DISSIMILARITY AND RELATED METHODS FOR FUNCTIONAL DATA

Shuichi Tokushige*, Koichi Inada† and Hiroshi Yadohisa‡

ABSTRACT

Functional data analysis, as proposed by Ramsay (1982), has been attracting many researchers. The most popular approach in recent studies of functional data has been to extend the statistical methods for usual data to functional data. Ramsay and Silverman (1997), for example, proposed regression analysis, principal component analysis, canonical correlation analysis, linear models, etc. for functional data. In this paper, we propose several dissimilarities of functional data. We discuss comparison of these dissimilarities by using the cophenetic correlation coefficient and the sum of squares. Our concern is the effect of dissimilarity on the result of analysis that is applied to dissimilarity data; e.g., cluster analysis.

1. Introduction

Some types of data, such as time series data, are essentially functional data, even though the observations themselves are not continuous. Typical examples of such data include weather data, lip movement data, human growth data, GDP data, and so on. Ramsay (1982) proposed functional data analysis as a data analysis method in cases where the data are a function rather than a vector. Treating data as a function rather than as discrete values has several merits, including the fact that this treatment ensures that such data are not limited to physical time, and the fact that the derivative itself can be treated as functional data with new information.

It is natural to develop multivariate methods of treating functional data by extending classical methods of multivariate analysis. In this paper, we define several dissimilarities for functional data. By using these dissimilarities, we apply hierarchical clustering to weather data.

2. Initial processing and representation of functional data

To treat observations of discrete data as functional data, some initial processing is necessary. In this section, we introduce smoothing and interpolation to join discrete points, and curve registration to align functional data.

*Graduate School of Science and Engineering, Kagoshima University, Korimoto 1-21-35, Kagoshima 890-0065, Japan E-mail: stoku@eniac.sci.kagoshima-u.ac.jp
†Department of Mathematics and Computer Science, Kagoshima University, Korimoto 1-21-35, Kagoshima 890-0065, Japan E-mail: inada@sci.kagoshima-u.ac.jp
‡Department of Mathematics and Computer Science, Kagoshima University, Korimoto 1-21-35, Kagoshima 890-0065, Japan E-mail: yado@sci.kagoshima-u.ac.jp

Key words: Dissimilarity; Distortion; Hierarchical clustering; Registration; Smoothing
2.1. Smoothing and interpolation

Even if data can be essentially described as a function, they are usually observed as discrete values. In this case, some method of smoothing is required in order to treat the data as functional data, because observations usually contain errors. Of course, there are cases when methods of interpolation may be more suitable than smoothing. In this paper, it is assumed that discrete data are transformed to functional data using a suitable method.

2.2. Curve registration

Smooth functions are not always appropriate for purposes of functional data analysis, and another transformation method, called curve registration, is often used to fit functions. In this paper, only simple curve registration is used to define dissimilarities between functional data. For details on curve registration, see Ramsay and Li (1998).

3. Dissimilarities for functional data

To apply cluster analysis or multidimensional scaling to functional data, we need a measure of dissimilarity between the functional data sets. In this section, we define several dissimilarities for functional data.

**Definition 1.** Let the functional data \( f_i(t) \) \( (i = 1, \ldots, N, \ t \in T \subseteq \mathbb{R}) \) be given, where the closed interval \( T \) is the domain of \( t \) and \( \mathbb{R} \) is the set of all real numbers. Then, a dissimilarity, \( d_{ij} \), between \( f_i(t) \) and \( f_j(t) \) is defined by

\[
    d_{ij} = \int_T (f_i(t) - f_j(t))^2 dt.
\]

In some cases, we are not interested in the physical time at which the data were observed. In these cases we may impose curve registration on the data, so we define another dissimilarity for this case.

**Definition 2.** Let the functional data \( f_i(t) \) be registered by the time transformation function \( h(t) \) as

\[
    f_i^*(t) = f_i(h(t)).
\]

The domain of \( f_i(t) \) is assumed to be transformed by some registration and represented by \( T^* \). Then, another dissimilarity between the registered functions \( f_i^*(t) \) and \( f_j^*(t) \) is defined by

\[
    d_{ij}^* = \int_{T^*} (f_i^*(t) - f_j^*(t))^2 dt.
\]

The transformation family \( h \) may be parametric, an example of which is \( h_i(t) = t + \delta_i \), or \( h_i(t) = (t + \delta_i)\beta_i, \beta > 0 \). In this case, \( h_i(t) \) is determined by the iterative minimization of

\[
    \sum_{i=1}^{N} \int_T [f_i(h_i(t)) - \hat{\mu}(t)]^2 dt,
\]

where \( \hat{\mu}(t) \) is a re-estimated mean of the curves \( f_i(h_i(t)) \) at each stage (see, Ramsay and Li, 1998).
Next, we define another dissimilarity for functional data expressed by basis expansion. Let the functional data be expressed by

\[ f_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), \]

where \( \phi_k(t) \) \((k = 1, \ldots, K)\) are basis functions, \( c_{ki} \) are real coefficients and \( K \) is a number of bases.

**Definition 3.** A dissimilarity can be defined as

\[
d_{ij} = \int_T \left( \sum_{k=1}^{K} c_{ik} \phi_k(t) - \sum_{k=1}^{K} c_{jk} \phi_k(t) \right)^2 dt
\]

Furthermore, if the basis is orthonormal, such as a Fourier basis, i.e., \( d_{ij} \) is rewritten in terms of only the coefficients as \( \int \phi_k^2(t) dt = 1 \) for \( k = k_1 = k_2 \) and \( \int \phi_{k_1}(t) \phi_{k_2}(t) dt = 0 \) for \( k_1 \neq k_2 \), then the dissimilarity can be simplified to

\[
d_{ij} = \sum_{k=1}^{K} (c_{ik} - c_{jk})^2 \int_T \phi_k^2(t) dt
\]

Finally, if we are interested in the shape of the function or wish to ignore vertical variation, we pay attention to the inclination, curvature, etc. of the function. In such cases, the derivative itself can be considered to be functional data.

Let the linear differential operator \( L \) of order \( m \) satisfy

\[ Lf = w_0 f + w_1 Df + w_2 D^2f + \cdots + w_m D^m f, \]

where \( D^l f \) is the \( l \)th derivative of \( f \) for \( t \) and \( w_m \) is the weight of each term.

We define dissimilarities between these functions as follows.

**Definition 4.** A dissimilarity between \( Lf_i(t) \) and \( Lf_j(t) \), in which the linear differential operator \( L \) is applied to \( f_i(t) \) and \( f_j(t) \), is defined as

\[
d_{ij}^L = \int_T (Lf_i(t) - Lf_j(t))^2 dt.
\]

**Definition 5.** A dissimilarity between \( Lf_i^*(t) \) and \( Lf_j^*(t) \), in which the linear differential operator \( L \) is applied to \( f_i^*(t) \) and \( f_j^*(t) \), is defined as

\[
d_{ij}^{L*} = \int_{\tau^*} (Lf_i^*(t) - Lf_j^*(t))^2 dt.
\]
4. Comparison of dissimilarities

In this section we compare the dissimilarities defined in Section 3 using weather data. Cluster analysis is also used for purposes of comparison.

Data for daily means of temperature, precipitation, and hours of sunlight for each of 56 stations distributed across Japan are used to create dissimilarities. Each raw data set observed at any given station consists of a vector of 365 observations.

We apply a smoothing method, based on Fourier expansion, to the raw data. Four kinds of dissimilarities are calculated for the smoothed data, based on definitions 1(=3), 2, 4, and 5, which are abbreviated to S, SR, SD, and SRD, respectively. Analyzing functional data, not only the functional data itself but also its derivative have some meaningful information. Here, as a sample, $w_1 = 1$ and $w_2 = 0$ ($i \neq 1$) are used for calculating of SD and SRD. For comparison, dissimilarities between the raw data are calculated by the usual vector norm.

Figure 1 represents the raw data and the four different functional data sets for the three types of weather data. These are used to calculate dissimilarities.

4.1. Distortion between dissimilarity matrices

The cophenetic correlation coefficient; $CC$ (Sokal and Rohlf, 1962), and the sum of squares, $SS$ (Hartigan, 1967), are calculated in Table 1. The upper values and lower values of Table 1 represent $CC$ and $SS$, respectively. The upper triangle shows distortions between dissimilarities of temperature and the lower triangle shows the dissimilarities of precipitation in the upper table and dissimilarities of hours of sunlight in the lower table.

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>S</th>
<th>SR</th>
<th>SD</th>
<th>SRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>1</td>
<td>0.99999</td>
<td>0.99997</td>
<td>0.81016</td>
<td>0.80278</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.00005×10^{12}</td>
<td>0.00017×10^{12}</td>
<td>1.54762×10^{12}</td>
<td>1.54762×10^{12}</td>
</tr>
<tr>
<td>S</td>
<td>0.18852</td>
<td>1</td>
<td>0.99998</td>
<td>0.81084</td>
<td>0.80351</td>
</tr>
<tr>
<td></td>
<td>1.35074×10^{12}</td>
<td>0</td>
<td>0.00005×10^{12}</td>
<td>1.53465×10^{12}</td>
<td>1.53465×10^{12}</td>
</tr>
<tr>
<td>SR</td>
<td>0.20304</td>
<td>0.92307</td>
<td>1</td>
<td>0.81112</td>
<td>0.80406</td>
</tr>
<tr>
<td></td>
<td>1.38233×10^{12}</td>
<td>0</td>
<td>0.00323×10^{12}</td>
<td>1.52703×10^{12}</td>
<td>1.52703×10^{12}</td>
</tr>
<tr>
<td>SD</td>
<td>0.35486</td>
<td>0.57870</td>
<td>0.45126</td>
<td>1</td>
<td>0.99888</td>
</tr>
<tr>
<td></td>
<td>1.54761×10^{12}</td>
<td>0.03362×10^{12}</td>
<td>0.02208×10^{12}</td>
<td>0</td>
<td>4.58824</td>
</tr>
<tr>
<td>SRD</td>
<td>0.23743</td>
<td>0.55109</td>
<td>0.57246</td>
<td>0.78830</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.54766×10^{12}</td>
<td>0.03362×10^{12}</td>
<td>0.02208×10^{12}</td>
<td>1248.41</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours of sunlight</th>
<th>Raw</th>
<th>S</th>
<th>SR</th>
<th>SD</th>
<th>SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>1</td>
<td>0.35485</td>
<td>0.29551</td>
<td>0.64958</td>
<td>0.47177</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.48319×10^{12}</td>
<td>1.50657×10^{12}</td>
<td>1.54768×10^{12}</td>
<td>1.54769×10^{12}</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>0.71335</td>
<td>0.78859</td>
<td>0.45086</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.00054×10^{12}</td>
<td>0.00195×10^{12}</td>
<td>0.00196×10^{12}</td>
<td></td>
</tr>
<tr>
<td>SR</td>
<td>1</td>
<td>0.62423</td>
<td>0.68407</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.00086×10^{12}</td>
<td>0.00086×10^{12}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>1</td>
<td>0.77430</td>
<td>81.5682</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRD</td>
<td>1</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Upper values: Cophenetic correlation; Lower values: Sum of squares.
Dissimilarity and Related Methods for Functional Data

Fig. 1: Weather data
All values of $CC$ for dissimilarities in the temperature data are large, that is, the distortions between dissimilarities from the raw data and other dissimilarities of smoothed data are small, and distortions within the dissimilarities of smoothed data are also small in terms of correlations compared to precipitation and hours of sunlight data. In other words, there is not much value to treating this data as functional compared to the two others. On the other hand, almost all values of $CC$ for dissimilarities in precipitation data and in the hours of sunlight data are small compared to temperature data. We believe that there are some means, or importance, which depend on the dissimilarities in the functional data.

We obtained similar results from evaluating the SS values.

4.2. Cluster analysis (the group average method)

Here, we applied hierarchical clustering (the group average method) to fifteen dissimilarities calculated from the data indicated in Figure 1. Figure 2 provides a dendrogram from $S$ of temperature. There are three clusters; Hokkaido, Okinawa, and others. We obtained similar clusters by analyzing SR of temperature. Figure 3 provides a dendrogram from SD of temperature, in which there are two clusters; Okinawa and others. This result was also obtained from SRD of temperature. And a dendrogram from Raw coincides with one from $S$. Figure 4 shows the dendrogram from $S$ of precipitation; we could not find any meaningful cluster corresponding to areas or latitude. We obtained the same results from other dissimilarities of the precipitation data. Figure 5 shows the dendrogram from $S$ of hours of sunlight. From this, 5 clusters can be obtained: i) the Pacific coast, ii) the Sea of Japan coast, iii) the Inland Sea of Japan region, iv) Hokkaido, and v) Okinawa. We consider these clusters to correspond to latitude. We obtained similar clusters by analysis of SR, SD, and SRD of hours of sunlight. A cluster obtained from Raw is different from the others. Almost objects are divided into 5 clusters; i) Pacific, ii) East Japan, iii) West Japan, iv) Hokkaido, and v) Okinawa.

Fig. 2: Dendrogram from $S$ of temperature
Fig. 3: Dendrogram from SD of temperature

Fig. 4: Dendrogram from S of precipitation
5. Concluding remarks

In this paper, we defined the dissimilarities between functional data and compared these using the cophenetic correlation coefficient and the sum of squares. For comparison, we analyzed data using a clustering method. Of course, we analyzed only a specific data set, so we cannot derive a general result from this analysis. For further work, we would like to reveal the effects of dissimilarities of functional data on the results of analysis, through a simulated study.

REFERENCES


