行政院國家科學委員會專題研究計畫 成果報告

函數型資料時間轉換函數模型使用對共同型態函數估計的影響
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2118-M-004-003-
執 行 期 間 ：98 年 08 月 01 日至 99 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：黃子銘

計畫參與人員：博士班研究生-兼任助理人員：鄭宇翔
博士班研究生-兼任助理人員：林永忠

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 99 年 07 月 29 日

1 Introduction, literature review and objectives

Functional data analysis is concerned of the problem of analyzing a set of random curves. Those curves are often functions of time, observed frequently at different time points. In many applications, the observed curves exhibit the same pattern, but vary in amplitude and in time progression. For instance, for speech signal curves, different individuals may speak the same words at different speeds and levels of loudness. More examples of applications can be found in [8] and [9].

For the estimation of the common shape of the curves, it is known that a usual estimator such as the cross-sectional mean is improper ([2]; [4]). To deal with the time variation nature of individual curves, different curve registration (warping) methods have been developed with the aim to align individual curves to a given template.

One well-known registration method is landmark registration, which involves selecting certain features (landmarks) and align the curves by identifying the timing of selected landmarks. Kneip and Gasser [4] described landmark registration in a statistical setting using structral functionals. As mentioned by various authors ([7]; [6]), to use the landmark registration approach, curves need to exhibit common features. If some landmarks are missing for certain curves, there can be a problem.

There are other approaches for curve registration. In shape invariant modeling (SIM), individual curves are modeled using shift and scaling transforms for the common shape function and for time transforms. This approach was first proposed by Lawton, Sylvestre and Maggio [5] and later considered by Kneip and Gasser [3]. Silverman [13] proposed a functional PCA model allowing for individual time-shift, where the shift effect was considered random. Ramsay and Li [7] proposed to estimate the time-warping function by minimizing a penalized squared error criterion, with a penalty term proportional to the relative curvature of the warping function. Wang and Gasser [15] proposed to align one curve to a reference curve using dynamic time warping, a technique that had been developed in the engineering literature (see [12] for example), but with a new cost function involving penalty for the roughness of the warping path and for mis-alignment. Rønn [10] considered a nonparametric maximum likelihood approach for a shape invariant model allowing for individual random time-shift.

Recently, Telesca and Inoue [14] proposed an interesting Bayesian hierarchical models for curve registration. To describe their model, let $[0, T]$ be the time interval where the individual curves are defined. Let N be the total number of curves. For $i = 1, \ldots, N$, let $Y_i(t)$ denote the value of the i-th individual curve at time t for $t \in [0, T]$ and let μ_i denote the warping function for the i-th curve, which is assumed to be increasing. Let f be the

common shape function. Then their model is as follows:

$$Y_i(t) = c_i + a_i f(\mu_i(t)) + \epsilon_i(t),\ t \in [0, T] \text{ and } i = 1, \ldots, N, \tag{1}$$

where the c_i and a_i are parameters for individual shift and scaling effects, and the $\epsilon_i(t)$'s are independent errors. The common shape function f and the warping functions μ_i's are modeled using B-splines. Priors are put on the spline coefficients as well as other parameters in the model, and the MCMC algorithm can be applied to sample from the posterior of the warping functions and the common shape function.

The model in (1) looks appealing since it offers great flexibility. However, identifiability is certainly a issue. First, the shift and scaling effects c_i's and a_i's cannot be separated from $f(\mu_i(t))$ since $c_i + c + a \cdot a_i \cdot g(\mu_i(t)) = c_i + a_i f(\mu_i(t))$ when $c + ag(t) = f(t)$. Such a non-identifiability problem may be solved by making an overall restriction like $\sum_{i=1}^N c_i = 0$ and $\sum_{i=1}^N a_i = N$, or by choosing one curve as the reference and making a restriction like $c_1 = 0$ and $a_1 = 1$. However, even in the case where $c_i = 0$ and $a_i = 1$ for all i, it is still possible to have $f(\mu_i(t)) = f(\tilde{\mu}_i(t))$ for different warping functions $\mu_i(t)$ and $\tilde{\mu}_i(t)$, unless f satisfies some shape constraints such as being unimodal or monotone. Therefore, a different type of non-identifiability may still exist.

Since the purpose for estimating the warping functions is to obtain a reasonable estimator for the common shape function, the following question is of interest.

- Question 1. Can the estimation of the common shape function be improved by using a more flexible family to model the warping functions than a simple family such as the location family?

Finding an aswer for the above question is the main objective of this project, and it is inevitable to deal with the identifiability issue.

2 Approach

To address Question 1, simulation studies will be carried out. For simplicity, the following simplified model is considered, which is obtained by setting $c_i = 0$ and $a_i = 1$ in Model (1):

$$Y_i(t) = f(\mu_i(t)) + \epsilon_i(t),\ t \in [0, T] \text{ and } i = 1, \ldots, N. \tag{2}$$

Here μ_i's are assumed to be IID random increasing functions such that $E\mu_i(t) = t$ for $t \in [0, T]$. Data curves will be generated according to (2) with f being a smooth function and the μ_i's are generated using B-splines with random coefficients, where $t \in \{t_1, \ldots, t_n\} \subset [0, T]$. f and μ_i's will

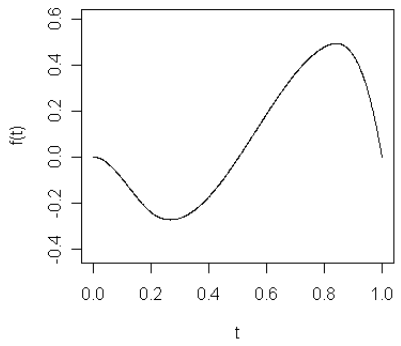
be estimated via maximum likelihood estimation assuming that the $\epsilon_i(t)$'s are normal. Both f and the μ_i's are approximated using B-splines. The estimate for f can be then obtained when μ_i's are modeled using (i) and (ii) respectively, and then comparisons can be made to see if the estimation of f is improved when (ii) is replaced by (i). The experimental details are given in Section 3.

3 Experimental details and results

In this experiment, two curves $Z_{1,t} = f(t) + \varepsilon_{1,t}$: $t = 1, \ldots, n$, and $Z_{2,t} = f(\mu(t)) + \varepsilon_{2,t}$: $t = 1, \ldots, n$ are generated, where the common shape function f is given by

$$f(x) = -0.7B_3(x) + 0.6B_4(x) + 0.5B_5(x),$$

where $B_1, \ldots, B_5$ are B-spline basis of order 4 (degree 3) with internal knots 0.2, 0.8 and boundary knots 0, 1, arranged by the lower/upper bounds of their supports. The graph of the common shape function f is shown below.

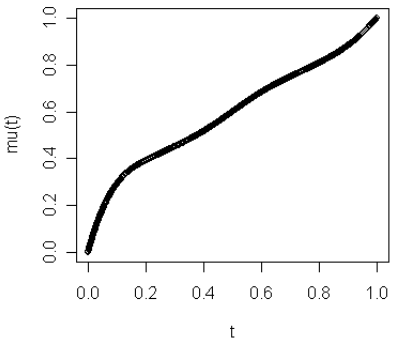


The warping function for the second curve is

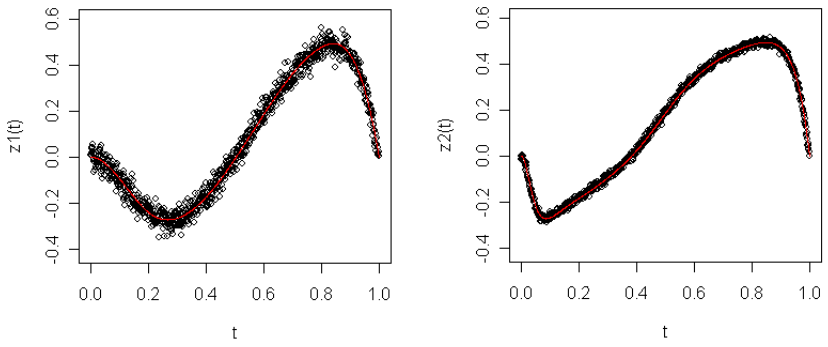
$$\begin{aligned}\mu(t) &= (1/15)B_{2,2}(t) + 0.2B_{2,3}(t) \\ &\quad +0.4B_{2,4}(t) + 0.6B_{2,5}(t) + 0.8B_{2,6}(t) + (14/15)B_{2,7}(t) + B_{2,8}(t),\end{aligned}$$

where $B_{2,1}, \ldots, B_{2,8}$ are B-spline basis of order 4 (degree 3) with internal knots 0.2, 0.4, 0.6, 0.8 and boundary knots 0, 1, arranged by the lower/upper bounds of their supports. The graph of the warping function μ is shown

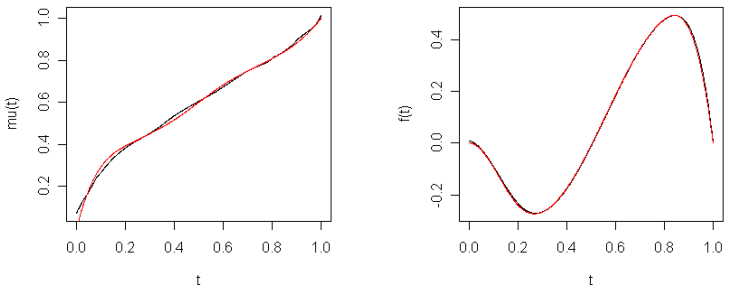
below.



The errors $\varepsilon_{1,t}$'s and $\varepsilon_{2,t}$'s are independent and normally distributed with mean zero and standard deviations 0.03 and 0.01 respectively. The graphs of the two generated curves are shown below.



To estimate the warping function for the second curve, the first curve is used as a reference curve. Then the two curves are aligned by first approximating the curves by spine functions and then matching their local extremes. For spine approximation, B-spline functions of order 4 (degree 3) are used and the algorithm proposed by Zhou and Shen (2001) is used to determine the knot locations. Using the spline approximations with the first curve as the reference, the warping time for the second curve can be estimated, and then the estimated warping function is approximated by a spline function again. Then the common shape function is estimated again treating the estimated warping function as the true warping function. Below are the graphs for the estimated common shape function and the estimated

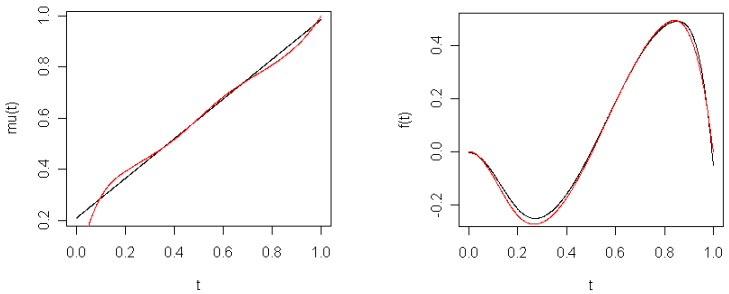


(a) Estimated $\mu(t)$ (b) Estimated $f(t)$

Figure 1: Estimated warping function and common shape function

warping function, where the red curves are the true common shape function $f(t)$ and the true warping function $\mu(t)$.

To see the estimation of the common shape function will be affected when using a simpler model for estimating the warping function, the above estimation procedure is repeated with the warping function replaced by a linear function. Below are the graphs for the estimated common shape function and the estimated warping function, where the red curves are the true common shape function $f(t)$ and the true warping function $\mu(t)$.



(a) Estimated $\mu(t)$ (b) Estimated $f(t)$

Figure 2: Estimated linear warping function and common shape function

4 Discussion

From the result of the previous experiment, it is found that using a more flexible family for the warping functions helps improve the estimation of the common shape function. An additional interest finding is that, it seems that some parameters involved may be nearly unidentifiable (different parameters yields the same common shape function, the estimate of the common shape function is stable. For future study, one might want to consider using stability of quantities that can be determined by the common shape function as an index for convergence.

References

[1] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[2] Theo Gasser, Hans-Georg Müller, Walter Köhler, Luciano Molinari, and Andrea Prader. Nonparametric regression analysis of growth curves (Corr: V12 p1588). *The Annals of Statistics*, 12:210–229, 1984.

[3] Alois Kneip and Theo Gasser. Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics*, 16:82–112, 1988.

[4] Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20:1266–1305, 1992.

[5] W. H. Lawton, E. A. Sylvestre, and M. S. Maggio. Self modeling nonlinear regression. *Technometrics*, 14:513–532, 1972.

[6] Xueli Liu and Hans-Georg Müller. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699, 2004.

[7] J. O. Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 60:351–363, 1998.

[8] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer-Verlag Inc, 1997.

[9] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag Inc, 2002.

[10] Birgitte B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(2):243–259, 2001.

[11] David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, Cambridge, U.K., 2003.

[12] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[13] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Series B: Methodological*, 57:673–689, 1995.

[14] Donatello Telesca and Lurdes Y. T. Inoue. Bayesian Hierarchical Curve Registration. *Journal of the American Statistical Association*, 103(481):328–339, 2008.

[15] Kongming Wang and Theo Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276, 1997.

[16] Shanggang Zhou and Xiaotong Shen. Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, 96(453):247–259, 2001.

無研發成果推廣資料

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 □撰寫中 ■無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

衍生研究成果: 本人進行一未對齊曲線資料分析, 用到本計畫的發現. 初步結果於 2010 複雜數據統計分析國際會議中報告.

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

主要發現為使用足夠複雜的時間轉換函數模型有助於估計共同形狀函數. 因此在未來研究中, 會建議使用較複雜的模型如 spline 模型, 而非簡單的線性模型來處理時間轉換函數. 這一發現對本人在曲線對齊方面的研究工作極有幫助, 例如本人和 Dr. Su-Fen Yang 一起分析了 一些嬰兒追蹤器數據, 其中牽涉了對齊曲線的問題. 因為有了本計畫的發現就使用了 spline 模型處理時間轉換函數, 效果不錯. 初步結果於 2010 年 7/1-3 於昆明大學舉辦的 SACD conference（複雜數據統計分析國際會議）中報告, 標題為 distinguishing an unusual curve from unaligned curves with a common shape with application to the babyfinder data.