Journal of the Chinese Statistical Association
Vol. 51, (2013) 327–351

應用切片逆迴歸法於區間型符號資料之維度縮減

陳業勛　　吳漢銘
淡江大學數學學系

摘　要

運用切片逆迴歸法 (sliced inverse regression, SIR) 可以找出有效的維度縮減方向來探索高維度資料的內在結構。針對單一反應變數逆迴歸問題, SIR 已發展並應用在各種資料型態上, 例如: 存活資料、時間序列資料、函數型資料及縱向資料等等。本研究中, 我們推展 SIR 方法到區間型符號資料。首先利用頂點法或中心法將區間資料做轉換, 再應用 SIR 於轉換後的資料上。模擬資料分析結果顯示, 不同的切片策略會產生不同的維度縮減方向及呈現不同的低維度視覺化結果, 因此找出合適的切片策略有助於正確地分析這類型高維度資料所隱含的結構與資訊。故我們進一步採用以群集爲基礎的切片逆迴歸法來分析區間型符號資料, 並和符號型主成份分析法相比較, 評估它們在低維度空間中區別能力及視覺化的表現。

關鍵詞: 資料視覺化, 逆迴歸法, 充份維度縮減法, 符號型資料分析法, 符號型主成份分析法。
JEL classification: C14, C63

1. 緒論

當資料所包含的變數越多, 代表此資料維度越大, 易造成統計分析上的困難, 而產生「維度的詛咒」(curse of dimensionality) 的問題。文獻上有一些維度縮減的統計分析方法, 目的是降低資料維度使其能在低維度空間中被視覺化, 藉以觀察此高維度資料所隱含的結構及資訊。

主成份分析 (principal component analysis, PCA) 就是最常見的維度縮減方法之一。傳統上, 維度縮減的研究是針對在每個維度 (或變數) 上以單一觀察值所組成的資料集合 (classical datasets) 爲分析對象。然而在現實生活中, 資料收集愈趨於巨量, 也愈爲複雜。爲了匯總管理並同時儘可能的保留資料原本所給予的資訊, 資料收集的變數格式不再只是單一點的數值, 而可能是以數列 (sequence) 、名單 (list) 或區間 (interval) 代表, 此爲文獻上所稱的符號型資料 (symbolic data)。例如: 天氣溫度每天的最高溫和最低溫、金融交易的最高跟最低交易價格等等的資料, 都是紀錄觀察值的最高和最低的資訊, 而形成一區間資料。而一般傳統資料也可轉換成符號型資料進行分析, 但相對地也會造成部份訊息損失。

符號型資料分析 (symbolic data analysis, SDA) (Billard & Diday, 2003) 的對象可分爲多元值 (multi-valued) 、給定機率的多元值 (modal multi-valued) 及區間值 (interval-valued) 等三種基本資料類型; 文獻上, 主成份分析在符號資料上的應用, 絕大部份限於區間值的資料, 針對其他符號資料類型的分析尚未有有效的推廣或方法發表。而本文研究也是以區間型資料爲對象。

以最常見的維度縮減方法–主成份分析法–來說, 早期應用在區間型符號資料上的方法有 Ichino (1988) 所提出的笛卡爾空間理論; Ichino & Yaguchi (1994) 用廣義 Minkowsky metrics 推廣主成份分析於混合特徵資料 (mixed feature data) 以及 Nagabhushan, Gowda & Diday (1995) 的泰勒級數想法; 目前較常見的方法有頂點式主成份分析 (vertices method PCA, V-PCA) 和中心式主成份分析 (center method PCA, C-PCA) (Cazes *et al.*, 1997; Douzal-Chouakria, Diday & Cazes, 1998; Gioia & Lauro, 2006) 兩種方法。前者可表示在每個變數中觀察值的區間長度變化, 而後者可表示觀測區間的重心變化。此兩方法都是先把符號型資料轉換成一般傳統資料集型式, 使得傳統主成份分析法可以直接應用。一般而言, 低維度空間的視覺化會以前兩個主成份來呈現資料的降維結果, 每個觀察值以方框方式呈現, 從方框的大小及群聚現象, 可觀察出資料在高維空間中的結構。主成份分析在區間型資料上發展迅速, 也有其他改進和新的方法陸續發表, 例如 Lauro & Palumbo (2000) 和 Palumbo & Lauro (2003) 陸續提出 S-PCA (symbolic PCA) 、RT-PCA (range transformation PCA)、MR-PCA (midpoint radius PCA) 等等這些對區間型資料做不同轉換的方法; Zuccolotto (2011) 則針對具遺失值資料的區間值, 提出相關係數矩陣的處理方法, 使得主成份分析得以應用。

Li (1991) 提出切片逆迴歸法 (SIR), 其目的在找出有效的維度縮減方向來探索高維度資料的內在結構。藉由將 p 維的解釋變數 $\mathbf{x}$ 投影在有效維度縮減 (effective dimension reduction, EDR) 方向上, 可使維度縮減且大量保有原本 $\mathbf{x}$ 對於 y 的迴歸訊息。SIR 已推廣及應用於其他類型資料上, 例如影像資料 (Wu & Lu, 2004; 2007), 存活資料 (Li, Wang & Chen, 1999;

Li & Li, 2004), 時間序列資料 (Becker & Fried, 2003), 函數型資料 (Ferre & Yao, 2003) 和縱向資料 (Li & Yin, 2009) 等等。本研究中, 我們採用以集群爲基礎的切片逆迴歸法 (Kuentz & Saracco, 2010) 來分析區間型資料, 原因是以頂點法或中心法將區間型資料轉換成一般資料矩陣後, 資料結構會以多群體組合在一起的方式呈現, 因而概念上使用以群集爲基礎的 SIR會較傳統的 SIR 較適合。文獻上, 目前尚未有切片逆迴歸法應用於區間型資料上, 我們嘗試比較幾種不同切片方式, 找出較適合的切片逆迴歸法於區間型資料上的應用。

本文組織如下: 第2節介紹主成份分析在區間型符號資料的推展, 包含頂點式主成份分析法和中心式主成份分析法。第3節介紹傳統的切片逆迴歸法、以群集爲基礎的切片逆迴歸法, 以及兩方法在區間型資料上的推展。第4節以模擬研究來比較不同切片策略; 第5 節模擬區間型符號之資料並比較三種維度縮減方向的準確性。第6 節則進行實際區間型資料分析。研究的結論及討論則列於第7節。

2. 區間型符號資料之維度縮減

2.1 頂點式主成份分析法 (V-PCA)

在$\Re^p$ 空間中, 傳統資料是以數值點的方式表現, 而區間型資料會以超立體 (hypercube) 的形式呈現, 所以 V-PCA 的做法是把區間型資料建構成一個包含所有超立體頂點座標的資料, 將之視爲一般傳統資料, 使得 PCA 得以直接應用; 亦即找出所有超立體變異數最大的維度縮減方向, 最後再依照超立體投影至此維度縮減方向上的區域, 重新建構其區間值。假設區間型資料記爲 $\boldsymbol{\xi}_u$,

$$\boldsymbol{\xi}_u = (\xi_{u1}, \cdots, \xi_{up}) = ([a_{u1}, b_{u1}], \cdots, [a_{up}, b_{up}]),$$

其中 ξ_{uj} 代表第u筆受試者 (subject) 的第j個變數的觀測區間$[a_{uj}, b_{uj}]$, ($u = 1, \cdots, m$, $j = 1, \cdots, p$), a_{uj}爲第j個區間變數的最小值,b_{uj}爲其最大值。若區間型資料是由某一傳統資料, 依群集或某一概念整合 (aggregate) 而來, 則$\boldsymbol{\xi}_u$可代表此一傳統資料的第u筆群集或概念。令 q_u爲非平凡 (nontrivial) 區間 (即滿足 $a_{uj} < b_{uj}$的區間) 個數。從 $\boldsymbol{\xi}_u$可建構出共 $m_u \equiv 2^{q_u}$ 個頂點座標所構成的超立體 H_u。若以 M_u 矩陣來表示此超立體 H_u, 則其維度爲 $2^{q_u} \times p$, 矩陣中每一列代表爲每一觀察值的頂點座標值 V_{k_u}, $k_u = 1, \cdots, 2^{q_u}$。以下爲頂點式主成份分析法之演算法。

V-PCA Algorithm:

i. 區間型資料之格式轉換: 從每個建構成超立體的 H_u 組合成矩陣 M, 維度爲 $N \times p$, $N = \sum_{u=1}^{m} 2^{q_u}$:

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix}, \quad \text{其中 } M_u = \begin{bmatrix} a_{u1} & \cdots & a_{uq_u} \\ a_{u1} & \cdots & b_{uq_u} \\ \vdots & \ddots & \vdots \\ b_{u1} & \cdots & a_{uq_u} \\ b_{u1} & \cdots & b_{uq_u} \end{bmatrix}, \quad u = 1, \cdots, m$$

ii. 特徵值分解: 將M矩陣中每一變數做標準化使得平均爲0, 標準差爲1。標準化後的矩陣記爲 $\mathbf{X}$, 將 $\mathbf{X}$ 的共變異數矩陣 $\Sigma_{\mathbf{x}}$ 做特徵值分解, 得到特徵值 λ_ν 和相對應的特徵向量 $\mathbf{e}_\nu$, $\nu = 1, \cdots, s$, $s \leq p$, 其中 s 個特徵向量排列依據爲 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s \geq 0$。

iii. 將$\mathbf{X}$ 與 $\mathbf{e}_\nu$ 做線性組合, 得到主成份PC_ν, 即降維後的資料:

$$PC_\nu = \mathbf{e}_\nu' \mathbf{X}, \quad \nu = 1, \cdots, s.$$

iv. 重建降維區間: 令 L_u 是用以指出矩陣M中超立體 H_u 內 2^{q_u} 個頂點座標的列指標集合。依上一步驟, 對於 L_u 裡的每一列 $k_u, k_u = 1, \cdots, 2^{q_u}$, 可得 到降維後的值, 記做$z_{\nu u k_u}$; 亦即 M_u 的第ν個主成份PC_ν之值。則第ν個區間型主成份, $Z_{\nu u}^V$, 可重建如下:

$$Z_{\nu u}^V = z_{\nu u} = [z_{\nu u}^a, z_{\nu u}^b], \quad \nu = 1, \cdots, s, \ u = 1, \cdots, m$$

其中

$$z_{\nu u}^a = \min_{k_u \in L_u} \{z_{\nu u k_u}\} \quad \text{和} \quad z_{\nu u}^b = \max_{k_u \in L_u} \{z_{\nu u k_u}\}.$$

v. 低維度視覺化: 若取 $s = 2$, 則於二維圖中畫出 Z_{1u}^V 和 Z_{2u}^V, $u = 1, \cdots, m$, 所形成之方框散佈圖。

2.2 中心式主成份分析法 (C-PCA)

中心式主成份分析法方法由 Cazes *et al.* (1997) 和 Douzal-Chouakria *et al.* (1998) 所提出。此方法是將區間型資料之每一區間變數以中心點代替, 可表示觀測區間的重心變化 (Douzal-Chouakria, Billard & Diday, 2011)。以下爲中心式主成份分析法之步驟:

C-PCA Algorithm:

i. 區間型資料之格式轉換: 將區間型資料 $\boldsymbol{\xi}_u = (\ [a_{u1}, b_{u1}], \cdots, [a_{up}, b_{up}]\ )$ 轉換爲

$$\mathbf{X}_u^c = (X_{u1}^c, \cdots, X_{up}^c),\ u = 1, \cdots, m,$$

其中

$$X_{uj}^c = \frac{a_{uj} + b_{uj}}{2},\ j = 1, \cdots, p.$$

最後藉由 $\mathbf{X}_u^c$ 向量組合成 $\mathbf{X}^c$ 矩陣 (亦即 $\mathbf{X}^c$ 的第 u 列元素爲 $\mathbf{X}_u^c$), 其維度爲 $m \times p$。

ii. 特徵值分解: 應用傳統的 PCA 於$\mathbf{X}^c$矩陣上; 亦即對 $\mathbf{X}^c$ 之變異數矩陣 $\Sigma_{\mathbf{x}}^c$ 做特徵值分解,

$$\lambda_\nu \mathbf{w}_\nu = \Sigma_{\mathbf{x}}^c \mathbf{w}_\nu,$$

其中, 特徵向量 $\mathbf{w}_\nu = (w_{\nu 1}, \cdots, w_{\nu p})$ 之順序是根據特徵值 λ_ν 由大至小排序, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s$, $s \leq p$。則區間觀察值$\boldsymbol{\xi}_u$的第ν個主成份之值爲

$$z_{\nu u}^c = \sum_{j=1}^{p} (X_{uj}^c - \bar{X}_j^c) w_{\nu j},\ \nu = 1, \cdots, s,\ u = 1, \cdots, m$$

其中$\bar{X}_j^c$爲變數X_{uj}^c之平均。

iii. 重建降維區間: 令 $Z_{\nu u}^C = [z_{\nu u}^{ca}, z_{\nu u}^{cb}]$爲降維後之區間, $\nu = 1, \cdots, s,\ s \leq p$, 其中

$$z_{\nu u}^{ca} = \sum_{j=1}^{p} \min_{a_{uj} \leqslant X_{uj}^c \leqslant b_{uj}} (X_{uj}^c - \bar{X}_j^c) w_{\nu j}\ ,$$

$$z_{\nu u}^{cb} = \sum_{j=1}^{p} \max_{a_{uj} \leqslant X_{uj}^c \leqslant b_{uj}} (X_{uj}^c - \bar{X}_j^c) w_{\nu j}.$$

iv. 低維度視覺化: 若取 $s = 2$, 則於二維圖中畫出 Z_{1u}^C 和 Z_{2u}^C, $u = 1, \cdots, m$, 所形成之方框散佈圖。

2.3 相關係數與貢獻度

相關係數與貢獻度是用來衡量區間型符號資料維度縮減後效益的兩個指標。以 V-PCA 爲例, 考慮矩陣 M 中的每一區間變數 X_j 和 PC_ν 之相關性, 公式如下:

$$C_{\nu j} = \mathrm{Cor}(PC_\nu, X_j) = e_{\nu j} \sqrt{\lambda_\nu / \hat{\sigma}_j^2}\ ,\ \nu = 1, \cdots, s,\ j = 1, \cdots, p,$$

其中 $e_{\nu j}$ 是第ν個主成份的特徵向量 $\boldsymbol{e}_\nu$ 中的第 j 個元素, λ_ν 爲特徵值, $\hat{\sigma}_j^2$ 爲 X_j 的變異數 (式子 (1))。藉由此公式判定原來區間變數與 PC_ν 之間的相關程度。若相關係數絕對值愈越接近 1, 則相關性越高; 越接近0代表相關性越小。

貢獻度的定義爲經由歐式距離計算出每一頂點k_u在第 ν 個有效維度縮減方向上所佔的分量 (Billard & Diday, 2006)。其計算步驟及公式如下:

i. 計算矩陣 M 中的每一行變數之平均數及標準差:

$$\bar{X}_j = \frac{1}{N}\sum_{u=1}^{m}\sum_{k_u=1}^{m_u} X_{k_u j}, \quad S_j^2 = \frac{1}{N-1}\sum_{u=1}^{m}\sum_{k_u=1}^{m_u}(X_{k_u j} - \bar{X}_j)^2, \tag{1}$$

其中$X_{k_u j}$ 代表超立體H_u中的頂點k_u在第j個變數之值。

ii. 計算頂點座標與中心點間的歐式距離平方:

$$d^2(k_u, \bar{X}_j) = \sum_{j=1}^{p}\left(\frac{X_{k_u j} - \bar{X}_j}{S_j}\right)^2.$$

iii. 計算頂點貢獻度 $Con(PC\nu, k_u)$:

$$Con(PC\nu, k_u) = \frac{z_{\nu u k_u}^2}{d^2(k_u, \bar{X}_j)}, \quad k_u = 1, \cdots, 2^{q_u}, \quad u = 1, \cdots, m,$$

其中 $z_{\nu u k_u}$ 代表區間觀察值 $\boldsymbol{\xi}_u$所構成的超立體H_k裡, 第 k_u 個頂點投影在第 ν 個維度縮減的方向上的值。

iv. 應用頂點貢獻度:

利用此貢獻度將 V-PCA 方法中的重建降維區間 $Z_{\nu u}^V = [z_{\nu u}^a, z_{\nu u}^b]$, 細分爲

$$Z_{\nu u}^V(\alpha) = [z_{\nu u}^a(\alpha), z_{\nu u}^b(\alpha)], \ 0 \le \alpha \le 1$$

其中

$$z_{\nu u}^a(\alpha) = \min_{k_u \in L_u}\{z_{\nu u k_u} | Con(PC\nu, k_u) \ge \alpha\},$$
$$z_{\nu u}^b(\alpha) = \max_{k_u \in L_u}\{z_{\nu u k_u} | Con(PC\nu, k_u) \ge \alpha\}.$$

給定α值之下, 藉此貢獻度的計算, 可刪除貢獻度較小的頂點, 使投影至有效維度縮減平面上所構成的範圍縮小, 讓資料視覺化後更容易辨識資料的樣態及群性。同樣的概念, 可以推廣至 C-PCA, 以及接下來要介紹的 V-SIR, C-SIR及 V-cbSIR 等方法上, 以求得降維後, 每個頂點的貢獻度。

3. 區間型符號資料之切片逆回歸法

3.1 切片逆迴歸法

Li (1991) 提出一個維度縮減的迴歸模型爲

$$y = f(\boldsymbol{\beta}_1'\mathbf{x}, \cdots, \boldsymbol{\beta}_K'\mathbf{x}, \epsilon),$$

其中 f 爲一隨意函數, y 是一個單變量的隨機變數, $\mathbf{x}$ 是維度爲 $p\times 1$ 的隨機向量, 其中 $K \leq p$, ϵ 是一個隨機誤差且和 $\mathbf{x}$ 獨立, $\boldsymbol{\beta}_i$ 是有效維度縮減方向, 其維度爲 $p \times 1$。藉由投影到有效維度縮減方向上, 可使維度縮減且大量保有原本 $\mathbf{x}$ 對於 y 的相關迴歸訊息; 而切片逆迴歸法即是從 $\mathbf{x}$ 和 y 變量中估計有效維度縮減方向的方法。假設所觀察到的資料表示爲 $\{y_i, \mathbf{x}_i\}_{i=1}^n$, SIR 的演算法如下。

SIR Algorithm:

i. 計算樣本平均數 $\bar{\mathbf{x}}$ 及樣本變異數矩陣 $\hat{\Sigma}_{\mathbf{x}}$。

ii. 對反應變數 y 之數值排序並切片: 將 $y_i's$ 由小到大排序後, 將之切割成 H 個切片 $I_h, h = 1, \cdots, H$。

iii. 計算每個切片中的平均數 $\bar{\mathbf{x}}^{(h)}$及加權共變異矩陣 $\hat{\Sigma}_{\mathbf{w}}$如下:

$$\bar{\mathbf{x}}^{(h)} = \frac{1}{n_h}\sum_{i\in I_h}\mathbf{x}_i, \quad \hat{\Sigma}_{\mathbf{w}} = \sum_{h=1}^{H}\frac{n_h}{n}(\bar{\mathbf{x}}^{(h)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(h)} - \bar{\mathbf{x}})',$$

其中 n_h 代表爲第 h 個切片內之觀察值個數。

iv. 對 $\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Sigma}_{\mathbf{w}}$ 做特徵值分解, 亦即對 $\hat{\Sigma}_{\mathbf{w}}$做廣義特徵值分解:

$$\hat{\Sigma}_{\mathbf{w}}\hat{\boldsymbol{\beta}}_\nu = \hat{\lambda}_\nu\hat{\Sigma}_{\mathbf{x}}\hat{\boldsymbol{\beta}}_\nu, \quad \nu = 1, \cdots, s,\ s \leq p$$

其中特徵值 $\hat{\lambda}_\nu$ 滿足 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_s$, 而 $\hat{\boldsymbol{\beta}}_\nu$ 即爲有效的維度縮減方向之估計。

v. 將資料投影到有效的維度縮減方向:

$$\text{SIR}_\nu = \hat{\boldsymbol{\beta}}_\nu'\mathbf{x}, \quad \nu = 1, \cdots, K.$$

vi. 低維度視覺化: 畫出 y 與 $\mathrm{SIR}_1(\mathrm{SIR}_2)$ 之二維 (三維) 散布圖或 SIR_1 與 SIR_2 之二維散佈圖, 藉以觀察維度縮減後資料之分佈。

3.2 以群集爲基礎的切片逆迴歸法 (cluster-based SIR)

以群集爲基礎的切片逆迴歸法 (cbSIR) 由 Kuentz & Saracco, (2010) 提出, 目的是先將 $\mathbf{x}$ 分爲 C 個群集, 應用 SIR 方法在每個群集中, 並整合其估計結果來尋找最後有效維度縮減空間, 藉以提高 SIR 對於維度縮減的估計準確性。在本研究中, 假設以下的線性條件皆成立,

> 對任意 $\mathbf{b} \in \Re^p$而言, 條件期望值$E[\mathbf{b}'\mathbf{x}^{(j)}|\boldsymbol{\beta}_1'\mathbf{x}^{(j)}, \cdots, \boldsymbol{\beta}_K'\mathbf{x}^{(j)}]$ 可以表示爲 $\boldsymbol{\beta}_1'\mathbf{x}^{(j)}$, $\cdots, \boldsymbol{\beta}_K'\mathbf{x}^{(j)}$的一線性組合, 其中 $j = 1, \cdots, C$。

我們敘述以群集爲基礎的切片逆迴歸法演算法如下, 方法的詳細理論說明請見 Kuentz & Saracco, (2010)。

cbSIR Algorithm:

i. 以群集分析 (例如 K 均值法 (K-means), 記作 KM) 將資料分成 C 群, $\mathbf{x} = \{\mathbf{x}^{(c)}, c = 1, \cdots, C\}$。此爲第一次的切片, 其切片數記爲$H_{(1)} = C$。

ii. 計算出每個群集內之前 K 個有效維度縮減方向: 應用 SIR 在每個群集所形成的子集合資料 $\mathbf{x}^{(c)}$, 計算各自的有效維度縮減方向, 其步驟同切片逆迴歸法。在此每一子集都採用一樣的切片數, 記爲 $H_{(2)}$。

iii. 定義矩陣 $B^{(c)} = [\mathbf{b}_1^{(c)}, \cdots, \mathbf{b}_K^{(c)}]$, 其中 $\mathbf{b}_1^{(c)}, \cdots, \mathbf{b}_K^{(c)}$ 爲第 c 群集資料之前 K 個有效維度縮減方向。將矩陣 $B^{(c)}B^{(c)\prime}$ 做特徵值分解並取前 K 個特徵向量 $\tilde{\mathbf{b}}_1^{(c)}, \cdots, \tilde{\mathbf{b}}_K^{(c)}$ 形成矩陣 $\tilde{B}^{(c)} = [\tilde{\mathbf{b}}_1^{(c)}, \cdots, \tilde{\mathbf{b}}_K^{(c)}]$, $c = 1, \cdots, C$。

iv. 合併所有 $\tilde{B}^{(c)}$ 矩陣得到 $\mathbb{B} = [\tilde{B}^{(1)}, \cdots, \tilde{B}^{(c)}]$。將矩陣 $\mathbb{B}\mathbb{B}'$ 做特徵值分解並取前 K 個特徵向量 $\tilde{\tilde{\mathbf{b}}}_1, \cdots, \tilde{\tilde{\mathbf{b}}}_K$。這些向量就是以群集爲基礎的切片逆迴歸法的有效維度縮減方向的估計。

v. 低維度視覺化: 畫出 y 與 $\mathrm{cbSIR}_1(\mathrm{cbSIR}_2)$ 之二維 (三維) 散布圖或 cbSIR_1 與 cbSIR_2 之二維散佈圖, 藉以觀察維度縮減後資料之分佈。

3.3 頂點式切片逆迴歸法 (V-SIR)

參照 V-PCA 的方法, V-SIR 方法的實現, 是以頂點法將區間型資料轉換成包含所有頂點座標的超立體資料, 再運用 SIR 的方法算出有效維度縮減方向; 將超立體矩陣與有效維度縮減方向做線性組合, 即可算出每組區間型資料 $\boldsymbol{\xi}_u$ 所構成的超立體在有效維度縮減方向上所投影的範圍並使其在低維度空間可被視覺化。V-SIR之演算法如下。

V-SIR Algorithm:

i. 區間型資料之格式轉換: 使區間型資料 $\boldsymbol{\xi}_u$ 建構成矩陣 M 並標準化, 方法同頂點式主成份分析法的步驟 i。

ii. 對 M 矩陣藉由 SIR 方法估計其有效的維度縮減方向, 其步驟同切片逆迴歸法的步驟 i 至 v。其中切片方式是以每一M_u爲一個切片單位 $(u = 1, \cdots, m)$。

iii. 將資料投影到有效的維度縮減方向: M矩陣與有效維度縮減方向做線性組合。

iv. 重建降維區間: 同 V-PCA 步驟 iv。

v. 低維度視覺化: 若取 $s = 2$, 則於二維圖中畫出 SIR_{1u}^{V} 和 SIR_{2u}^{V}, $u = 1, \cdots, m$, 所形成之方框散佈圖。

3.4 中心式切片逆迴歸法 (C-SIR)

C-SIR 是 SIR 演算法應用在以中心式方法來做區間型資料轉換成資料矩陣。其演算法如下所示。

C-SIR Algorithm:

i. 將區間型資料之格式轉換成 $\mathbf{X}^c$ 矩陣, 方法同 C-PCA 之步驟 i。

ii. 以 SIR 演算法應用在 $\mathbf{X}^c$ 資料矩陣上, 其步驟同切片逆迴歸法的步驟 i 至 v。其中切片方法可藉由分割排序後之反應變數 y, 或由對 $\mathbf{X}^c$ 做群集分析而得到。

iii. 重建降維區間及低維度視覺化, 同 C-PCA 之步驟 iii 及 iv。

表 1　本研究中不同維度縮減方法所採用的切片策略。

資料類型	傳統資料	區間型資料		
維度縮減法	SIR/cbSIR	V-SIR	C-SIR	V-cbSIR
切片策略	y已知: NE(y) 或 KM(y) y未知: KM($\mathbf{x}$)	M_u $u = 1, \cdots, m$	KM($\mathbf{X}^c$)	$H_{(1)}$: $M_u, u = 1, \cdots, m$ $H_{(2)}$: KM(M_u)

3.5 以群集爲基礎的切片逆迴歸法於區間型符號資料之維度縮減

以群集爲基礎的切片逆迴歸法應用在以頂點式或中心式轉換後的區間型資料, 即可達成區間型資料之維度縮減目的。相較於頂點式法是將原資料個數增多, 中心式之轉換法則是保持總觀察個數不變。因此若資料的觀察個數不多的狀況之下, 採用中心式之轉換法來應用 cbSIR 會有計算上的問題, 因此本研究中, 僅提出以頂點式轉換法的 cbSIR 演算法 (V-cbSIR)。假設區間資料爲 $\boldsymbol{\xi}_u, u = 1 \cdots, m$。

V-cbSIR Algorithm:

i. 區間型資料之格式轉換: 以頂點式將區間型資料 $\boldsymbol{\xi}_u = (\xi_{u1}, \cdots, \xi_{up})$ 架構成超立體 H_u, 組合成矩陣 M, H_u 即爲 M 之子資料集合, $u = 1, \cdots, m$。此步驟同 V-PCA 之步驟 i。

ii. 應用 SIR 在每個區間觀察值所形成的頂點子集合資料 H_u, $u = 1, \cdots, m$, 計算各自的有效維度縮減方向並整合, 其降維方向估計及低維度視覺化的步驟同 cbSIR 的步驟 i 至 v。

V-cbSIR步驟中的第一次切片數爲$H_{(1)} = m$; 即每一M_u爲一個切片單位。第二次切片則採取群集分析的方式 (例如 K 均值法)。以切片逆迴歸法進行維度縮減的分析時, 其反應變數y的角色是做切片。對傳統資料而言, 若y已知, 可以對反應變數採用每群內之觀察個數趨近相等 (記爲 NE) 或 KM 法來做切片。若y未知, 則可對解釋變數採用 KM 法來得到切片。對區間型資料的頂點式切片逆迴歸法的切片方式, 則是每一$\boldsymbol{\xi}_u$所形成的M_u矩陣爲一切片, 所以共有m個切片。本研究中不同維度縮減方法所採用的切片策略則摘錄於表 1。

4. 以群集爲基礎的切片逆迴歸法模擬研究

不同的切片策略會導致不同的維度縮減方向估計。在本節, 我們在反應變數y已知的狀況下, 以模擬的方式比較四種不同的切片 (群集) 方法如下, 將結果當成在下一節區間型資料分析時的指引。

(1) 在傳統 SIR 的演算過程中, 針對排序後的連續變數 y 做切片, 給定切片群數爲 C, 其切片選取方式分可爲兩種: (1) 每群內之觀察個數趨近相等 (NE 法) (2) 以 K 均值法應用在 y (KM 法)。

(2) 以群集爲基礎兩種切片逆迴歸法中, 需要有兩次的切片 (或分群), 每一次可根據 NE 或 KM 的方式來做分群。第一次分群記爲 $H_{(1)}$, 第二次分群記爲 $H_{(2)}$。本小節中, $H_{(1)}$ 與 $H_{(2)}$選用一致的切片方式。

表 2　針對 Li model (6.1) 之模擬資料, 採用傳統及群集切片逆迴歸法, 在不同切片方式之下的充分維度縮減方向估計, 重覆模擬 100 次, 計算方向估計之平均及標準差 (眞實方向爲 (1,1,1,1,0), $H_{(1)}$=4, $H_{(2)}$=3)。

SIR	Slicing	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
classical	NE	0.501	0.502	0.497	0.498	-0.003
		(0.025)	(0.026)	(0.026)	(0.026)	(0.029)
	KM	0.499	0.499	0.499	0.498	-0.001
		(0.028)	(0.027)	(0.031)	(0.029)	(0.028)
cluster-based	NE	0.505	0.494	0.492	0.493	-0.001
		(0.055)	(0.051)	(0.059)	(0.058)	(0.059)
	KM	0.501	0.495	0.492	0.496	0.010
		(0.052)	(0.054)	(0.055)	(0.056)	(0.066)

4.1 Li model 6.1, 6.2 和 6.3

以下的線性模型來自 Li (1991) 的 model (6.1), 目的是藉由模擬來評估所估計的充分維度縮減方向的準確性:

$$y = x_1 + x_2 + x_3 + x_4 + 0x_5 + \epsilon, \qquad \text{(Li model 6.1)}$$

其中 x_i 與 ϵ 互相獨立且都來自於標準常態分配。我們生成 $n = 500$ 且維度爲 $p = 5$ 的資料。因爲此模型爲只有一個維度方向 ($K = 1$), 即 $\boldsymbol{\beta} = (1, 1, 1, 1, 0)$ 。我們比較 SIR、cbSIR 在

NE 及 KM 切片方式下, 有效維度縮減方向估計的準確性。取 $H_{(1)} = 4$ 和 $H_{(2)} = 3$ 並重覆實驗 100 次後取平均數和標準差, 結果列於表 2。結果顯示四種方法估計的方向皆靠近眞正的維度縮減方向 $(0.5, 0.5, 0.5, 0.5, 0)$, 而以集群爲基礎的切片逆迴歸法有較高的變異數。

表 3　針對 Li model (6.2) 及 (6.3) 之模擬資料, 採用傳統及群集切片逆迴歸法, 以不同切片方式之下的充分維度縮減方向估計, 重覆模擬 100 次, 計算典型相關係數之平均及標準差 (兩模型之眞實方向爲 (1,0,$\cdots$,0) 和 (0,1,0,$\cdots$,0)。Li model 6.2之設定爲$H_{(1)}$=5, $H_{(2)}$=3; Li model 6.3之設定爲$H_{(1)}$=4, $H_{(2)}$=3)。

		Li model (6.2)				Li model (6.3)			
SIR	Slicing	σ=0.5		σ=1		σ=0.5		σ=1	
classical	NE	0.969	0.850	0.959	0.698	0.983	0.896	0.962	0.661
		(0.020)	(0.115)	(0.025)	(0.206)	(0.010)	(0.062)	(0.023)	(0.188)
	KM	0.973	0.852	0.964	0.741	0.986	0.941	0.968	0.791
		(0.018)	(0.087)	(0.022)	(0.180)	(0.008)	(0.028)	(0.018)	(0.115)
cluster-based	NE	0.814	0.376	0.780	0.301	0.865	0.478	0.790	0.332
		(0.115)	(0.222)	(0.126)	(0.194)	(0.085)	(0.249)	(0.131)	(0.220)
	KM	0.786	0.320	0.727	0.243	0.932	0.561	0.793	0.354
		(0.126)	(0.198)	(0.159)	(0.165)	(0.048)	(0.265)	(0.149)	(0.208)

接下來我們討論眞實維度方向只有 2 ($K = 2$) 的情況。模型來自 Li (1991) 的 model (6.2) 和 (6.3) :

$$y = x_1(x_1 + x_2 + 1) + \sigma \cdot \epsilon \qquad \text{(Li model 6.2)}$$

$$y = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + \sigma \cdot \epsilon. \qquad \text{(Li model 6.3)}$$

模型內之 $x_1, \cdots, x_p$, 及 ϵ 皆互相獨立且來自於標準常態分配。我們設定資料數爲 $n = 400$, $p = 10$, 且有 $\sigma = 0.5$ 和 $\sigma = 1$ 的兩種狀況。兩模型之眞正有效維度縮減方向應在向量 $(1, 0, \cdots, 0)$ 和 $(0, 1, 0, \cdots, 0)$ 所產生的平面上。我們運用典型相關性分析法 (Canonical correlation analysis, CCA) 計算資料投影在前兩個所估計的主要維度縮減向量, 跟實際有效維度縮減向量之典型相關係數。當資料投影在兩方向上相近時, 典型相關係數值會接近 1, 我們比較 SIR、cbSIR 在 NE 及 KM 切片方式下, 有效維度縮減方向估計的準確性。對模型 (6.2) 取 $H_{(1)} = 5$ 和 $H_{(2)} = 3$, 而對模型 (6.3) 取 $H_{(1)} = 4$ 和 $H_{(2)} = 3$, 並都重複實驗 100 次, 計算典型相關係數的平均數和標準差, 列於表 3。結果顯示, 傳統的切片逆迴歸方法較好, 而群集爲基礎的維度縮減第二個方向相關性大約只有 0.3 左右, 其值偏低, 表示估計第二個維度縮減方向不夠準確。

表 4　針對 Kuentz model (8) 之模擬資料, 採用傳統及群集切片逆迴歸法, 以不同切片方式之下的充分維度縮減方向估計, 重覆模擬 100 次, 計算典型相關係數之平均及標準差 (眞實方向爲 $(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, 0)$ 和 $(0, 0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$, $H_{(1)}$=2, $H_{(2)}$=5)。

SIR	Slicing	θ=0		θ=0.5		θ=1	
classical	NE	0.999	0.973	0.999	0.993	1.000	0.995
		(0.002)	(0.033)	(0.001)	(0.011)	(0.001)	(0.004)
	KM	0.999	0.819	0.999	0.985	0.999	0.993
		(0.002)	(0.089)	(0.001)	(0.022)	(0.001)	(0.006)
cluster-based	NE	0.994	0.627	0.993	0.685	0.991	0.754
		(0.009)	(0.292)	(0.010)	(0.263)	(0.012)	(0.217)
	KM	0.993	0.857	0.998	0.912	0.998	0.896
		(0.008)	(0.144)	(0.003)	(0.092)	(0.003)	(0.097)

4.2 Kuentz model (8)

以下模型, 來自於 Kuentz & Saracco (2010) 的 equation (8)

$$y = (\mathbf{x}'\beta_1 + \varepsilon_1)\mathbb{I}_{[\mathbf{x}'\beta_2+\varepsilon_2>0]}\ , \qquad \text{(Kuentz model 8)}$$

其中 $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)'$, $x_j \sim (1-\theta) \times Exp(1) + \theta \times N(0,1)$ 且 $\varepsilon_1 \sim N(0, 0.1^2)$ 和 $\varepsilon_2 \sim N(0, 0.1^2)$獨立, θ 的範圍爲 [0, 1]之間。變數 x_j 皆獨立且殘差項 ε_1 跟 ε_2 也都與之獨立, 此模型之眞實維度方向應爲 $\boldsymbol{\beta}_1 = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, 0)'$ 和 $\boldsymbol{\beta}_2 = (0, 0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})'$。我們比較 SIR、cbSIR 在 NE 及 KM 切片方式下, 計算實驗 100 次所估計的前兩個主要維度縮減方向跟實際有效維度縮減向量之典型相關係數的平均和標準差。我們取切片數爲 $H_{(1)} = 2$, 集群內的切片數爲 $H_{(2)} = 5$。結果列於表 4。對於 SIR 方法於模型中, $\theta = 0$ 的狀況下, 以 NE 切片方式, 其前兩個典型相關係數平均爲 (0.999, 0.973); 另外用 KM 切片方式則其前兩個典型相關係數平均爲 (0.999, 0.819); 而對於 cbSIR, 其 NE 切片方法計算之典型相關係數的平均爲 (0.994, 0.627), 是四種切片方式在 $\theta = 0$ 狀況下表現最差的, 而 KM 切片方式計算典型係數的平均則爲 (0.993, 0.857); 當 $\theta = 1$ 在的狀況下, 傳統 SIR 方法的兩種切片結果計算典型係數的平均都是 (0.99, 0.99) 以上, 而 cbSIR 於 NE 的切片方式計算典型係數的平均爲 (0.991, 0.754), 雖然有比在 $\theta = 0$ 情況下的估計還來的準確, 但仍然是四種方法在 $\theta = 1$ 狀況下表現最差的, 另外一種 KM 切片方式所得爲 (0.998, 0.896)。

綜合上述結果, 可看出 cbSIR 中用 K 均值法做爲群集選取的方式結果會較好, 但在 Kuentz

表 5　針對 Li model (6.3) 模擬資料轉換後之區間型資料, 採用傳統及群集切片逆迴歸法, 以不同切片方式之下的充分維度縮減方向估計, 重覆模擬 100 次, 計算典型相關係數之平均及標準差 (眞實方向爲 (1,0,$\cdots$,0) 和 (0,1,0,$\cdots$,0))。

	(a) $H_{(1)}$=5, $H_{(2)}$=4				(b) $H_{(1)}$=5, $H_{(2)}$=20			
Method	σ=0.5		σ=1		σ=0.5		σ=1	
V-SIR	0.973	0.638	0.882	0.350	0.991	0.941	0.965	0.694
	(0.017)	(0.239)	(0.089)	(0.235)	(0.005)	(0.040)	(0.026)	(0.208)
C-SIR	0.646	0.244	0.650	0.221	0.946	0.765	0.914	0.511
	(0.131)	(0.149)	(0.185)	(0.157)	(0.029)	(0.135)	(0.060)	(0.230)
V-cbSIR	0.854	0.390	0.702	0.248	0.881	0.484	0.795	0.294
	(0.093)	(0.234)	(0.170)	(0.171)	(0.082)	(0.248)	(0.113)	(0.207)

model (8) 模型下以 NE 切片方式的 SIR 即可獲得很好的結果了。

5. 區間型符號之資料維度縮減模擬分析

本節是使用傳統及群集切片逆迴歸法, 以頂點式和中心式等不同區間轉換方法做區間型符號資料的維度縮減方向估計。首先依第 4 節的四個模型各生成一般傳統資料, 再依據下列所述方法各整合成一區間型資料。第 4 節中的四個模型生成樣本數爲 500 的資料, 對變數 y 排序後, 再依照此排序以每 100 個觀察值爲一群集 ($H_{(1)} = 5$)。依此設定, 我們比較兩種實驗: (1) 第一種爲每 25 個資料轉變爲區間型態, 所以每個子群集下會有 $H_{(2)} = 4$ 組區間型資料; 亦即每個解釋變數在每一群集內取其最小值及最大值, 便可構成一區間型資料。(2) 第二種是每 5 個資料轉變爲區間型態, 所以每個子群集下會有 $H_{(2)} = 20$ 組的區間型資料。

我們採用頂點式切片逆迴歸法 (V-SIR)、中心式切片逆迴歸法 (C-SIR) 和頂點式以群集爲基礎的切片逆迴歸法 (V-cbSIR) 等三種區間型切片逆迴歸法對上述兩種實驗做充分維度縮減方向估計。應用 V-cbSIR 時, 先以 $H_{(1)}$ 做第一次分群, 第二次切片則會依照第一次群集的結果在每一群集內再依 $H_{(2)}$ 做第二次分群。因爲每次實驗起始生成樣本數 500 個的資料都是重新生成, 所以每次實驗轉換成區間型資料會是獨立的, 我們重複模擬 100 次, 若採用模型中眞正維度縮減方向只有一個的話, 就取第一個維度縮減方向的估計量, 計算此 100 次的平均數和標準差; 若維度縮減方向爲 K 個時, 則取前 K 個主要維度縮減方向的估計量, 計算此 100 次典型相關係數的平均數和標準差做爲比較的依據。

模擬過程中, 資料第二次切片 ($H_{(2)}$) 的步驟會依照區間型資料內所構成的超立體做分群,

表 6　針對 Kuentz model (8) 模擬資料轉換後之區間型資料, 採用傳統及群集切片逆迴歸法, 以不同切片方式之下的充分維度縮減方向估計, 重覆模擬 100 次, 計算典型相關係數之平均及標準差 (眞實方向爲 $(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, 0)$ 和 $(0, 0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$)。

(a) $H_{(1)}$=5, $H_{(2)}$=4

method	$\theta = 0$		$\theta = 0.5$		$\theta = 1$	
V-SIR	0.992	0.894	0.996	0.964	0.998	0.976
	(0.011)	(0.092)	(0.005)	(0.032)	(0.003)	(0.018)
C-SIR	0.972	0.737	0.947	0.689	0.960	0.724
	(0.041)	(0.189)	(0.069)	(0.187)	(0.048)	(0.151)
V-cbSIR	0.919	0.398	0.960	0.598	0.962	0.590
	(0.113)	(0.251)	(0.053)	(0.294)	(0.054)	(0.259)

(b) $H_{(1)}$=5, $H_{(2)}$=20

method	$\theta = 0$		$\theta = 0.5$		$\theta = 1$	
V-SIR	0.998	0.974	0.999	0.987	0.999	0.990
	(0.003)	(0.021)	(0.002)	(0.009)	(0.001)	(0.007)
C-SIR	0.994	0.954	0.996	0.962	0.997	0.966
	(0.008)	(0.032)	(0.005)	(0.024)	(0.004)	(0.027)
V-cbSIR	0.927	0.514	0.974	0.603	0.962	0.529
	(0.096)	(0.245)	(0.035)	(0.264)	(0.055)	(0.287)

其結果不易發生區間內有最大值和最小值相同 (即 trivial interval) 的情況, 因此每個超立體所構成的個數是相同的, 亦即我們不需要考慮第二次切片是要採用每個切片趨近相等大小的方法還是 K 均值分群方法。此處我們只需要考慮 V-SIR 、C-SIR 和 V-cbSIR 三種區間型切片逆迴歸法方法對於估計方向是否準確即可。四個模型之模擬結果分析中, 因篇幅限制, 我們僅列出 Li model (6.3) 和 Kuentz model (8) 的結果於表5及6; 綜合這些模擬結果可知, 當 $H_{(2)} = 4$ 這種切片個數較少時, 以 V-SIR 方法較佳, 其估計維度縮減爲一個方向時與正確方向很接近且誤差較小, 但若估計維度縮減方向爲兩個方向時, 三個方法在對於第 2 個方向的典型相關係數估計都偏低。在 $H_{(2)} = 20$ 較高的切片個數狀況下, C-SIR 和 V-SIR 兩者估計都會準確估計維度縮減爲一個方向的模型, 特別是 Kuentz model (8) 中 C-SIR 的估計, 較不易受到模型中 θ 的影響估計; 對於維度縮減方向爲兩個時, V-SIR 和 C-SIR 兩種方法估計其典型相關係數的結果都較高, 尤其 V-SIR 方法較 C-SIR 佳; 而 C-SIR 要有較好的準確性還需要區間型資料個數較多的條件才能達成。至於 V-cbSIR 的表現並沒有明顯優於 V-SIR 及

表 7 Finance data (Vu, Vu & Foo, 2003)

$\boldsymbol{\xi}_u$	觀察變數 (縮寫)	Job Cost (Co)	Job Code (JC)	Activity Code (AC)	Monthly Cost (MC)	Annual Budget (AB)	
w_u	部門種類	$[a_1, b_1]$	$[a_2, b_2]$	$[a_3, b_3]$	$[a_4, b_4]$	$[a_5, b_5]$	n_u
w_1	快速消費類 (FMCG)	[1168, 4139]	[1, 7]	[1, 20]	[650, 495310]	[1000, 1212300]	605
w_2	汽車類 (Automotive)	[2175, 4064]	[1, 6]	[1, 20]	[400, 377168]	[400, 381648]	48
w_3	電信類 (Telecommu.)	[2001, 4093]	[1, 7]	[1, 20]	[1290, 250836]	[1290, 500836]	63
w_4	出版類 (Publishing)	[2465, 3694]	[1, 3]	[4, 20]	[5000, 42817]	[5000, 79367]	12
w_5	金融類 (Finance)	[2483, 4047]	[1, 7]	[2, 20]	[6500, 248629]	[6500, 313501]	40
w_6	顧問類 (Consultants)	[2484, 4089]	[1, 6]	[9, 20]	[4510, 55300]	[4510, 74400]	15
w_7	消費品類 (Consumer)	[2532, 4073]	[1, 6]	[2, 20]	[2900, 77500]	[2901, 77500]	34
w_8	能源類 (Energy)	[2542, 3685]	[1, 6]	[2, 20]	[2930, 54350]	[2930, 54350]	17
w_9	製藥類 (Pharmaceutical)	[2547, 3688]	[1, 7]	[1, 20]	[1350, 49305]	[12450, 1274700]	31
w_{10}	旅遊類 (Tourism)	[2604, 3690]	[1, 5]	[9, 20]	[1600, 31700]	[1600, 31700]	13
w_{11}	紡織類 (Textiles)	[2697, 4012]	[1, 5]	[9, 9]	[12800, 54850]	[12800, 94000]	15
w_{12}	服務類 (Services)	[3481, 4058]	[1, 1]	[9, 9]	[8400, 31500]	[8400, 41700]	3
w_{13}	耐用消耗品類 (Durable)	[2726, 4068]	[1, 5]	[10, 20]	[5800, 28300]	[3700, 55400]	3
w_{14}	其它類 (Others)	[3042, 4137]	[1, 1]	[1, 20]	[458, 19400]	[458, 19400]	34

C-SIR。

6. 區間型符號資料之維度縮減實例分析

根據上述模型研究結果, 在實例分析這部分中, 我們採用每群內之觀察個數趨近相等或 K 均值法的切片方式來比較 V-SIR、C-SIR 及 V-cbSIR。

6.1 金融紀錄資料: finance data

Finance data 爲一個金融紀錄的區間型資料 (Vu, Vu & Foo, 2003), 資料內已按照部門種類分群成 14 個組, 其部門名稱如表 7 所示, 而觀察變數已用區間值記錄下來, 分別爲 Co: 工作成本 (Job Cost)、JC: 工作代碼 (Job Code)、AC: 活動代碼 (Activity Code)、MC: 每月費用 (Monthly Cost)、AB: 年度預算 (Annual Budget) 和每一群組觀察到的公司數目 n_u。我們以維度縮減方法進行分析, 期望由兩個降維方向觀察此 14 組部門種類之分群及歸類狀況。首先我們將 Co $\sim$ AB 等五個區間變數進行 C-PCA 分析, 其二維主成份圖形如圖 1(a) 所示。

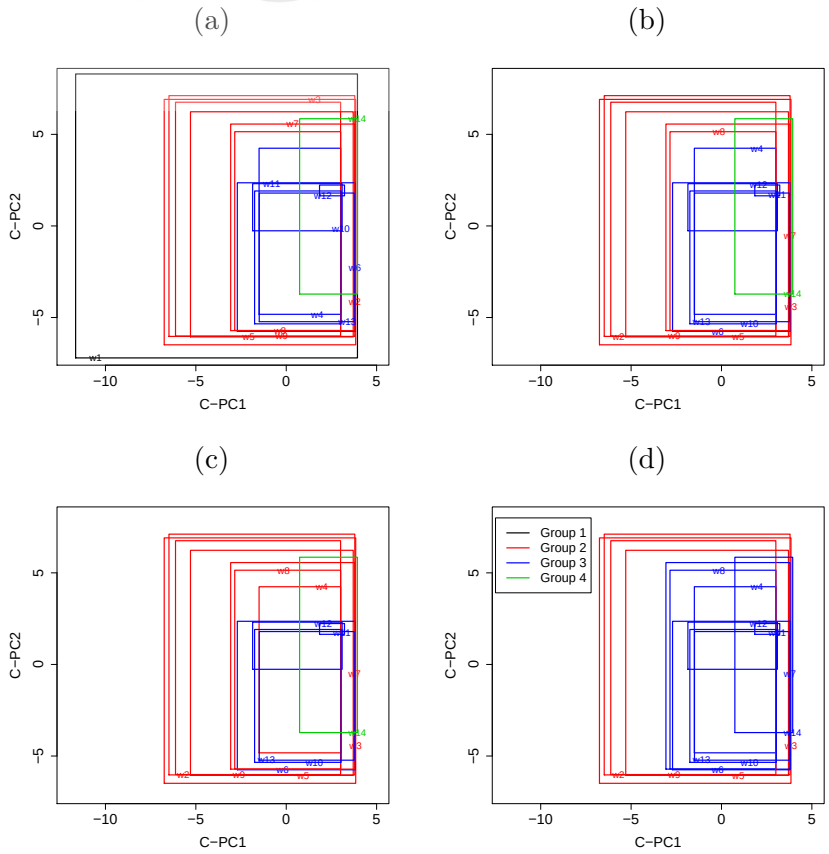

圖 1　Finance data 之 C-PCA 分析。(a) 在 Co、JC、AC、MC、AB 等五個變數下, faniance data 內 14 組區間型資料做 C-PCA 的結果; (b) 移除快速消費類 (w_1) 之後的 13 組區間型資料做 C-PCA 的結果; (c) 爲將出版類 (w_4) 從第二群轉爲第一群其圖示結果; (d) 爲將製藥類 (w_9)、消費品類 (w_7) 和其他類 (w_{14}) 通通歸爲第二群其圖示結果。

很明顯地, 快速消費類項目所繪出方框的形狀較大, 導致其他項目所繪出圖形較密集且較小, 我們可以只繪出除了快速消費類 以外的項目所表示的方框再做比較, 其結果列於圖 1 (b)(c)(d) 中, 可以發現出版類、製藥類、消費品類和其他類別的方框圖形的位置和形狀很接近, 於是可考慮將分群改爲兩種, 一種爲將出版類從第二群轉爲第一群, 如圖 1(c); 亦或是將製藥類、消費品類和其他類通通歸爲第二群, 如圖 1(d)。

將本資料的部門種類以低維度視覺化分群的考量可以很多種, 但將圖 1(c) 和 1(d) 兩圖做個比較後, 因製藥類和消費品類兩個項目皆與第一群其他項目很相近, 而出版類圖形較其他第二群長些, 所以可將出版類歸爲第一群之中, 另外一種合理解釋爲出版類的觀察的公司數目較其他第一群小 (服務類) , 以至於方框內差異小使方框較小讓人誤以爲屬於第二群的類別, 所以

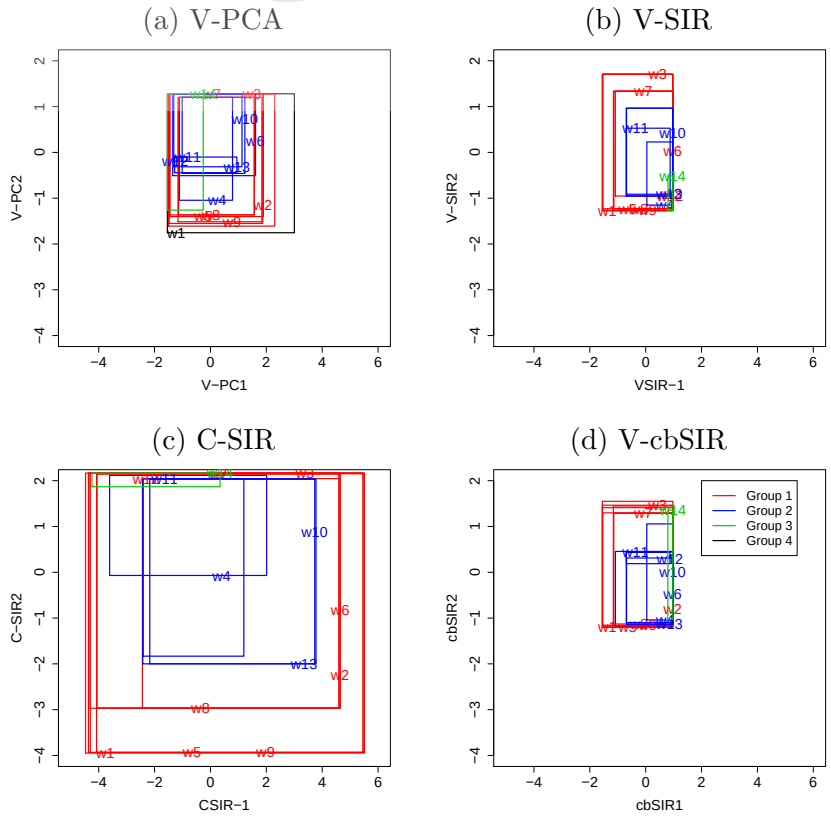

圖 2　使用 V-PCA 、V-SIR 、C-SIR 及 V-cbSIR 四種方法對 finance data 降維後之視覺化結果。

圖 1(c) 的分群結果較爲適當。

接著我們僅以 Co、JC、AC 三個變數進行 V-PCA 分析 (圖 2(a)), 由二維主成份圖中可依照群組位置和大小來進行圖形分群或歸類: 第一群爲汽車、電信、金融、製藥、消費品、能源 (w_2 、w_3 、w_5 、w_7 、w_8 、w_9) , 第二群爲出版、顧問、旅遊、紡織、服務和耐用消耗品類, (w_4 、w_6 、w_{10} 、w_{11} 、w_{12} 、w_{13}) , 第三群爲其他 (w_{14}), 第四群爲快速消費類 (w_1) 。比較這些分群可發現快速消費類比紡織和服務類圖形大了許多, 原因爲快速消費類的公司數目爲 605 個, 使其方框與其他項目相較之下有較大的內部差異; 另外一方面, 紡織和服務類這兩個觀察資料數目皆爲 3 筆, 與其他項目相比少之下許多, 因群內的差異小導致方框圖形較小。使用 V-SIR 、C-SIR 跟 V-cbSIR 這三種方法對此資料全部變數 Co ∼ AB 分析, 其 2 維視覺化結果列於圖 2(b)(c)(d), 對這些圖的分群的結果來說, 與上述使用 V-PCA 或 C-PCA 最大差異爲此資料可分爲三群, 因爲快速消費類與第二群的方框相近, 所以可將此兩群分類視爲同一群

集; 而 V-SIR 和 C-SIR 的分群結果相似, 其第一群爲快速消費、汽車、電信、金融、顧問、製藥、消費品和能源類, 第二群爲出版、旅遊、紡織、耐用消耗品類, 第三群爲其他類。在這結果中, 出版類 (w_6) 分群較爲困難, 因爲其視覺化結果所繪出的方框位置和大小介於第一群和第二群之間, 但考慮其觀察的公司的數目這個因素, 因出版類的個數較其他第一群內其他種類觀察的公司的數目少, 會使群內變化量較小, 造成方框形狀較小, 所以將出版類歸納於第一群。

而以群集爲基礎的切片逆迴歸法, 其分類爲第一群是快速消費、汽車、電信、金融、製藥和能源類, 第二類爲出版、顧問、旅遊、紡織、服務和耐用消耗品類, 第三類爲其他類。在此分群中, 出版類 (w_4) 分類比較有問題, 因爲其位置介於第二群和第三群之間, 但考量其種類觀察的公司的數目較第三類少可是方框形狀卻大很多, 所以將出版類歸納於第二群。

6.2 人臉自動辨識資料: face data

近年來因爲對安全監視的重視使得影像辨識受到關注。在人臉辨識過程中, 由於無法對人臉做實際測量, 只能識別臉部之主要特徵, 例如鼻子、眼睛和嘴巴的寬度等等, 而影像所選取到的臉孔是以圖片結果呈現, 其所劃定提取特徵邊界的具體點是局部的, 使資料的觀察值爲特徵點到特徵點之間的測量距離。這個距離的量測會依照人臉圖像的像素來表示, 因此資料所收集的是這些測量距離的區間值。

本研究中的 face data 是來自於 Leroy *et al.* (1996), 其收集了人臉自動辨識的資料, 共包含六個區間變數: 兩眼間最大的長度 (AD)、兩眼間最短的距離 (BC)、右眼外到上唇和鼻的中點之間的距離 (AH)、左眼外到上唇和鼻的中點之間的距離 (DH)、唇和鼻的中點至嘴唇最右邊的點之間距離 (EH) 及唇和鼻的中點至嘴唇最左邊的點之間距離 (GH)。此資料的觀察對象共九個人 (FRA、HUS、INC、ISA、JPL、KHA、LOT、PHI 和 ROM), 但由於角度、照明亮度、姿勢等不相同條件, 使每個人會有三組區間型資料, 所以整體觀察資料共有 27 組。

我們將此資料用 V-PCA、V-SIR、C-SIR 和 V-cbSIR 方法做辨識比較, 其二維視覺化結果列於圖 3 左欄, 比較這些圖的結果, V-SIR 方法較能對於此資料做辨識分群的工作, 原因在於 V-SIR 方法可將不同觀察對象所表示的方框, 其彼此間的距離較以 V-PCA 的圖形所呈現的遠, 使我們容易辨識哪些觀察對象爲相同群集; 若因圖形方框數目過多而造成難以辨識群集狀況, 我們還可從每個頂點所提供的貢獻度幫助辨識。我將四種維度縮減方法中, 設定貢獻度臨界值爲$\alpha = 0.2$, 其視覺化結果列於圖 3 右欄。比較不同 α 值的 V-PCA 視覺化, 其結果可略分爲三

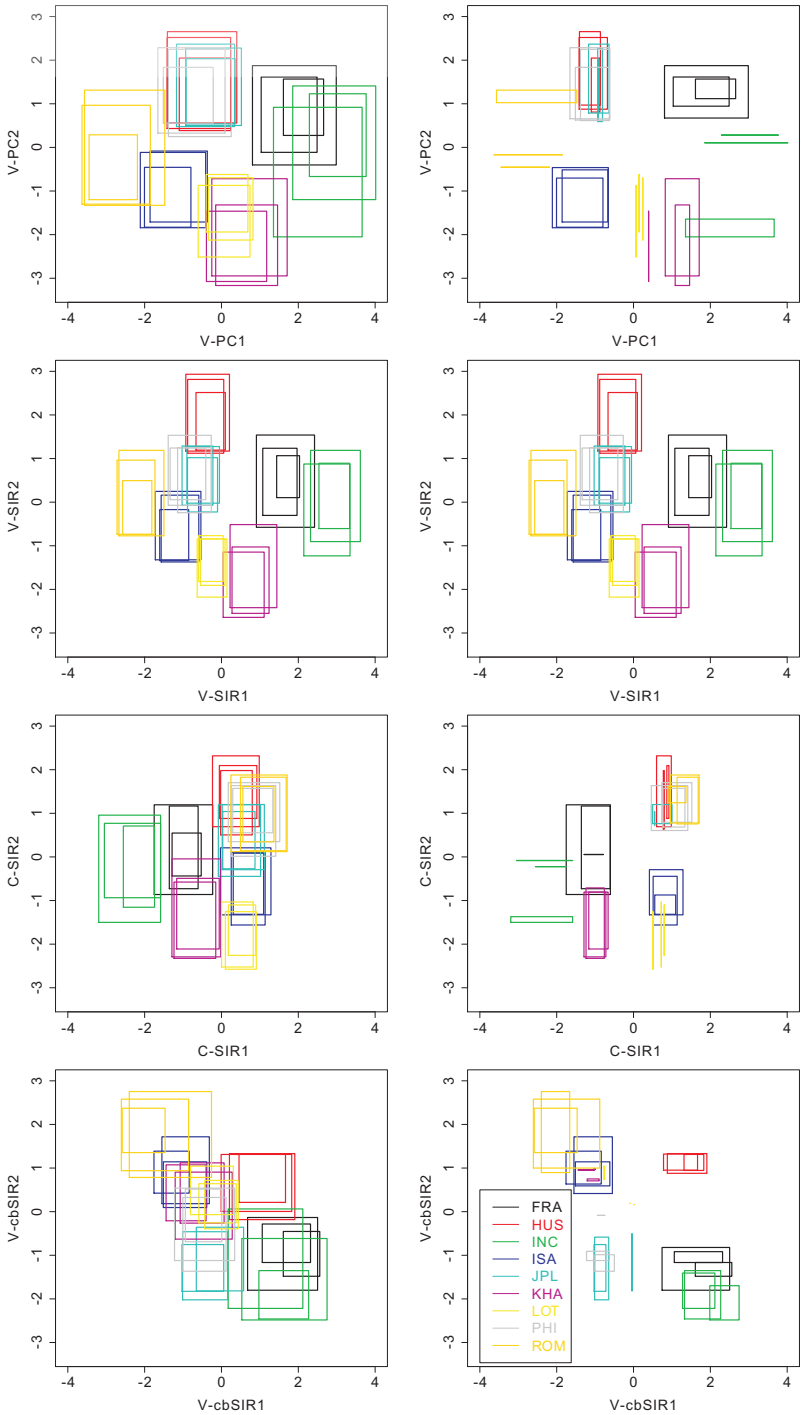

圖 3　以不同貢獻度(左欄: $\alpha = 0$; 右欄: $\alpha = 0.2$) 觀察 V-PCA、V-SIR、C-SIR 及 V-cbSIR 對 face data 降維後之視覺化的結果。

表 8　對 face data 區間型資料採用 V-PCA 及 V-SIR 方法, 其前3 個有效維度縮減方向與各變數的相關係數。

	V-PCA1	V-PCA2	V-PCA3	V-SIR1	V-SIR2	V-SIR3
AD	-0.644	0.589	-0.172	-0.238	0.539	-0.269
BC	-0.490	0.666	0.140	-0.105	0.304	-0.229
AH	-0.837	-0.197	0.371	-0.699	-0.559	-0.486
BH	-0.891	0.088	-0.165	-0.639	0.278	0.550
EH	0.475	0.625	0.561	0.051	0.123	-0.337
GH	0.428	0.755	-0.338	0.058	0.328	-0.100

群, 第一群爲{FRA、INC}, 第二群爲{KHA、LOT、ISA、ROM}, 第三群爲{HUS、JPL、PHI}; 至於 V-SIR 方法的視覺化, 其分群結果與 V-PCA 相同; 而 C-SIR 方法是藉由 $\alpha = 0.2$ 的視覺化圖形來看比較容易分辨, 其分群結論爲三群, 第一群爲{FRA、INC、KHA}, 第二群爲{LOT、ISA}, 第三群爲{HUS、JPL、PHI、ROM}; 然而 V-cbSIR 方法也是藉由 $\alpha = 0.2$ 的視覺化圖來看比較容易分辨, 其分群結果可爲四群, 第一群爲{FRA、INC}, 第二群爲{JPL、PHI}, 第三群爲{LOT、ISA、ROM、KHA}, 第四群爲{HUS}。

最後, 考慮資料維度縮減方向與變數之間的相關性, 我們將 V-PCA 方法和 V-SIR 方法所估計的維度縮減方向與變數之間的相關性結果列於表 8。就其結果來看, V-PCA 方法其相關性最高的變數爲 AD、AH 和 BH; 而 V-SIR 方法則是 AH 和 BH 兩個較高, 原因在於一個人的臉孔最大的特徵會是在於臉的大小, 而 AD、AH 和 BH 這三個變數與其他變數相較下來是比較容易決定臉孔大小的變數, 而剩下的變數與維度縮減第二個方向相關性較高, 其所代表的含意爲人臉辨識部位的細節。

7. 結論與討論

本研究中, 以頂點法及中心法, 實現了切片逆迴歸法於區間型資料的分析, 我們進一步採用以群集爲基礎的切片逆迴歸法來分析, 並與 C-PCA 、V-PCA 等方法相比較。我們所提出的方法, 其主要優點之一是透過區間資料的轉換, 而不改變演算法本身之下達成維度縮減的目的。因此各軟體套件如R 的 dr 套件即可直接應用分析, 無形中增加了新方法的推廣性。從模

擬實驗中可看出 V-SIR 方法的準確性較高, 其次爲 C-SIR 方法; 而在實例中也得到了驗證。若要提高用 V-SIR 和 C-SIR 估計有效維度縮減方向的準確性, 就需要有更多的區間型資料個數。

以群集爲基礎的切片逆迴歸法其所估計出的維度縮減方向存在著不穩定性的問題。以一般傳統點值資料爲分析對象, 應用 cbSIR 時, 通常建議以 K均值法的方式決定群集內切片的區域, 其所選取的分群點會是比較相近的; 但對於區間型資料的分析時, 雖有按照產生超立體的所有頂點的方式轉換使得 SIR 相關方法即可應用, 但其關鍵在於分群方式的決定。我們將一組區間資料所生成的超立體視爲一個分群, 但所構成的每個超立體在 $\Re^p$ 空間中可能會有重疊或相近的狀況發生, 在此情況下, 若是從每個超立體生成的頂點當作切片的範圍, 其切片之間所選取的內容是相似的, 也因爲如此, 以 V-cbSIR 對於區間型資料的正確維度縮減方向的估計準確性偏低。要改進此缺點, 可能的做法是要從生成所有頂點座標後, 再對所有轉換後的應變數做K均值分群法, 使相近的頂點座標以之做爲切片的根據; 另外, K 均值方本身是適用於連續型的變數, 而對於區間型的資料可能需要改變其距離計算公式, 才能改善切片方法對於區間型符號資料的準確性。本研究僅對於區間型資料做維度縮減及分析, 尚有其他類型的符號資料可供應用或推廣。而針對區間型資料, 其切片及維度個數的決定也需要進一步研究。

致謝

作者感謝主編及審稿委員的指正與建議, 使本論文更加完善。本研究由行政院國家科學委員會經費補助支持, 特此誌謝 (計畫編號: NSC 101-2118-M-032-012-)。

參考文獻

Becker, C. and Fried, R. (2003). Sliced inverse regression for high-dimensional time series. *Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifickation*, University of Munich, 3–11.

Billard, L. and Diday, E. (2006). *Symbolic data analysis: conceptual statistics and*

data mining. John Wiley & Sons.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, **98**, 470–487.

Cazes, P., Douzal-Chouakria, A., Diday, E. and Schecktman, Y. (1997). Extension de l'analyse en Composantes Principales des donnes de Type Intervalle," *Revue Statistique Applique*, **45**, 5–24.

Douzal-Chouakria, A., Diday, E. and Cazes, P. (1998). An improved factorial representation of symbolic objects. In: *Advances in Data Science and Classification* (eds. A. Rizzi, M. Vichi and H.-H. Bock), Springer-Verlag, Rome, 397–402.

Douzal-Chouakria, A., Billard, L. and Diday, E. (2011). Principal components for interval-valued observations. *Statistical Analysis and Data Mining*, **4**(2), 229–246.

Ferré, L. and Yao, A. F. (2003). Functional sliced inverse regression analysis. *Statistics*, **37**, 475–488.

Gioia, F. and Lauro, C. N. (2006). Principal component analysis on interval data. *Computational Statistics*, **21**, 343–363.

Ichino, M. (1988). General metrics for mixed features - the cartesian space theory for pattern recognition. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, **1**, 494–497.

Ichino, M. and Yaguchi, H. (1994). Generalized minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, **24**, 698–708.

Kuentz, V. and Saracco, J. (2010). Cluster–based sliced inverse regression. *Journal of the Korean Statistical Society*, **39**(2), 251–267.

Lauro, C. N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic analysis approach. *Computational Statistics*, **15**, 73–87.

Leroy, B., Chouakria, A., Herlin, I. and Diday, E. (1996). Approche géométrique et classification pour la reconnaissance de visage. *RFIA'96: Reconnaissance des Formes et Intelligence Artificielle*, pp. 548-557. Rennes.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of The American Statistical Association*, **86**, 316–342.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**(18), 3406–3412.

Li, K. C., Wang, J. L. and Chen, C. H. (1999). Dimension reduction for censored regression data. *Annals of Statistics*, **27**(1), 1–23.

Li, L. and Yin, X. (2009). Longitudinal data analysis using sufficient dimension reduction method. *Computational Statistics & Data Analysis*, **53**(12), 4106–4115.

Nagabhushan, P., Gowda, K. C. and Diday, E. (1995). Dimensionality reduction of symbolic data. *Pattern Recognition Letters*, **16**, 219–233.

Palumbo, F. and Lauro, N. C. (2003). A PCA for interval valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J.J. Meulman, editors, *New developments in Psychometrics*, Springer-Verlag, Tokyo.

Vu, T.H.T, Vu, T.M.T. and Foo, R.W.S. (2003). Analyse de données symboliques sur des projects markering. Technical Report, CEREMADE, Dauphine. Université Paris IX.

Wu, H. M. and Lu, H.H.-S. (2004). Supervised motion segmentation by spatial-frequential analysis and dynamic sliced inverse regression. *Statistica Sinica*, **14**, 413–430.

Wu, H. M. and Lu, H.H.-S. (2007). Iterative sliced inverse regression for segmentation of ultrasound and MR Images. *Pattern Recognition*, **40**(12), 3492–3502.

Zuccolotto, P. (2011). Principal component analysis with interval imputed missing values. *AStA Advances in Statistical Analysis*, **96**(1), 1–23.

THE APPLICATION OF SLICED INVERSE REGRESSION FOR DIMENSION REDUCTION OF THE INTERVAL-VALUED SYMBOLIC DATA

Ye-Shiun Chen and Han-Ming Wu

Department of Mathematics, Tamkang University

ABSTRACT

Sliced inverse regression (SIR) was introduced by Li (1991) to find the effective dimension reduction directions for exploring the intrinsic structure of high-dimensional data. For univariate response regression, SIR has been extended and applied to different data types. Examples were the cases of the survival data, the time series data, the functional data and the longitudinal data. This study intends to develop SIR for the interval-valued symbolic data. Firstly, the interval-valued data was transformed into the conventional data matrix using the vertices method or the centers method. Then the classical SIR algorithm was directly applied to the transformed data. The simulation results shown that using different slicing schemes produced different projection directions and different lower-dimensional visualization. Therefore, a suitable slicing scheme is needed for correctly investigating the embedded structure and information of the high-dimensional interval-valued symbolic data in the lower-dimensional plots. The results motivated us to adopt the clustered-based SIR to improve the implementation of the symbolic SIR. We compared and evaluated the results with those obtained with several existing symbolic dimension reduction techniques (such as the symbolic principal component analysis) for discriminative and visualization purposes.

Key words and phrases: data visualization, inverse regression, sufficient dimension reduction, symbolic data analysis, symbolic principal component analysis.

JEL classifications: C14, C63