furthermore, using the same tools one can show that:

$$\inf_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\| = \inf_{\gamma \in \Gamma} \|(q_1, \gamma_1) - ((q_2, \gamma_2), \gamma)\| \,,$$

for any $\gamma_1, \gamma_2 \in \Gamma$.

Figure 7 shows some results from using this alignment framework using some simple LC-MS chromatograms. In each of the four examples shown there, the left panel shows the original chromatograms f_1, f_2, the middle panel shows the aligned chromatograms $f_1, f_2 \circ \gamma^*$, and the last panel shows the optimal warping function γ^*. In the first row, the level of misalignment is relatively small, and a small warping is able to align the peaks. However, as we go down this figure, the level of misalignment increases and it takes an increasing amount of warping to align the peaks. This increasing warping is visible in the corresponding warping functions in terms of their deviations from γ_{id}, the 45° line. Noticeably, the algorithm is quite successful in alignment of peaks for all these datasets. Furthermore, it is fully automatic and requires no parameter tuning or any kind of manual intervention.

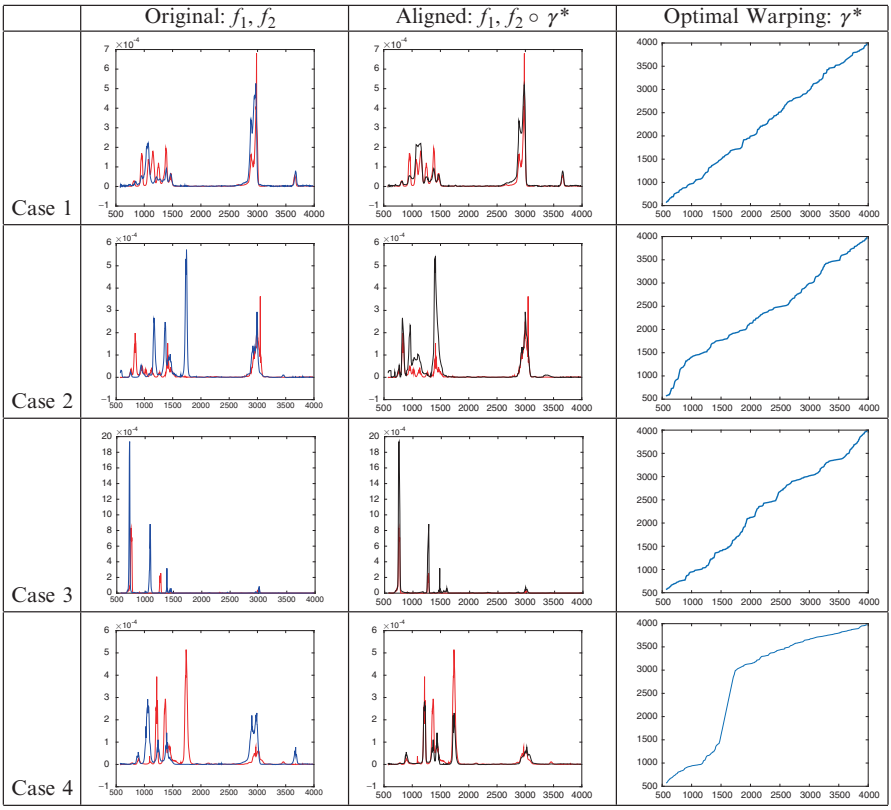


Fig. 7 Pairwise alignment of LC-MS spectra using optimization given in Eq. (5). In each *row* we display the original and the aligned chromatograms, along with the optimal warping functions

3.3 Multiple Alignment Algorithm

With the ability to align chromatograms in a pairwise fashion, we now extend this idea to simultaneous alignment of multiple chromatograms. We will use a template-based approach, where we iteratively define a template chromatogram and align the given chromatograms to this template in a pairwise manner [using Eq. (5)]. The template is created by taking an average of the aligned chromatograms in the SRSF space, at each iteration. The full alignment algorithm is as follows:

Algorithm 1 (Alignment of Multiple Chromatograms). *Given a set of chromatograms in functional form* $f_1, f_2, \ldots f_n$ *on* $[0, 1]$*, let* $q_1, q_2, \ldots, q_n$ *denote their SRSFs, respectively.*

1. *Initialize* $\tilde{q}_i = q_i$ *for* $i = 1, 2, \ldots, n$.
2. *Compute their mean according to* $\mu = \frac{1}{n} \sum_{i=1}^{n} \tilde{q}_i$.
3. *For* $i = 1, 2, \ldots, n$, *find* γ_i^* *by solving:*

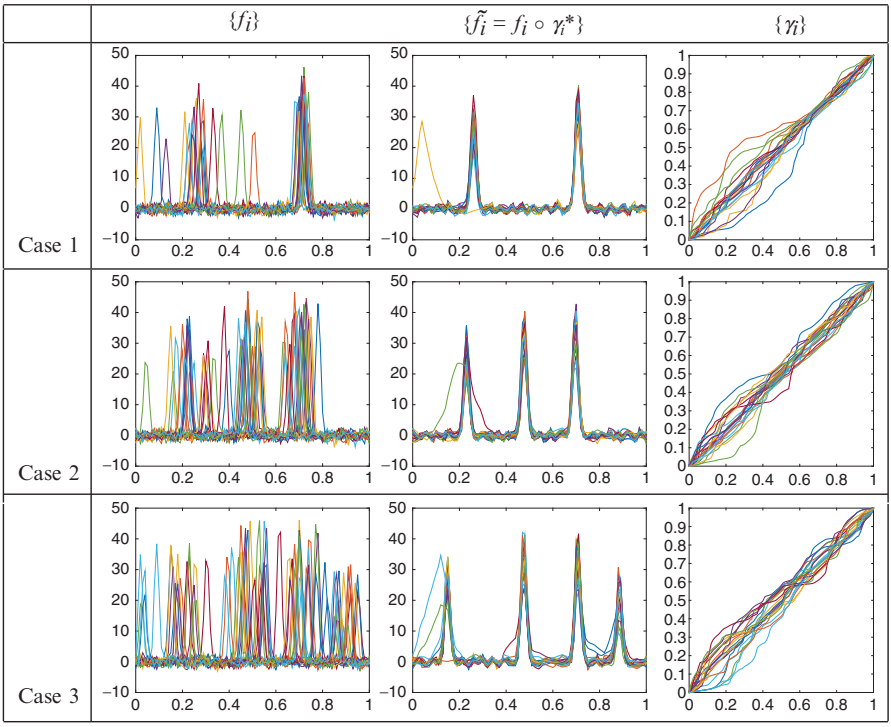


Fig. 8 Alignment of simulated chromatograms using Algorithm 1. In each *row* we see the original functions, the aligned functions, and the corresponding warping functions

$$\gamma_i^* = \underset{\gamma \in \Gamma}{\operatorname{arginf}} \|\mu - (q_i, \gamma)\|^2 .$$

This minimization is approximated using the dynamic programming algorithm.

4. *Compute the aligned SRSFs* $\tilde{q}_i = (q_i, \gamma_i^*)$ *and aligned functions* $\tilde{f}_i = f_i \circ \gamma_i^*$.
5. *Check for convergence. If not converged, go to Step 2.*
6. *Return the template* μ*, the warping functions* $\{\gamma_i^*\}$*, and the aligned functions* $\{\tilde{f}_i\}$.

We show some illustrations of this example using simulated data in Fig. 8. In this experiment, we simulate chromatograms as superpositions of Gaussian probability density functions with random shifts in heights and locations. Each chromatogram is made up of either two (top row), three (middle row), or four (bottom row) Gaussian probability density functions, and we use 20 such chromatograms in each case. Additionally, we corrupt these chromatograms by adding white Gaussian noise at each time t. As can be seen in the middle panels, the algorithm is quite successful in finding and aligning the corresponding peaks across chromatograms. In the process, the algorithm discovers non-trivial, nonlinear warping functions that are required for alignments. We reemphasize that this algorithm is fully automated and does not

require any manual input, nor does it involve any parameter tuning for superior performance. Interestingly, it not only does a good job in aligning major (taller) peaks but it also aligns smaller peaks that can be attributed mainly to noise.

4 Experimental Results on Real Data

In this section we present some alignment results on several LC-MS datasets taken from various sources.

1. As the first result, we utilize the proteomics dataset introduced by Koch et al. [7]. As described there, protein profiling can be used to study changes in protein expression in reference to therapeutic treatments for diseases, and this data involves protein profiles of patients with Acute Myeloid Leukemia. The original data with markers corresponding to the key peaks in the data is presented in Fig. 9 (left panel). An interesting part about this dataset is that it comes with an expert-labeling that provides a unique number to each of the major peaks. This numbering, from 1 to 14 for each spectrum, is used only to study the alignment but are not used in the alignment process itself. As can be seen in the left panel, the peaks in the data are not well aligned as the corresponding numbers demonstrate. The results of applying our alignment method are presented in Fig. 9 (right panel). The aligned functions exhibit good registration with almost all of the peaks lining up. There are a few exceptions involving peaks numbered 1 and 2. Since they have a very low amplitude, their registration is relatively difficult.

   We can also quantify the alignment performance using the decrease in the cumulative cross-sectional variance of the aligned functions. For any functional dataset $\{g_i(t), i = 1, 2, \ldots, n, t \in [0, 1]\}$, let

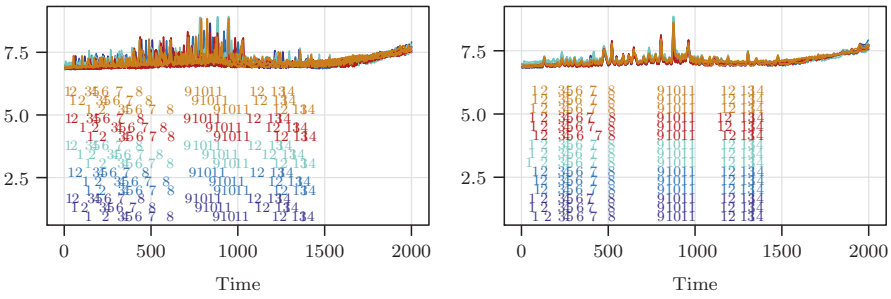


Fig. 9 Alignment of proteomics data using the square-root slope framework with original data in *left panel* and aligned functions in *right panel*. The aligned functions exhibit good registration of marked peaks. Picture taken from [16]

$$\mathrm{Var}(\{g_i\}) = \frac{1}{n-1}\int_0^1 \sum_{i=1}^{n}\left(g_i(t) - \frac{1}{n}\sum_{i=1}^{n} g_i(t)\right)^2 dt\,,$$

denote the cumulative cross-sectional variance in the given data. For the proteomics data, we found

$$\text{Original Variance} = \mathrm{Var}(\{f_i\}) = 4.05, \quad \text{Aligned Variance} = \mathrm{Var}(\{\tilde{f}_i\}) = 1.13$$

$$\text{Warping Variance} = \mathrm{Var}(\{\mu_f \circ \gamma_i^*\}) = 3.04\,.$$

where $\{f_i\}$ is the set of original functions, $\{\tilde{f}_i\}$ is the set of aligned functions, μ_f is the mean of the aligned functions, and $\{\mu_f \circ \gamma_i^*\}$ is the result of applying the warping functions $\{\gamma_i^*\}$ to μ_f. From the decrease in the aligned variance and increase in the warping variance we can quantify the level of alignment.

Figure 10 presents a zoom-in on a region of the data on the time interval [615, 911]. The top panel is the original data where we see very poor alignment of the peaks. The bottom panel is the corresponding aligned data using the extended Fisher-Rao framework, where very tight alignment of the peaks and valleys have occurred.

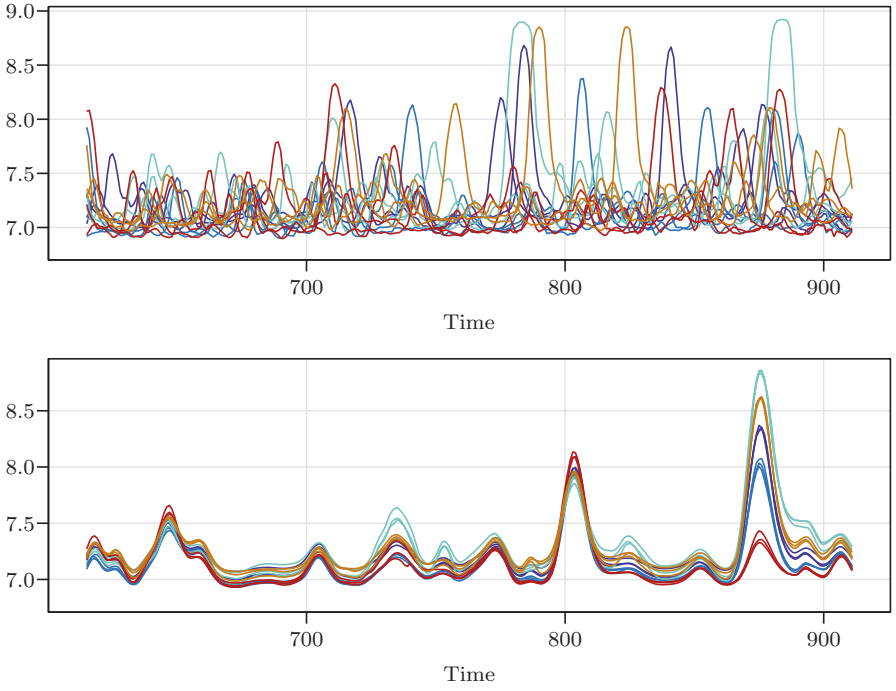


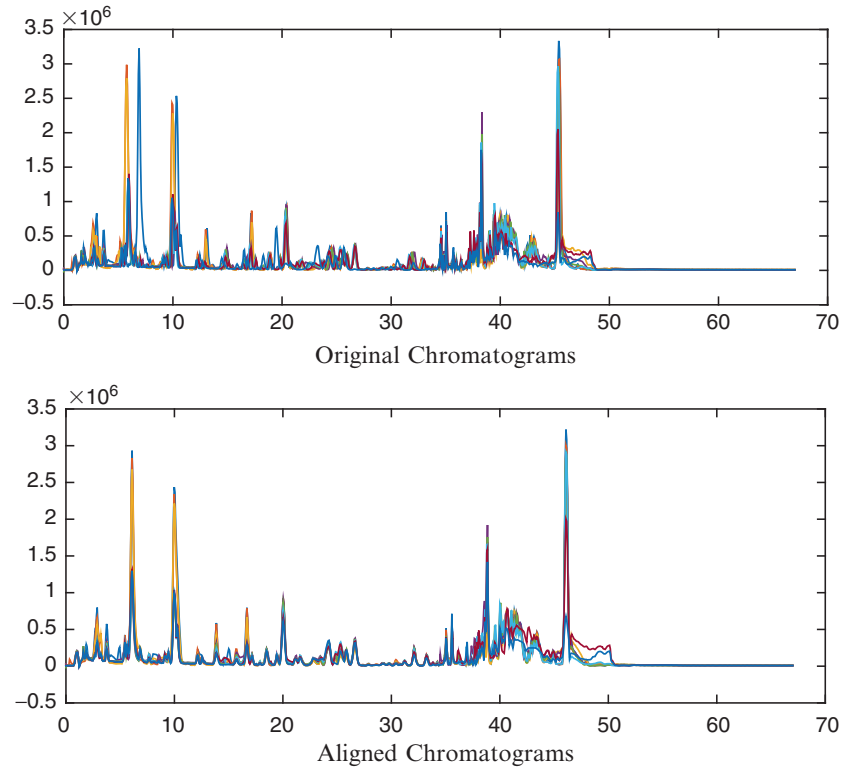
Fig. 10 Zoom-in of the region (600–910) to demonstrate accurate alignment of smaller as well larger peaks. The *top picture* is before and *bottom* is after alignment. Picture taken from [16]

2. Another experiment involves a set of eight chromatograms shown in the top panel in Fig. 11. In this data, some of the major peaks are not aligned, especially in the domain range [15, 25]. The outcome of Algorithm 1 on this data is shown in the second row where the peaks appear to be sharply aligned throughout the spectrum. To emphasize the quality of alignment we look at a couple of smaller intervals in the spectrum more carefully. The zoom-ins of these smaller regions are shown in the last two row of the figure. In each of the last two rows, we show the before and after alignment spectra for these two domains: 0–20 and 28–50. It can be seen in these zoom-ins that the algorithm aligns both the major and minor peaks remarkably well. Once again, this procedure does not require any prior peak detection or matching to reach this alignment.
3. In the third and final example, we study a set of 14 chromatograms associated with urine samples collected at NIST. The top row of Fig. 12 shows the full chromatograms before and after alignment. Since the misalignments in this examples are relatively small, compared with the full range of retention times, it is difficult to evaluate the quality of alignment in this full view. In the bottom row, we look at magnified view of a smaller region—5 to 15—and find the peaks are very closely aligned after the algorithm has been applied.

5 Conclusions

The problem of alignment of mass spectrometry data is both important and challenging task. We have utilized a recent comprehensive approach that treats chromatograms as real-valued functions, uses extended Fisher-Rao metric to perform alignment of peaks and valleys in these functions. The key idea is to form SRSFs of the given chromatograms and then to use the standard $\mathbb{L}^2$ norm between these functions to perform both pairwise and groupwise alignment. We demonstrate this framework using a number of examples involving real and simulated database taken from different spectrometry applications. The success of this alignment procedure is clearly visible in all experiments where the peaks are nicely aligned across observation. This procedure is fully automated and does not require any user input. Furthermore, it aligns full chromatograms (functions) rather than simply matching a few dominant peaks.

Acknowledgements The author is very thankful to the people who provided data for experiments presented in this paper—Prof. I. Koch of Adelaide, South Australia and Dr. Yamil Simon of National Institute of Standards and Technology (NIST), Gaithersburg, Maryland. This research was supported in part by the grants NSF DMS-1208959 and NSF CCF 1319658, and support from the Statistics Division at NIST.



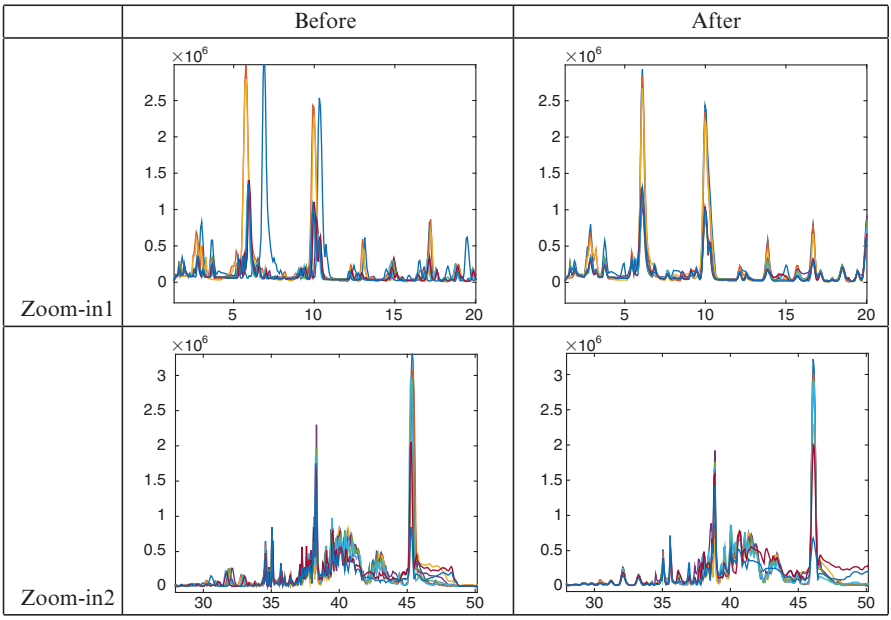


Fig. 11 Alignment of multiple LC-MS chromatograms using Algorithm 1

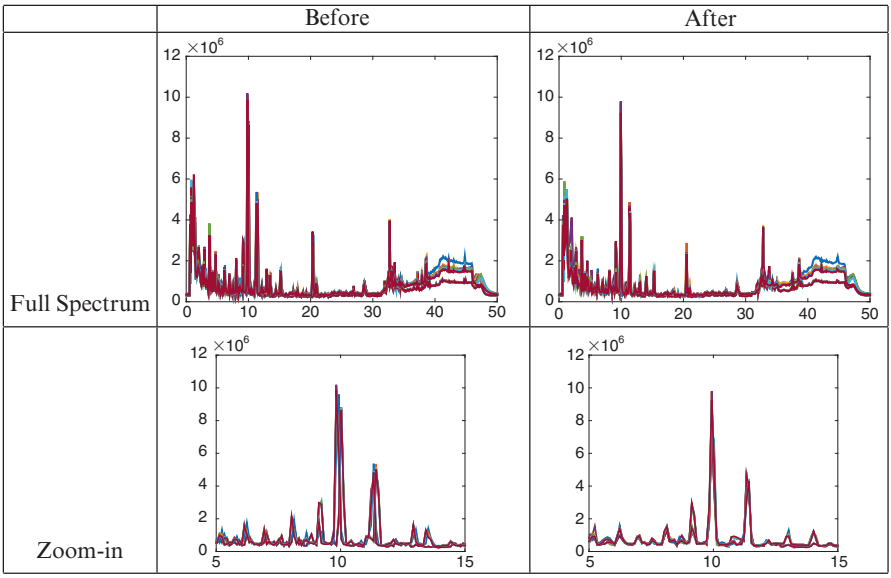


Fig. 12 Alignment of LC-MS chromatograms associated 14 urine samples using Algorithm 1. The *top row* shows the full chromatograms while the *bottom row* shows a magnified region to facilitate a closer look at the alignment performance

References

1. Bertsekas, D. P. (1995). *Dynamic programming and optimal control*. Boston: Athena Scientific.
2. Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., & Buydens, L. M. (2013). Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. *Analytica Chimica Acta, 781*, 14–32.
3. Browne, W. J., Dryden, I. L., Handley, K., Mian, S., & Schadendorf, D. (2010). Mixed effect modelling of proteomic mass spectrometry data by using Gaussian mixtures. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 59*(4), 617–633.
4. James, G. (2007). Curve alignments by moments. *Annals of Applied Statistics, 1*(2), 480–501.
5. Joshi, S. H., Klassen, E., Srivastava, A., & Jermyn, I. H. (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. In *Proceedings of IEEE CVPR* (pp. 1–7).
6. Kneip, A., & Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association, 103*(483), 1155–1165.
7. Koch, I., Hoffmann, P., & Marron, J. S. (2013). Proteomics profiles from mass spectrometry. *Electronic Journal of Statistics, 8*(2), 1703–1713.
8. Kurtek, S., Srivastava, A., & Wu, W. (2011). Signal estimation under random time-warpings and nonlinear signal alignment. In *Proceedings of Advances in Neural Information Processing Systems (NIPS), Grenada, Spain* (pp. 676–683).
9. Liu, X., & Muller, H. G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association, 99*, 687–699.
10. Marron, J. S., Ramsay, J. O., Sangalli, L. M., & Srivastava, A. (2015). Functional Data analysis of amplitude and phase variation. *Statistical Science, 30*(4), 468–484.
11. Robinson, D. (2012, August). *Functional Analysis and Partial Matching in the Square Root Velocity Framework*. PhD thesis, Florida State University.

12. Srivastava, A., Klassen, E., Joshi, S. H., & Jermyn, I. H. (2011, July). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(7), 1415–1428.
13. Srivastava, A., Wu, W., Kurtek, S., Klassen, E., & Marron, J. S. (2011). Registration of functional data using Fisher-Rao metric. arXiv:1103.3817v2 [math.ST].
14. Tang, R., & Muller, H. G. (2008). Pairwise curve synchronization for functional data. *Biometrika, 95*(4), 875–889.
15. Tucker, J. D., Wu, W., & Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis, 61*, 50–66.
16. Tucker, J. D., Wu, W., & Srivastava, A. (2014). Analysis of proteomics data: Phase amplitude separation using an extended Fisher-Rao metric. *Electronic Journal of Statistics, 8*(2), 1724–1733.
17. Wallace, W. E., Srivastava, A., Telu, K. H., & Simon-Manso, Y. (2014). Pairwise alignment of chromatograms using an extended Fisher-Rao metric. *Analytica Chimica Acta, 841*, 10–16.
18. Wong, J. W., Cagney, G., & Cartwright, H. M. (2005). SpecAlign – processing and alignment of mass spectra datasets. *Bioinformatics, 21*(9), 2088–2090.

http://www.springer.com/978-3-319-45807-6

Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry

Datta, S.; Mertens, B.J.A. (Eds.)

2017, VIII, 295 p. 106 illus., 83 illus. in color., Hardcover
ISBN: 978-3-319-45807-6