Analysis of Distribution Valued Data using Techniques of FDA

Masahiro Mizuta*
Hokkaido University, Sapporo Japan mizuta@iic.hokudai.ac.jp

Abstracts

In this article, we discuss methods for analysis of distribution valued data as a kind of symbolic data. Most methods for data analysis assume that the data are sets of numbers with structure. For example, typical multivariate data are identified as a set of *n* vectors of real numbers and dissimilarity data on pairs of *n* objects are as matrix. However, requests for analysis of data with new models become higher, as the kind and quantity of the data is increased. In order to overcome this problem, Symbolic Data Analysis (SDA) supplies various data descriptions; interval valued data, modal interval data, categorical data, distribution valued data, *etc*. Distribution valued data is fruitful because of its ability of expression. An approach to analyze distribution valued data is the use of functions (density function, cumulative distribution function, quantile function *etc*.) We focus on quantile functions to describe the objects or data. Another great approach to deal with such complex data is Functional Data Analysis (FDA.) In FDA, the objects or data are described by functions. Then we can use techniques of FDA to analyze the quantile functions. We introduce methods of clustering and multidimensional scaling (MDS) for distribution valued data and related topics.

Key Words: Symbolic data analysis, functional data analysis, quantile function

1. Introduction

Distribution is a fundamental concept in Statistics. Prof. Diday, the founder of symbolic data analysis (SDA), quoted Schweizer as saying "distributions are the number of the future" (Schweizer;1985, Diday; 2008). From a viewpoint of SDA, distributions are adequate descriptions of the objects or concepts. We call these kinds of data distribution valued data. Most of distribution valued data can be represented by functions; density function, distribution function, quantile function *etc*. Functional Data Analysis (FDA) is another great approach to deal with complex data, which treats data as functions. This means that FDA has potential to deal with distribution valued data. In this paper, we will discuss mathematical tools for analysis of distribution valued data basis of FDA and will introduce related topics.

2. Distribution Valued Data

The targets of SDA are called as concepts. Typical case for concepts is a set of individuals, which have scalars (*i.e.* variables) for their attributes. If we use all scalars of the individuals that belong to the concept, we need much computation time and it is difficult to depict its feature. The distribution of the scalars in the concept can be utilized. In other words, distribution valued data; descriptions of concepts are distributions.

There are several approaches for analysis of distribution valued data. The simplest approach is to use the averages of the distributions. We can use most conventional methods on them. Of course, this approach ignores internal variations of the distribution valued data. Another approach is that we restrict distributions to a certain family of distributions, for example the family of normal distributions. We can describe any normal distribution with a few parameters. The third approach is the use of empirical quantile functions (inverse distribution functions). The domain of quantile function is the closed interval [0,1]. We adopt the third approach in this paper.

There are mainly two typical types of distribution valued data. One is that the descriptions of objects are distribution values; x_i $(i = 1,2,\cdots,n)$ are distributions. The other is that the descriptions of dissimilarities are distribution values; s_{ij} $(i,j = 1,2,\cdots,n)$ are distributions. We mainly use the latter type.

3. Functional Data Analysis

One type of the complex data is functional data structure; data themselves are represented as functions. Typical functional data are time series data. There are many other functional data, of course. Ramsay and Silverman have studied function data analysis (FDA) as the analysis method to function data from the 1990's. Ramsay and Silverman (1997) is an excellent textbook on FDA. They published another book for applications on FDA (Ramsay and Silverman, 2002).

Typical functional data are given by a set of functions: $x_i\ (t)$ $(i = 1,2,\cdots,n)$. There are many techniques in FDA including functional regression analysis, functional principal components analysis, functional discriminant analysis, functional canonical correlation and functional clustering. We can get excellent lists of bibliography on FDA from *http://ego.psych.mcgill.ca/misc/fda/index.html*. We have proposed several methods for functional data: functional multidimensional scaling Mizuta (2000), functional clustering Mizuta (2003a,b), *etc*. Functional data can be considered as infinity-dimensional data. Most methods in the book (Ramsay and Silverman, 1997 and 2002) for functional data are based on an approximation with finite expansions of the functions with basis functions. Once the functional data can be thought as finite linear combinations of the basis functions, functional data analysis methods for the functional data are almost the same as those of conventional data analysis methods. But, there are some possibilities of using different approaches.

4. Clustering and MDS for Distribution Valued Data

{Quantile Function}

When we adopt the approach with empirical distributions, it is difficult to represent them with a few parameters. An effective mathematical tool is a quantile function. The quantile function for a distribution is defined as

$$\mathrm{Q}(p) = \inf\{\mathrm{x} \in \mathrm{R}; p \leq \mathrm{F}(\mathrm{x})\}$$

where, $\mathrm{F}(\mathrm{x})$ is the (empirical) distribution function. It is essentially an inverse of $\mathrm{F}(\mathrm{x})$. A distribution can be transformed a quantile function on [0,1]. This means that distribution valued data can be regarded as functional data. We focus on distribution valued dissimilarity data s_{ij} . Typical methods for these data are multidimensional scaling (MDS) and cluster analysis.

{Single Linkage Clustering for Distribution Valued Data}

The first step of the method is to calculate quantile functions of s_{ij}: $\mathrm{Q}_{ij}(p)$ $(i,j = 1,2,\cdots,n)$. Then, we apply the algorithm of Single Linkage for functional dissimilarity data (Mizuta, 2003a). When the value p is fixed, the distribution valued dissimilarities are represented by nonnegative real values. The basic idea of this method is that we apply conventional Single Linkage to $\mathrm{Q}_{ij}(p)$ for each p and get functional Minimum Spanning Tree (MST), say $\mathrm{MST}(p)$. Then we calculate functional configuration and adjust labels of objects. The results of functional single linkage are represented as motions of MST using dynamic or interactive graphics.

{MDS for Distribution Valued Data}

The aim of MDS is to construct the configuration $X = \{x_i; i = 1,2,\cdots,n\}$ that represents the relations among n objects *i.e.* s_{ij}. The configuration is a set of n points on the Euclidean space. There are several methods of conventional MDS including Torgerson's method, Kruskal's method. We have proposed an MDS method for functional dissimilarity data $s_{ij}(p)$ (Mizuta, 2003b). The proposed method needs a conventional MDS method for the first step, *e.g.* Torgerson's method. Then we get primitive configuration $X(p)$. The key idea of the method is to find out functional orthogonal matrix $T(p)$ that adjusts $X(p)$ to $T(p)X(p)$ for almost continuous. If we use quantile functions $s_{ij}(p)$, we can get the configuration by quantile functions.

5. Topics on distribution valued data

In this section, we will introduce actual situations that distribution valued data are important. Radiotherapy plays an important role in the treatment of solid tumors. Fractionated irradiation is performed in most clinical cases to kill tumor cells effectively. A typical fraction schedule is 1.8 to 2.0 Gy per day to a total of 60 to 70 Gy. Many studies so far have discussed alternative treatment regimens, e.g. varying the number of fractions or dose per fraction. We introduce a mathematical method for selecting a single or fractionated irradiation regime based on physical dose distribution adding to biological consideration (Mizuta *et al.*, 2012a,b).

{LQ model}

There are many models in the radiation survival responses of human tumor cells including Linear-Quadratic model (*LQ model*). The LQ model is commonly used to evaluate and compare different fractionation schedules in radiotherapy. The basic assumption in this study relies on the LQ model for both tumors and normal tissues; the formula $\mathrm{E}(d) = \alpha d + \beta d^2$ is used for the effect as a function of absorbed dose d, where α and β are parameters. We can regard $\exp(-\mathrm{E}(d))$ as survival rate. We use the notations α_1 and β_1 for the tumor and α_0 and β_0 for the Organ At Risk (OAR) or normal tissue as the parameters, respectively. In general, these parameters satisfy $\frac{\alpha_0}{\beta_0} < \frac{\alpha_1}{\beta_1}$.

{Radiation Effect on Tumor}

For multifractionated radiation therapy with n-fraction dose d, the radiation effect on the tumor is represented by

$$E_1(d,n) = n(\alpha_1 d + \beta_1 d^2)$$

and fixed as $E_1 = 5ln10$, *i.e.* the survival rate of tumor is $\mathrm{S} = \exp(-E_1) = 10^{-5}$.

{Damage Effect on OAR}

It should be reasonable to consider that the dose for the OAR is proportional to the dose for the tumor, that is, the dose for the OAR is given by δd, where the dose for the tumor is d and the proportionality factor $\delta > 0$.

If we assume that the δ is constant on OAR, the damage effect on OAR is represented by

$$E_0(d,n) = n(\alpha_0 \delta d + \beta_0 (\delta d)^2).$$

If we assume that the δ is not constant on OAR and the density function, *i.e.* differential DVH, is $f_i(\delta)$, the damage effect on OAR is

$$E_0(d,n) = -ln \int_0^1 \exp(-n(\alpha_0 \delta d + \beta_0 (\delta d)^2)\, f_i(\delta) \mathrm{d}\delta.$$

This formulation shows that the damage effect on OAR is determined by n, d and $f_i()$. It is not difficult to solve the constraint optimization problem for each i. We can interpret $E_0(d,n)$ as functional or operator; the feasible range of the effect on Tumor versus the damage effect on OAR corresponds a function. Functional and operator are important mathematical tools for analysis of distribution valued data.

6. Concluding Remarks

In this paper, we dealt with distribution valued data, methods for analysis, and actual application. Of course, there are important problem related distribution valued data. For example, we can access databases on the research of radioactive substance distribution. We can regard these objects (or concepts) as distribution valued data.

References

Diday, E., Noirhomme-Fraiture, M. eds.(2008) *Symbolic Data Analysis and the SODAS Software*, Wiley.

Mizuta, M. (2000) “Functional multidimensional scaling,” *Proceedings of the Tenth Japan and Korea Joint Conference of Statistics*, 77-82.

Mizuta, M.(2003a) “Hierarchical clustering for functional dissimilarity data,” *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, Volume* V, 223-227.

Mizuta, M.(2003b) “K-means method for functional data,” *Bulletin of the International Statistical Institute, 54th Session, Book 2,* 69-71.

Mizuta, M., Takao, S., Date, H., Kishimoto, N., Sutherland, K. L., Onimaru, R., Shirato, H. (2012a) “A Mathematical Study to Select Fractionation Regimen based on Physical Dose Distribution and the Linear-Quadratic Model,” *International Journal of Radiation Oncology, Biology, Physics*, 84(3), 829-833.

Mizuta, M., Date, H., Takao, S., Kishimoto, N., Sutherland, K.L., Onimaru, R., Shirato, H. (2012b) “Graphical Representation of the Effects on Tumor and OAR for Determining the Appropriate Fractionation Regimen in Radiation Therapy Planning,” *Medical Physics*, 39(11):6791-6795.

Ramsay, J. O., Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer-Verlag.

Ramsay, J. O., Silverman, B. W. (2002) *Applied Functional Data Analysis - Methods and Case Studies* -. New York: Springer-Verlag.

Schweizer, B. (1985) “Distributions are the numbers of the future”, *Proceedings of the Mathematics of Fuzzy Systems Meeting*, pp.137-149.