Journal of the Korean Data & Information Science Society
2014, **25**(5), 1069–1077

http://dx.doi.org/10.7465/jkdi.2014.25.5.1069
한국데이터정보과학회지

클럽발 자료를 위한 함수적 군집 분석: 사례연구[†]

이미애[1] · 임요한[2] · 박천건[3] · 이경은[4]

[1]롯데카드 신용기획팀 · [2]서울대학교 통계학과 · [3]경기대학교 수학과 · [4]경북대학교 통계학과

접수 2014년 6월 27일, 수정 2014년 8월 19일, 게재확정 2014년 8월 27일

요 약

클럽발은 발이 안쪽으로 굽어있는 상태로 태어나는 선천적인 발 기형의 일종이다. 본 연구에서는 한 쪽 클럽발 환자들의 수술 후 시간에 따른 양 쪽 발의 상대적인 차이 커브들을 군집분석 하려고 한다. 관측값들이 일정하지 않은 (irregular) 시점에서 희박하게 (sparsely) 관측되어서 일반적인 함수적 군집모형을 사용할 수 없어 James와 Sugar (2003) 가 제안한 희박한 자료의 함수적 군집 모형 (functional clustering model)을 이용하여 모수들을 추정하였다. 그리고 Sugar와 James (2003)의 왜곡함수 (distortion function)를 이용하여 군집의 수를 결정하여 군집분석하여 두 개의 군집을 발견하였다.

주요용어: 군집분석, 모형에 기반한 군집분석, 클럽발, 희박한 함수적 자료.

1. 서론

클럽발 (club foot, 혹은 congenital talipes equinovarus)은 한 쪽 혹은 양 쪽 발목이 안쪽으로 굽어 있는 상태로 태어나는 선천적 발 기형의 일종으로 유전자의 결함과 같은 유전적 원인 혹은 양수부족으로 인한 자궁압착과 같은 외부적 원인으로 인한 것으로 알려져 있다. 1000명 당 한 명 정도가 선천적 기형을 가지고 태어나며, 그 중 약 50% 정도는 양 쪽 발목 (양측 기형)에 생긴다고 보고되고 있다. 또한 여아보다 남아에서 발생 빈도가 두 배 정도 높은 것으로 알려져 있다 (Wikipedia, Figure 1.1).

증상이 심한 정도와 관계없이 적절한 치료를 반드시 받아야 하며 적절한 치료 없이는 정상적으로 걸을 수 없다. 증상이 가벼운 유아들은 깁스 붕대 (plaster cast), 테이핑 (taping), 물리치료 등 복합적인 치료를 통해 6주에서 8 주 정도 안에 회복될 수 있으나 아주 심한 중증을 앓고 있는 유아들 (약 5% 미만)은 앞에서 언급한 치료로는 회복이 불가능하며 반드시 수술을 받아야 회복될 수 있다. 또한, 초기에 성공적인 치료에도 불구하고 재발 가능성이 높아 재발을 방지하기 위하여 몇 년 동안 버팀목 (bracing)을 대고 있어야 한다 (OrthoInfo). 한 쪽 클럽발 환자들은 클럽발 쪽의 종아리와 발의 크기가 정상 쪽보다 작으며, 적절한 치료 후에도 완전한 크기로 회복되기 힘들다고 알려져 있다. 적절한 치료 후 시간이 경과할 수록 상대적인 차이가 어느 정도 작아지면 치료방법이 성공적인 것으로 간주한다.

본 논문에서는 미국 캘리포니아 주에 있는 한 대학병원에서 한 쪽 클럽발 환자들이 수술 후 시간이 경과함에 따라 발 크기가 정상적으로 돌아오는 지, 어떠한 패턴을 띄고 있는 지 알고자 하여 2001년 수집

[†] 본 연구는 2012 년도 한국연구재단 연구비 (NRF-2012R1A1A3013075) 의하여 연구되었음.
[1] (100-778) 서울특별시 중구 소월로 7, 롯데카드(주) 신용기획팀, 대리.
[2] (777-111) 서울특별시 관악구 관악 1로, 서울대학교 통계학과, 교수.
[3] (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 수학과, 조교수.
[4] 교신저자: (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과, 부교수. Email: artlee@knu.ac.kr

Table 1.1 Club foot data

ID	Weeks	leg1 (cm)	leg2 (cm)	ft1(cm)	ft2 (cm)	rel. leg diff.	rel. ft. diff.
1	10	17	19	12	13.5	10.5	11.1
2	155	29.5	35.5	24.5	27.4	16.9	10.5
3	106	26	27	20	20.5	3.7	2.4
4	5	16	17.5	10	10.5	8.6	4.8
4	8	19.5	19.5	10.3	11.7	0	12
4	13	18	18	11.5	12	0	4.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

된 환자자료를 분석하려고 한다. 이 자료는 수술을 받은 67명 한 쪽 클럽발 환자들의 수술 후 경과시간 (주단위), 클럽발 쪽 다리길이 (leg1), 정상 쪽 다리 길이 (leg2), 클럽발 쪽 발 길이 (ft1), 정상쪽 발의 길이 (ft2), 정상 쪽과 클럽발 쪽의 다리길이 상대차이 (rel. leg diff.), 정상 쪽과 클럽발 쪽의 발 길이 상대차이 (rel. ft. diff.)를 포함하고 있다 (참고 Table 1.1, 자료 중 일부). 그 중에서 관측치가 2개 이상인 20명 환자들만 고려하였다.

Figure 1.1 Club foot (Source:Wikipedia)

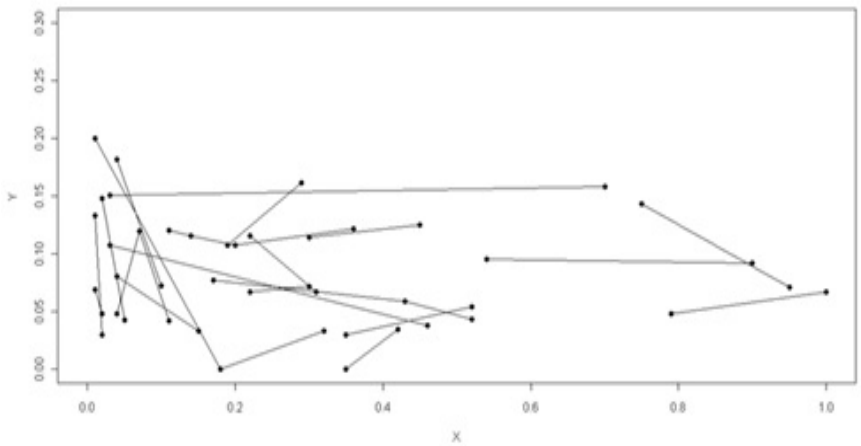

Figure 1.2 Plot of the relative differences of feet length

앞에서 언급한 것처럼 본 논문의 주 관심사는 수술 후 다양한 회복 패턴들을 발견하는 것이다. 즉, 상대적 발 길이 차이 커브들을 군집 분석하고자 한다. 수술 후 회복 패턴에 따라 수술 후 추후 치료법을 결정하는 데 도움이 될 것으로 기대된다.

Figure 1.2는 수술 후 시간 (단위:주)이 경과함에 따라 정상발과 클럽발의 상대적인 차이인 (정상발 크기- 클럽발 크기)/정상발크기를 그린 것으로 각 환자마다 단지 2 ∼ 4개 관측 값만 있을 정도로 자료가 굉장히 희박하고 또한 관측 시간들도 매우 불규칙하다.

종단형 자료 (longitudinal data), 예를 들어, 청소년 패널자료 (Lee와 Kang, 2013)은 각 서브젝트 (subject)가 다양한 시점에서 반복된 측정치를 가지고 있는 자료들을 의미하며 종단형 자료 분석 방법 중 하나로 각 서브젝트의 궤적 (trajectory)을 모형화하기 위한 비모수적 방법으로 함수적 자료 분석으로 보나, 연구 관심사에 따라 종단형 자료 분석 혹은 함수적 자료 분석으로 구별하기도 한다 (Rice, 2004).

클럽발 자료처럼 희박한 함수적 자료를 모형화하는 방법은 혼합효과 (mixed-effects) 를 이용한 모형들과 국소적 평활기 (local smoother)를 이용한 모형들로 크게 나눌 수 있다. 예를 들어, James 등 (2001)과 James와 Sugar (2003)는 B-spline을 통한 축소된 계수 (reduced-rank) 혼합모형 (mixed-effects) 을 이용하여 남녀 280명의 척추에서 무기질 함유량을 각각 2 ∼ 4 시기에 관측한 자료에 적용하여 분석하였다. Yao 등 (2005)은 국소적 평활기를 이용한 PACE (principal components analysis through conditional expectation) 추정치들을 토대로 함수적 주성분 점수 (functional principal component scores)를 이용하여 283명 동성애자들의 CD4 퍼센티지를 1 ∼ 14 시기에 관측한 자료에 적용하여 분석하였다.

일반적으로 함수적 자료를 군집 분석하는 방법은 크게 세 가지로 나누어 볼 수 있다. 첫 번째로, 각 커브마다 관측된 시간이 동일한 경우에 사용될 수 있는 레귤라이제이션 (regularization) 방법은, 다변량 자료처럼 취급하여 군집 분석하는 방법으로, 특별히 고차원자료인 경우에는 추정할 모수가 너무 많아 공분산 행렬에 제약조건을 걸기도 한다. 두 번째 방법인 필터링 (filtering) 방법은 각 커브를 저 차원 베이시스 함수에 사영 (projection)하여 생성된 계수들을 이용하여 군집 분석하는 방법이다. 세 번째 방법은, 앞의 두 방법들을 혼합하여 사용하는 것이다 (James와 Sugar, 2003).

이 자료처럼 희박한 함수적 자료인 경우는 이러한 군집 분석 방법들을 사용하기가 어려워 본 논문에서는 James와 Sugar (2003)가 제안한 희박한 함수적 자료 (sparsely sampled functional data)를 위한 군집 모형을 클럽발 자료에 적용하려고 한다. 이 모형은 사실상 모든 형태의 함수적 자료들을 군집분석 할 수 있으나 특별히 자료가 희박한 경우에 더 유용하다라고 알려져 있다.

본 논문의 구조는 다음과 같다. James와 Sugar (2003)가 제안한 함수적 군집 모형과 군집의 수를 결정하는 방법을 2절에서 살펴보고 3절에서 클럽발 자료에 적용하여 보고 4절에서는 본 연구에 대한 결론을 제시하려고 한다.

2. 함수적 자료를 위한 군집 모형

이번 절에서는 James와 Sugar (2003)가 제안한 희박한 함수적 자료를 위한 군집모형에 대하여 설명하려고 한다.

$g_i(t)$ 를 t 시점에서의 i번째 개체(곡선)의 참값이라 하자. $\mathbf{g}_i$, $\mathbf{Y}_i$와 $\boldsymbol{\epsilon}_i$은 시점 $t_{i1}, t_{i2}, \ldots, t_{in_i}$에 대응되는 참값, 관측값, 측정오차라 하자. 그러면

$$\mathbf{Y}_i = \mathbf{g}_i + \boldsymbol{\epsilon}_i, i = 1, \ldots, n,$$

여기서, n은 개체의 수를 나타낸다. 측정오차는 평균이 $\mathbf{0}$이고 서로 무상관 (uncorrelated)하며 $\mathbf{g}_i$와도

무상관하다. 이것은 미 관측된 시점이 임의 결측이라는 것을 의미한다. 수학적으로 좋은 성질을 가지고 있는 자연 3차 스플라인 (natural cubic spline) 기저함수 (basis function)를 이용해 $g_i(t)$를 모형화하고 있다.

$$g_i(t) = \mathbf{s}(t)^T \boldsymbol{\eta}_i,$$

여기서 $\mathbf{s}(t)$는 p-차원의 스플라인 기저 벡터이고 $\boldsymbol{\eta}_i$는 스플라인 계수 벡터이다. 가우시안 분포를 이용해 $\boldsymbol{\eta}_i$가 다음과 같이 모형화 되었다.

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i, \boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{\Gamma}),$$

여기서 전체 군집의 수가 G일 때, $\mathbf{z}_i = \big(z_{i1}, \ldots, z_{iG}\big)/$이며, 미지의 군집 멤버쉽을 나타낸다.

군집의 평균을 한 단계 더 모수화 함으로 커브들을 유용한 저 차원표현 (lower dimensional representation)이 가능하며 $\boldsymbol{\mu}_k$를 다음과 같이 나타낼 수 있다:

$$\boldsymbol{\mu}_k = \boldsymbol{\lambda}_o + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k, \tag{2.1}$$

여기서 $\boldsymbol{\lambda}_o$와 $\boldsymbol{\alpha}_k$는 각각 p와 h 차원 벡터이고, $\boldsymbol{\Lambda}$는 $p \times h$ 행렬이다 (단, $h \leq \min(p, G-1)$). 정리를 하면, 함수군집모형은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \mathbf{Y}_i = \mathbf{S}_i(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_{z_i} + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i, i = 1, 2, \ldots, n, \\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}), \boldsymbol{\gamma}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}). \end{aligned} \tag{2.2}$$

여기서 $\mathbf{S}_i = (\mathbf{s}(t_{i1}), ..., \mathbf{s}(t_{in_i}))^T$는 i번째 곡선의 스플라인 기저행렬이다. $\boldsymbol{\epsilon}_i$와 $\boldsymbol{\gamma}_i$의 공분산 행렬인 $\mathbf{R}$과 $\boldsymbol{\Gamma}$가 가능한 형태가 많이 있지만, 여기에서는 단순하게 $\mathbf{R} = \sigma^2\mathbf{I}$를 가정하고 모든 군집에서 공통 $\boldsymbol{\Gamma}$를 사용한다. 특별한 제약조건이 없으면, $\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}$와 $\boldsymbol{\alpha}_k$는 서로 교락된다. 따라서 다음과 같은 제약조건들을 사용한다:

$$\sum_k \boldsymbol{\alpha}_k = \mathbf{0} \tag{2.3}$$

$$\boldsymbol{\Lambda}^T\mathbf{S}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Lambda} = \mathbf{I}, \tag{2.4}$$

여기서 $\mathbf{S}$는 자료의 모든 시점을 포함하고 있는 가장 세분된 시점에서 계산된 기저 행렬이고 $\boldsymbol{\Sigma} = \sigma^2\mathbf{I} + \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^T$. 제약조건 (2.1)로 인하여 $\mathbf{s}(t)^T\boldsymbol{\lambda}_0$가 전체 평균커브라고 생각할 수 있다. 함수적 군집모형 적합은 $\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}, \boldsymbol{\alpha}_k, \boldsymbol{\Gamma}$와 σ^2을 추정하는 것이다. 이것은 모형 (2.2)의 가정하에서, i번째 커브가 k번째 군집에 속했다는 조건부 하에서, 혼합 우도함수를 최대화함으로 얻어질 수 있다:

$$\mathbf{Y}_i \sim N(\mathbf{S}_i(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k), \boldsymbol{\Sigma}_i),$$

여기서 $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I} + \mathbf{S}_i\boldsymbol{\Lambda}\mathbf{S}_i^T$.

혼합 우도함수를 적합시키는 표준적인 접근 방법은 미지의 군집 멤버십을 결측변수로 간주하여 EM 알고리즘을 이용하는 것이다. $\mathbf{z}_i$와 $\boldsymbol{\gamma}_i$의 상호독립성 가정 때문에 결합분포는 $f\big(\mathbf{Y}, \mathbf{z}, \boldsymbol{\gamma}\big) = f\big(\mathbf{Y}|\mathbf{z}, \boldsymbol{\gamma}\big)f\big(\mathbf{z}\big)f\big(\boldsymbol{\gamma}\big)$이다. $\mathbf{z}_i$가 모수가 $\big(\pi_1, \pi_2, \ldots, \pi_G\big)$인 다항분포를 따른다는 조건하에, $\boldsymbol{\gamma}_i$는 다변량 정규분포 $N(\mathbf{0}, \mathbf{R})$를 따르고, $\mathbf{Y}_i$의 조건부 분포는 $N\big(\mathbf{S}_i\big(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k + \boldsymbol{\gamma}_i\big), \sigma^2\mathbf{I}\big)$이므로, 로그 우

도함수 (상수항제외)는 다음과 같다:

$$l\big(\pi_k,\boldsymbol{\Gamma},\sigma^2,\boldsymbol{\lambda}_0,\boldsymbol{\Lambda},\boldsymbol{\alpha}_i\big) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}\log(\pi_k) \tag{2.5}$$

$$-\frac{1}{2}\sum_{i=1}^{n}\Big(\log|\boldsymbol{\Gamma}|+\boldsymbol{\gamma}_i^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_i\Big) \tag{2.6}$$

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}\left[n_i\log\sigma^2+\frac{1}{\sigma^2}\|\mathbf{Y}_i-\mathbf{S}_i\big(\boldsymbol{\lambda}_0+\boldsymbol{\Lambda}\boldsymbol{\alpha}_k+\boldsymbol{\gamma}_i\big)\|^2\right]. \tag{2.7}$$

EM 알고리즘은 $\mathbf{Y}_i$와 현재 모수 추정값이 주어진 상태에서 (2.5), (2.6), (2.7)의 기댓값을 반복적으로 최대화하는 과정을 포함하고 있다. 세 부분이 각각 다른 모수들을 포함하고 있으므로 서로 독립적으로 최대화 할 수 있다. (2.5)의 기댓값은 다음처럼 세팅함으로 최대화 된다:

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\pi_{k|i}, \tag{2.8}$$

여기서 $\pi_{k|i}=P\big(z_{ik}=1|\mathbf{Y}_i\big)=\frac{f\left(\mathbf{y}|z_{ik}=1\right)\pi_k}{\sum_{j=1}^{G}f\left(\mathbf{y}|z_{ij}=1\right)\pi_j}$. 또한, (2.6)의 기댓값은 다음의 사실을 이용하여

$$\boldsymbol{\gamma}_i|\mathbf{Y}_i,z_{ik}=1\sim N\big((\sigma^2\boldsymbol{\Gamma}^{-1}+\mathbf{S}_i^T\mathbf{S}_i)^{-1}\mathbf{S}_i^T(\mathbf{Y}_i-\mathbf{S}_i\boldsymbol{\lambda}_0-\mathbf{S}_i\boldsymbol{\Lambda}\boldsymbol{\alpha}_i),(\boldsymbol{\Gamma}^{-1}+\mathbf{S}_i^T\mathbf{S}_i/\sigma^2)^{-1}\big) \tag{2.9}$$

다음처럼 세팅함으로 최대화 된다:

$$\boldsymbol{\Gamma}=\frac{1}{n}\sum_{i=1}^{n}E[\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T|\mathbf{Y}_i]=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{k|i}E[\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T|\mathbf{Y}_i,z_{ik}=1]. \tag{2.10}$$

다음으로, (2.7)의 기댓값을 최대화한다. 이 과정은 $\boldsymbol{\lambda}_0,\boldsymbol{\alpha}_k$와 $\boldsymbol{\Lambda}$의 열벡터가 각각 순차적으로 최적화하는 것이다. 첫 번째로,

$$\boldsymbol{\lambda}_0=\left(\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{S}_i^T\left(\mathbf{Y}_i-\sum_{j=1}^{G}\pi_{i|j}\mathbf{S}_i\left(\boldsymbol{\Lambda}\boldsymbol{\alpha}_k+\hat{\boldsymbol{\gamma}}_{ik}\right)\right), \tag{2.11}$$

여기서 $\hat{\boldsymbol{\gamma}}_{ik}=E\left(\boldsymbol{\gamma}_i|z_{ik}=1,\mathbf{Y}_i\right)$은 식 (2.9)을 이용해서 얻어진 것이다. 다음으로 $\boldsymbol{\alpha}_k$를 구할 수 있다:

$$\boldsymbol{\alpha}_k=\left(\sum_{i=1}^{n}\pi_{k|i}\boldsymbol{\Lambda}^T\mathbf{S}_i^T\mathbf{S}_i\boldsymbol{\Lambda}\right)^{-1}\sum_{i=1}^{n}\pi_{k|i}\boldsymbol{\Lambda}^T\mathbf{S}_i^T\left(\mathbf{Y}_i-\mathbf{S}_i\boldsymbol{\lambda}_0-\mathbf{S}_i\hat{\boldsymbol{\gamma}}_{ik}\right). \tag{2.12}$$

다음으로, 다음의 식을 이용하여 다른 모수들이 고정된 상태에서 $\boldsymbol{\Lambda}$의 각 열 벡터를 최적화시킨다:

$$\boldsymbol{\lambda}_m=(\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{k|i}\alpha_{km}^2\mathbf{S}_i^T\mathbf{S}_i)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{k|i}\alpha_{km}\mathbf{S}_i^T(\bar{\mathbf{Y}}_i-\sum_{l\neq m}\alpha_{kl}\mathbf{S}_i\boldsymbol{\lambda}_l-\mathbf{S}_i\hat{\boldsymbol{\gamma}}_{ik}), \tag{2.13}$$

여기서, $\boldsymbol{\lambda}_m,\boldsymbol{\lambda}_l$은 각각 $\boldsymbol{\Lambda}$의 m번째, l번째 열 벡터이고, α_{km}은 $\boldsymbol{\alpha}_k$의 m번째 원소이며, $\bar{\mathbf{Y}}_i=\mathbf{Y}_i-\mathbf{S}_i\boldsymbol{\lambda}_0$.

마지막 단계는 σ^2의 값을 다음과 같이 두는 것이다:

$$\begin{aligned}\sigma^2 =&\frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_k E\left[\left(\bar{\mathbf{Y}}_i-\mathbf{S}_i\boldsymbol{\lambda}\boldsymbol{\alpha}_k-\mathbf{S}_i\boldsymbol{\gamma}_i\right)^T\left(\bar{\mathbf{Y}}_i-\mathbf{S}_i\boldsymbol{\lambda}\boldsymbol{\alpha}_k-\mathbf{S}_i\boldsymbol{\gamma}_i\right)\right]\\ =&\frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_k\left[\left(\bar{\mathbf{Y}}_i-\mathbf{S}_i\boldsymbol{\lambda}\boldsymbol{\alpha}_k-\mathbf{S}_i\hat{\boldsymbol{\gamma}}_i\right)^T\left(\bar{\mathbf{Y}}_i-\mathbf{S}_i\boldsymbol{\lambda}\boldsymbol{\alpha}_k-\mathbf{S}_i\hat{\boldsymbol{\gamma}}_i\right)+\mathbf{S}_i Cov\left(\boldsymbol{\gamma}_i|\mathbf{Y}_i,z_{ik}=1\right)\mathbf{S}_i^T\right],\end{aligned} \tag{2.14}$$

여기서 $N=\sum_{i=1}^{n}n_i$.

이 알고리즘은 모든 모수가 수렴할 때까지 (2.8), (2.10), (2.11), (2.12), (2.13), (2.14)을 반복하는 것이다.

3. 자료 분석 및 결과

James와 Sugar (2003)의 R 코드 (http://www-bcf.usc.edu/∼gareth/research/fclust)를 이용하여 분석하였다. 고차원 자료의 군집분석의 어려움 중에 하나가 시각화이다. 함수적 자료인 경우는 차수가 높지만 시간을 x축으로 하고 y축을 함수 값으로 그림을 그리는 것은 쉬운 일이나 두 커브 사이의 상대적 거리를 정의하고 측정하기는 쉽지 않기 때문에 고차원 자료의 군집분석 결과를 시각화하는 것 또한 어렵다. 이러한 문제는 함수의 측정값이 희박한 경우는 더 어렵게 된다. James와 Sugar (2003)는 베이시스 함수들의 계수들을 다시 저 차원 공간으로 사영시킴으로써 각 커브를 저 차원 공간의 한 점으로 나타냄으로 시각화를 통해 군집들을 쉽게 파악할 수 있도록 했다. Figure 3.1은 각 커브들의 α와 수술 후 평균 경과 주 (week)의 산점도로 두 군집 (빨간색과 파란색)이 잘 구별된 것을 쉽게 알 수 있다.

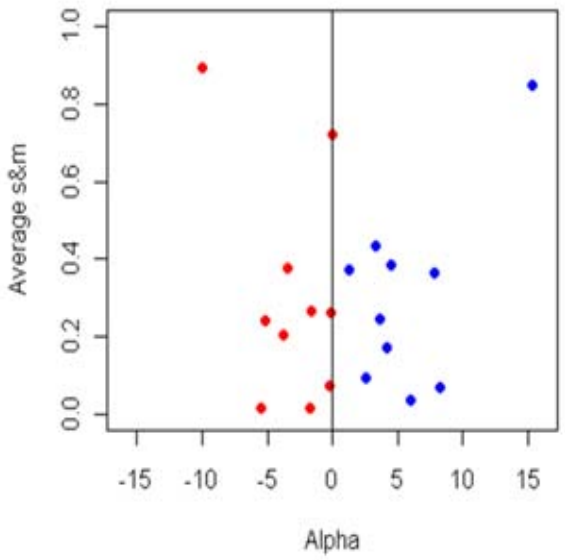

Figure 3.1 A Linear discriminant plot for club foot data

군집분석의 어려움 중 하나가 군집의 수를 정하는 것이다. 아직까지 최고의 방법은 알려져 있지 않고 있다. 우리는 Sugar와 James (2003)가 제안한 뒤틀림 함수 (distortion function)를 기초로 하여 군집의 수를 정하려고 한다:

$$d_k = \frac{1}{p}\min_{\mathbf{c}_1,\ldots,\mathbf{c}_k} E\left(\boldsymbol{\eta}_i - \mathbf{c}_{\boldsymbol{\eta}_i}\right)^T \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}_i - \mathbf{c}_{\boldsymbol{\eta}_i}\right),$$

여기서, $\boldsymbol{\eta}_i$들은 함수적 군집모형에서 스플라인 계수이다. 뒤틀림 함수 d_k는 각 $\boldsymbol{\eta}_i$와 $\boldsymbol{\eta}_i$와 가장 가까운 군집의 중심 $\mathbf{c}_i$와의 평균 마할라노비스의 거리를 뜻한다. Sugar와 James (2003) 는 혼합 분포 (mixture distribution)에서 정확한 혼합 요소 (mixture component)의 수에서 뒤틀림 함수의 가장 큰 점프를 가짐을 이론적으로 보였다. Figure 3.2는 군집의 수 $k = 2$에서 $k = 4$까지 점프 $d_k^{-1} - d_{k-1}^{-1}$를 그린 것으로 가장 큰 값인 2를 군집의 수로 정한다.

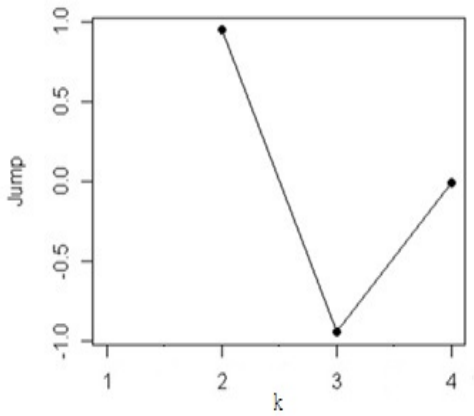

Figure 3.2 Jump plot for the club foot Data

앞 절에서 소개한 James와 Sugar (2003)의 함수적 군집모형의 장점 중 하나는 각 커브에서 관측되지 않은 부분도 예측이 가능하다는 것이다. 사실 이 점은 레귤라이제이션 방법이나 필터링 방법에서는 불가능한 일이다. 함수적 군집모형 하에서 가장 작은 MSE (mean squared error)를 가지는 $g_i(t)$의 추정량은 $\hat{g}_i(t) = s(t)^T E(\boldsymbol{\eta}_i|\mathbf{Y}_i)$이며 $E(\boldsymbol{\eta}_i|\mathbf{Y}_i)$는 다음과 같다 (James와 Sugar, 2003):

$$\hat{\boldsymbol{\eta}}_{M_i} = E(\boldsymbol{\eta}_i|\mathbf{Y}_i) = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\sum_{k=1}^{G}\boldsymbol{\alpha}_k\pi_{k|i} + \left(\sigma\boldsymbol{\Gamma}^{-1} + \mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\left(\mathbf{Y}_i - \mathbf{S}_i\left(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\sum_{k=1}^{G}\boldsymbol{\alpha}_k\pi_{k|i}\right)\right),$$

여기서,

$$\pi_{k|i} = P(z_{ik} = 1|\mathbf{Y}_i) = \frac{f(y|z_{ik}=1)\pi_k}{\sum_{j=1}^{G} f(y|z_{ij}=1)\pi_j}.$$

Figure 3.3은 두 군집에 속하는 각 커브들의 추정 커브와 평균 커브를 그린 것으로 수술 후 시간이 지남에 따라 패턴이 다른 것을 보여주고 있다. 첫 번째 군집에서는 전반적으로 완만한 회복을 보이고 있고 두 번째 군집에서는 빠른 회복의 패턴을 보이다가 다시 안 좋아지는 현상을 보이는 것으로 보이지만 사실상 관측된 값들이 초반에 모여 있고 후반에 소수의 관측 값들로 인하여 그런 영향을 받은 것으로 보인다. 두 번째 군집에서 평평한 한 커브는 첫 번째 군집의 속성에 가깝지만 커브의 y 값들이 두 번째 군집에 있는 커브들과 가깝기 때문에 두 번째 군집에 할당 된 것으로 보인다.

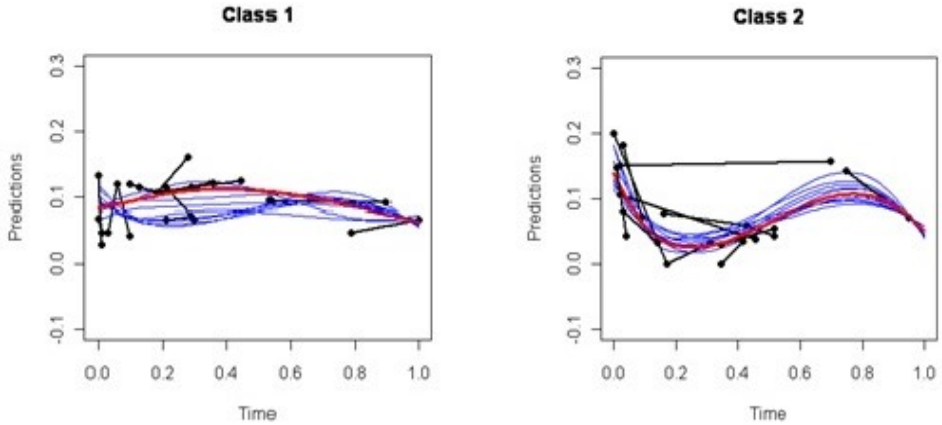

Figure 3.3 The comparison between cluster 1 and cluster 2

4. 결론

본 연구에서는 클럽발을 가지고 태어난 아기 환자들의 수술 후 경과하는 패턴을 알기 위하여 수술 후 시간에 따른 클럽발의 상대적인 발차이의 커브들을 군집분석하여 보았다. 특별히, 각 환자당 관측값 수가 2 ∼ 4개 정도 밖에 되지 않아 일반적인 함수적 자료의 군집분석 방법들을 이용하지 못하여 James와 Sugar (2003)가 제안한 희박한 함수적 자료를 위한 군집모형을 이용하여 군집분석을 하여보았고 점프함수를 이용하여 군집의 수를 2개로 결정하여 각 군집의 평균 커브를 추정하였다. 각 커브에서 관측값 수도 작고 환자도 20명 밖에 되지 않았고 무엇보다 관측치 대부분이 수술 후 경과시간 초반에 몰려 있어 분석의 한계가 있었다.

References

de Boor, C. (1978). *A Practical guide to splines*, Springer, New York.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611-631.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*, Chapman and Hall, London.

Hastie, T. J., Buja, A. and Tibshirani, R. J. (1995). Penalized discriminant analysis. *The Annals of Statistics*, **23**, 73-102.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society B*, **63**, 533-550.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**, 397-408.

Lee, H-J. and Kang, S-B. (2013). Analysis of latent growth model using repeated measures ANOVA in the data from KYPS. *Journal of theKorean Data & Information Science Society*, **24**, 1409-1419.

OrthoInfo. Children's clubfoot: treatment with casting or operation? Retrieved May, 2014, from `http://orthoinfo.aaos.org/topic.cfm?topic=a00296`.

Rice, J. A. (2004). Functional and longitudinal data analysis perspectives on smoothing. *Statistica Sinica*, **14**, 413-629.

Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, **90**, 928-934.

Yao, F., Muller, H-G. and Wang, J-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical Association*, **100**, 577-590.

Clubfoot. (n.d.) In Wikipedia. Retrieved May, 2014, from `http://en.wikipedia.org/wiki/Club\_foot`.

Journal of the Korean Data & Information Science Society
2014, **25**(5), 1069–1077

http://dx.doi.org/10.7465/jkdi.2014.25.5.1069
한국데이터정보과학회지

Functional clustering for clubfoot data: A case study[†]

Miae Lee[1] · Johan Lim[2] · Chungun Park[3] · Kyeong Eun Lee[4]

[1]Credit Planning Team, Lotte Card
[2]Department of Statistics, Seoul National University
[3]Department of Mathematics, Kyonggi University
[4]Department of Statistics, Kyungpook National University

Received 27 June 2014, revised 19 August 2014, accepted 27 August 2014

Abstract

A clubfoot is a kind of congenital deformity of foot, which is internally rotated at the ankle. In this paper, we are going to cluster the curves of relative differences between regular and operated feet. Since these curves are irregular and sparsely sampled, general clustering models could not be applied. So the clustering model for sparsely sampled functional data by James and Sugar (2003) are applied and parameters are estimated using EM algorithm. The number of clusters is determined by the distortion function (Sugar and James, 2003) and two clusters of the curves are found.

Keywords: Club foot, clustering, model-based clustering, sparse functional data.

[†] This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (No. NRF-2012R1A1A3013075).
[1] Senior assistant, Credit Planning Team, Lotte Card, Seoul 100-778, Korea.
[2] Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea.
[3] Assistant professor, Department of Mathematics, Kyonggi University, Gyeonggi-do 443-760, Korea.
[4] Corresponding author: Associate professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: artlee@knu.ac.kr