The Korean Journal of Applied Statistics
2022, Vol. 35, No. 2, 265–283

DOI: https://doi.org/10.5351/KJAS.2022.35.2.265
Print ISSN1225-066X / Online ISSN 2383-5818

Pairwise fusion approach to cluster analysis with applications to movie data

Hui Jin Kim[a,b], Seyoung Park[1,a]

[a]Department of Statistics, Sungkyunkwan University; [b]Department of Fintech, Sungkyunkwan University

Abstract

MovieLens data consists of recorded movie evaluations that was often used to measure the evaluation score in the recommendation system research field. In this paper, we provide additional information obtained by clustering user-specific genre preference information through movie evaluation data and movie genre data. Because the number of movie ratings per user is very low compared to the total number of movies, the missing rate in this data is very high. For this reason, there are limitations in applying the existing clustering methods. In this paper, we propose a convex clustering-based method using the pairwise fused penalty motivated by the analysis of MovieLens data. In particular, the proposed clustering method execute missing imputation, and at the same time uses movie evaluation and genre weights for each movie to cluster genre preference information possessed by each individual. We compute the proposed optimization using alternating direction method of multipliers algorithm. It is shown that the proposed clustering method is less sensitive to noise and outliers than the existing method through simulation and MovieLens data application.

Keywords: MovieLens, convex clustering, alternating direction method of multipliers (ADMM), missing imputation, optimization, pairwise fused penalty

1. 서론

현대 사회는 정보 과잉 시대로 진입하여 인터넷에 검색만 해도 수많은 관련 정보를 얻을 수 있게 되었다. 이에 따라 방대한 정보 중에서 특정 사용자가 관심을 가질 만한 정보들을 추천해주는 추천 시스템들 역시 활발히 생성되고 있다. 추천 시스템은 사용자의 선호도, 그리고 과거 행동을 바탕으로 개인에 맞는 관심사를 제공하는 분야를 말한다. 이러한 제공은 어떤 상품을 구매할지, 어떤 음악을 들을지 또는 어떤 영화를 볼 것인지와 같은 다양한 의사결정과 연관이 있다. Netflix의 경우, 사용자가 시청한 목록들을 분석한 내용을 바탕으로 사용자가 선호할만한 콘텐츠를 예측하여 추천해줌으로써 사용자의 만족도를 높여주고, 기업에는 콘텐츠 운영의 효율성을 높여 사용자 이탈을 막는 데 효과적으로 사용되고 있다. 본 논문에서는 추천 시스템이 개인화된 추천, 특히 영화 추천에 효율적으로 구성하는 데 도움을 주기 위한 연구를 진행하였다. 본 논문에서 제안하는 방법론은 미네소타 대학의 GroupLens Research Project에서 사용자들의 영화정보를 기록한 MovieLens 데이터의 특성을 반영한 방법론이다. MovieLens 데이터는 추천시스템 연구에서 아이디어를 탐색하고 검증하는 데 상당한 가치가 있는 데이터 (Harper와 Konstan, 2015)로, 기존 데이터 분할 및 군집화 알고리즘을 사용하여 사용자 평가 데이터를 기반으로 항목 집합을 분할하는 연구 (O'Connor와 Herlocker, 1999) 등에 사용되는 데이터이다.

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Seonggyungwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail : ishspsy@skku.edu

Published 30 April 2022 / journal homepage: http://kjas.or.kr

본 논문에서는 기존 연구에서 대표적으로 사용되었던 영화 평점 데이터와 영화 장르 데이터를 통해 사용자의 영화 장르 선호도를 예측하고 선호도 패턴을 기반으로 사용자를 군집화(clustering)하여 유의미한 정보를 얻는 연구를 진행하였다. 군집화란, 관측값의 몇 가지 공통적인 특성에 따라 다른 그룹으로 나누는 방법론으로 오래전부터 사용되었으며, 현재까지 다양한 군집화 기술이 개발되었다. Hartigan과 Wong (1979)의 K-means, Friedman과 Russell (2013)의 가우스 혼합 모델(Gaussian mixture model), Sibson (1973)의 계층적 군집화(hierarchical clustering), Ng 등 (2002)의 스펙트럼 군집화(spectral clustering)와 같은 알고리즘을 통해 다양한 군집화 모양을 인식할 수 있게 되었다. 이러한 방법은 비 볼록 최적화(non-convex optimization)이거나 거리의 임곗값에 의존하기 때문에 불안정성이 존재한다. 본 논문에서는 결측 비율이 높은 MovieLens 데이터 특성에 모티브를 얻어 결측치 대체(missing imputation)가 동시에 진행되는 쌍별 규합 벌점함수(pairwise fused penalty)를 활용한 볼록 군집화(convex clustering) 기반의 방법을 제안한다.

본 논문의 구성은 다음과 같다. 제2장에서는 볼록 군집화 기법과 MovieLens 데이터에 대해 소개하고, 새로운 군집화 모델을 제안한다. 제3장에서는 반복 알고리즘인 ADMM에 대해 설명하고, 제안하는 최적화 문제를 풀어 진행한 다음, 수렴 기준과 군집 개수의 추정에 대해 소개한다. 4장에서는 시뮬레이션 및 실데이터 분석을 진행하고, 마지막으로 제5장에서 본 논문에 대한 요약 및 향후 연구 과제에 대하여 논의한다.

2. 제안

본 장에서는 볼록 군집화 기법에 대해 소개하고, MovieLens데이터와 제안하는 모델에 대해 설명한다.

2.1. 볼록 군집화 기법

주어진 데이터 행렬 $X \in \mathbb{R}^{n\times p}$에서 n은 표본의 크기 그리고 p는 변수의 개수를 의미한다. 여기서 X의 표본을 군집화 하기 위해 Lindsten 등 (2011)과 Hocking 등 (2011)은 쌍별 규합벌점함수(pairwise fused penalty)를 활용하여 군집화를 유도하는 작업을 볼록 최적화 문제로 공식화했다. 이때 최적화 문제는 다음과 같다.

$$\min_{\alpha\in\mathbb{R}^{n\times p}} \frac{1}{2}\|\alpha - X\|_F^2 \quad \text{subject to} \quad \sum_{i<j} 1_{\alpha_i\neq\alpha_j} \le t. \tag{2.1}$$

여기서 $\|\cdot\|_F^2$는 프로베니우스 노름(Frobenius norm)의 제곱 형태이고, $\alpha_i \in \mathbb{R}^n$는 α의 i번째 행 벡터이다. $1_{\alpha_i\neq\alpha_j}$은 $\alpha_i \neq \alpha_j$이면 1이고 그렇지 않으면 0의 값을 가진다. $\sum_{i<j} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}$ 은 $n(n-1)/2$개의 행 벡터 쌍을 합한다. 만약 $t \ge n(n-1)/2$로 고정한다면 최적화 문제 (2.1)은 제약이 없고, 해는 모든 i에 대해 $\alpha_i = X_i$이다.

$t = 0$일 경우의 해는 $\alpha_i = \bar{X} = \sum_{i=1}^{n} X_i/n$이 된다. 일반적으로 식 (2.1)은 비볼록 최적화 문제로 간단한 방법으로 해를 도출할 수 없다. 따라서 다음의 볼록 완화(convex relaxation)를 제안하였다.

$$\min_{\alpha\in\mathbb{R}^{n\times p}} \frac{1}{2}\|\alpha - X\|_F^2 \quad \text{subject to} \quad \Omega_q(\alpha) = \sum_{i<j} w_{ij}\left\|\alpha_i - \alpha_j\right\|_q \le t. \tag{2.2}$$

여기서, $w_{ij} > 0$이고 $\|\cdot\|_q, q \in \{1, 2, \infty\}$는 $\mathbb{R}^n$의 l_q 노름(norm)이다. 식 (2.2)에서의 제약조건은 α의 행 벡터 차이에 희소성(sparsity)을 유도하는 역할을 한다. α의 행 벡터가 같아지면 해당되는 표본이 같은 군집을 형성한다고 볼 수 있으며, t를 변경하여 형성된 최적 해의 연속적인 정규화 경로(regularization path)를 군집화 경로(clusterpath)라고 부른다. 예를 들어 $t = 0$인 경우 α의 모든 행은 같은 벡터를 갖게 되며, 모든 표본은 같은 군집에 속하게 된다. 여기서 t에 대한 매개변수화는 $0 \le t \le \Omega_q(X)$를 취하지만, 계산의 편의를 위해 $0 \le s \le 1$인 다음의 매개변수화를 도입한다.

$$\min_{\alpha\in\mathbb{R}^{n\times p}} \frac{1}{2}\|\alpha - X\|_F^2 \quad \text{subject to} \quad \frac{\Omega_q(\alpha)}{\Omega_q(X)} \le s. \tag{2.3}$$

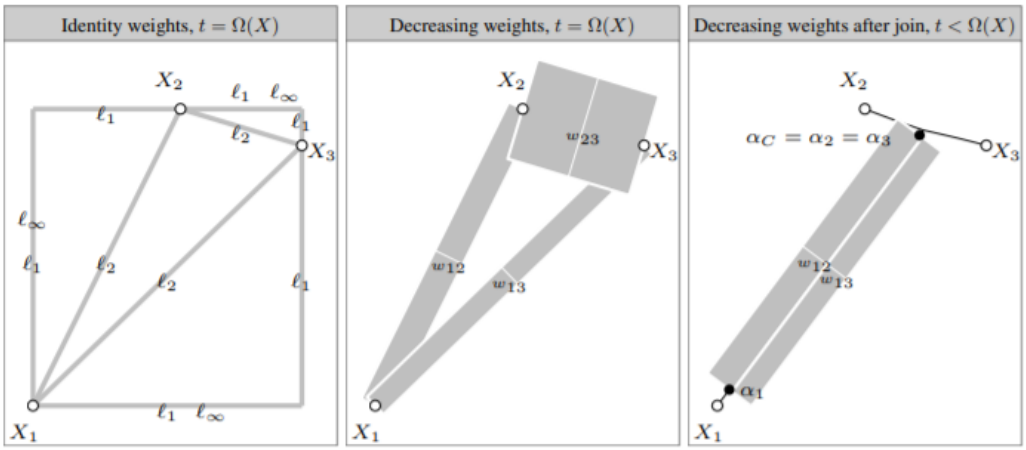


Figure 1: *Geometric interpretation of the optimization problem* (2.2) *for data* $X \in \mathbb{R}^{3\times 2}$ *(Hocking et al., 2011).*

또한 식 (2.3)의 라그랑지안 함수(Lagrangian function)를 고려할 수 있다.

$$\min_{\alpha\in\mathbb{R}^{n\times p}} \frac{1}{2}\|\alpha - X\|_F^2 + \lambda\Omega_q(\alpha). \tag{2.4}$$

여기서 λ는 조정 상수이고, $\lambda = 0$일 경우, $\alpha_i = X_i$일 때 최소값에 도달하고, X_i 벡터값에 따라 i의 군집이 결정된다. λ가 증가하면 서로 다른 군집의 중심이 합쳐지기 시작한다. 만약 $\alpha_i = \alpha_j$일 경우 i와 j는 동일한 군집 내에 속한다고 볼 수 있다. 식 (2.4)는 λ에 대해 고유한 최소점을 가지고 있다. Lindsten 등 (2011)은 식 (2.4)에서 l_q 노름(norm) 패널티를 고려하였고, Hocking 등 (2011)은 l_1 노름, l_2 노름 그리고 l_∞ 노름을 고려하였다. 위의 최적화 문제에서 $w_{ij} > 0$은 대부분 데이터 사이의 거리에 따라 감소하는 가중치 $w_{ij} = \exp(-\gamma\|X_i - X_j\|_2^2)$를 사용한다. K-means는 지역 최적화 알고리즘(local optimization algorithm)이기 때문에 초기값 결과에 따라 군집화 결과가 크게 변하는 반면, 식 (2.4)와 같은 볼록 이완 (convex relaxation)은 고유한 최소값(global minimizer)으로 수렴할 수 있는 알고리즘을 고려할 수 있다. 한편 식 (2.4)는 Tibshirani 등 (2005)의 fused lasso signal approximator와 유사하다.

Hocking 등 (2011)은 $X \in \mathbb{R}^{3\times 2}$ 인 경우에 대해서 최적화 식 (2.2)의 기하학적인 해석을 Figure 1로 나타내었다. 첫 번째는 $w_{ij} = 1$인 경우로, α의 최적해는 X_i와 X_j의 거리 합계에 대한 제약 조건 에 따라 X에 가까운 α에 해당한다. 제약 조건 $\Omega_q(\alpha) = \sum_{i<j} w_{ij}\|\alpha_i - \alpha_j\|_q \le t$는 회색 선으로 표시된 모든 X의 행 벡터 사이의 l_q거리를 의미한다. 두 번째는 일반 가중치 w_{ij}인 경우, l_q 제약 조건은 X의 행 벡터 사이의 너비가 w_{ij}인 직사각형의 전체 면적을 의미하고 이를 제한한다. 세 번째는 α의 최적해를 제한한 후, α_2와 α_3가 융합하여 $\alpha_C = \alpha_2 = \alpha_3$인 군집 C를 형성하고 두 개의 가중치는 더해져 $w_{1C} = w_{12} + w_{13}$가 된다. Figure 1에서 l_2경로는 데이터 X의 회전에 변하지 않는 반면 다른 경로는 그렇지 않다는 것을 확인할 수 있다.

2.2. MovieLens 데이터

본 논문에서는 영화 추천 서비스인 MovieLens의 영화 평점 데이터와 영화 장르 데이터를 이용하였다. 먼저, 영화 평점 데이터는 1995년 1월 9일부터 2019년 11월 21일 사이에 162,541명의 사용자가 62,423개의 영화에 대해 최저점인 0.5점에서 최고점인 5점 사이의 평점을 매긴 데이터로, 총 25,000,095개의 평점이 존재한다. 영화 장르 데이터는 62,423개의 영화에 대해 장르 정보가 존재한다. 본 논문에서는 영화 평점이 영화별 장르 가중치와 사용자의 고유한 장르 선호 점수의 곱으로 구해질 것이라고 가정한다. 따라서, 영화별 장르 가중치

Movie 1	Animation | Comedy | Romance
Movie 2	Animation | Romance
Movie 3	Romance
Movie 4	Comedy | Romance | Drama
Movie 5	Romance | Drama

Genre weight (Z)	Animation	Comedy	Romance	Drama
Movie 1	1/3	1/3	1/3	0
Movie 2	1/2	0	1/2	0
Movie 3	0	0	1	0
Movie 4	0	1/3	1/3	1/3
Movie 5	0	0	1/2	1/2

Figure 2: *Genre data preprocessing.*

Rating (Y)	User 1	User 2	User 3	User 4
Movie 1	2	3		2
Movie 2	1		4	
Movie 3		3		3
Movie 4	5		3	
Movie 5	2	1	4	5

$\simeq$

Genre weight (Z)	Animation	Comedy	Romance	Drama
Movie 1	1/3	1/3	1/3	0
Movie 2	1/2	0	1/2	0
Movie 3	0	0	1	0
Movie 4	0	1/3	1/3	1/3
Movie 5	0	0	1/2	1/2

$\times$

Genre preference (X)	User 1	User 2	User 3	User 4
Animation	x_{11}	x_{12}	x_{13}	x_{14}
Action	x_{21}	x_{22}	x_{23}	x_{24}
Comedy	x_{31}	x_{32}	x_{33}	x_{34}
Romance	x_{41}	x_{42}	x_{43}	x_{44}
Drama	x_{51}	x_{52}	x_{53}	x_{54}

Figure 3: *Relationship diagram between movie rating data Y, genre weight data Z, and genre preference data X.*

데이터와 사용자의 고유한 장르 선호 점수 데이터가 필요하므로 영화별 장르 가중치 데이터는 Figure 2와 같이 생성한다. Figure 2에서 영화1을 예로 들면, Animation, Comedy, Romance로 총 3개의 장르 정보를 가지고 있으므로, 오른쪽의 표와 같이 Animation, Comedy, Romance에 각각 가중치 1/3을 부여하여 가중치의 합이 1이 되도록 만든다. 영화 평점 데이터와 영화 장르 가중치 데이터 그리고 사용자의 고유한 장르 선호도 데이터에 대한 가정은 Figure 3에서 확인할 수 있다. Figure 3에서 사용자2가 영화1에 평가한 3점을 예로 들면, 영화1에 대한 장르 가중치 정보와 사용자2의 추정된 장르 선호도의 곱으로 구해진 점수라고 가정을 하는 것이다. 데이터의 관계를 식으로 표현하면 다음과 같이 정의 된다. 먼저, 영화의 개수를 p개, 사용자를 n명, 장르 개수를 k개로 정의할 때, 영화 평점 데이터 행렬 $Y \in \mathbb{R}^{p\times n}$과 영화 장르 가중치 데이터 행렬 $Z \in \mathbb{R}^{p\times k}$은 다음과 같이 정의된다.

$$Y = \begin{array}{c} \\ \text{movie1} \\ \text{movie2} \\ \vdots \\ \text{movie}p \end{array} \begin{array}{c} \begin{array}{cccc} \text{user1} & \text{user2} & \cdots & \text{user}n \end{array} \\ \begin{pmatrix} y_{11} & NA & \cdots & y_{1n} \\ NA & y_{22} & \cdots & NA \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{pmatrix} \end{array}, \quad Z = \begin{array}{c} \\ \text{movie1} \\ \text{movie2} \\ \vdots \\ \text{movie}p \end{array} \begin{array}{c} \begin{array}{cccc} \text{genre1} & \text{genre2} & \cdots & \text{genre}k \end{array} \\ \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pk} \end{pmatrix} \end{array}$$

행렬 Y의 원소 y_{ij}는 j번째 사용자의 i번째 영화에 대한 평점값으로 다음의 식을 만족한다고 가정한다.

$$y_{ij} = \sum_{h=1}^{k} z_{ih}x_{hj} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, p, \quad j = 1, \ldots, n. \tag{2.5}$$

여기서 행렬 Z는 영화별로 존재하는 장르 정보를 가중치화하여 다음의 식을 만족한다.

$$\sum_{h=1}^{k} z_{ih} = 1, \quad i = 1, \ldots, p.$$

2.3. 모델

본 논문에서는 쌍별 규합 벌점함수를 활용한 볼록 군집화 기법에 결측치가 존재하는 데이터의 특성을 고려한 결측치 대체 볼록 군집화 방법을 제안한다. 먼저, 목적 식에 추가로 필요한 행렬 $Y_1, \widetilde{Y}$에 대해 설명한다. 영화 평점 데이터 행렬 Y의 결측치 인덱스 쌍의 집합을 $\Omega = \{(i, j) : y_{ij}$는 결측치$\}$로 정의할 때, $Y_1 \in \mathbb{R}^{p\times n}$은 아래와 같이 정의하며, 여기서 $\overline{Y}_{i\cdot}$은 결측치를 제외한 i번째 행 평균이다. $\widetilde{Y}$는 본 논문에서 사용한 반복 알고리즘을 통해서 아래와 같이 결측치가 대체된 행렬로 정의한다. 알고리즘에 대한 자세한 설명은 제3장에서 확인할 수 있다.

$$\widetilde{Y}^{(t)} = \begin{cases} y_{ij}, & \text{if} \quad (i,j) \notin \Omega, \\ \widetilde{Y}_{ij}^{(t)}, & \text{if} \quad (i,j) \in \Omega, \end{cases} \qquad Y_1 = \begin{cases} y_{ij}, & \text{if} \quad (i,j) \notin \Omega, \\ \bar{Y}_{i\cdot}, & \text{if} \quad (i,j) \in \Omega. \end{cases}$$

$\widetilde{Y}$와 Y_1을 행렬로 표현하면 아래와 같이 나타낼 수 있으며, 식 (2.6)과 같이 목적식에 결측치를 대체하는 항을 추가하고 최적화 문제를 풀어 군집화를 진행한다.

$$\widetilde{Y}^{(t)} = \begin{array}{c} \\ \text{movie1} \\ \text{movie2} \\ \vdots \\ \text{movie}p \end{array} \begin{array}{c} \begin{array}{cccc} \text{user1} & \text{user2} & \cdots & \text{user}n \end{array} \\ \begin{pmatrix} y_{11} & \widetilde{Y}_{12}^{(t)} & \cdots & y_{1n} \\ \widetilde{Y}_{21}^{(t)} & y_{22} & \cdots & \widetilde{Y}_{2n}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{pmatrix} \end{array} \qquad Y_1 = \begin{array}{c} \\ \text{movie1} \\ \text{movie2} \\ \vdots \\ \text{movie}p \end{array} \begin{array}{c} \begin{array}{cccc} \text{user1} & \text{user2} & \cdots & \text{user}n \end{array} \\ \begin{pmatrix} y_{11} & \overline{Y}_{1\cdot} & \cdots & y_{1n} \\ \overline{Y}_{2\cdot} & y_{22} & \cdots & \overline{Y}_{2\cdot} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pn} \end{pmatrix} \end{array}$$

$$\left(\widehat{X}, \widehat{Y}\right) = \arg\min_{X,\widetilde{Y}} \frac{1}{2}\left\|\widetilde{Y} - ZX\right\|_F^2 + \gamma_1 \left\|\left(Y_1 - \widetilde{Y}\right)_\Omega\right\|_F^2 + \gamma_2 \sum_{i<j} w_{ij} \left\|X_i - X_j\right\|. \quad i = 1, 2, \ldots, n, \tag{2.6}$$

여기서 결측치 대체를 최적화 단계에서 동시에 진행하는 방법은 Park과 Zhao (2019)에서의 방법의 아이디어와 유사하다. 여기서, $\|(Y_1 - \widetilde{Y})_\Omega\|_F^2 = \sum_{(i,j)\in\Omega}((Y_1)_{ij} - \widetilde{Y}_{ij})^2$이며, $\sum_i \|(Y_1)_i - \widetilde{Y}_i\|^2$로 표현될 수 있다. 결측치 대체 범위는 $\gamma_1 > 0$을 통해 조정할 수 있으며, γ_1이 증가할수록 범위가 좁혀진다. 마찬가지로 $\gamma_2 > 0$를 통해 벌점항의 벌점 크기를 조정할 수 있으며, γ_2가 증가할수록 $\widehat{X}_i$의 열 벡터들은 서로 근접한 값을 가지게 된다. 여기서 $w_{ij} = w_{ji}$는 X_i와 X_j사이의 음이 아닌 가중치로 다음과 같이 설정한다.

$$w_{ij} = \iota_{i,j}^k \exp\left(-\phi \left\|X_i - X_j\right\|_2^2\right).$$

가중치 식의 첫 번째 요소인 지시 함수 $\iota_{i,j}^k$는 j와 i가 K-최근접 이웃(K-nearest neighbors)이면 1이 되고, 그렇지 않으면 0이 된다. 두 번째 요소는 가우스 커널 함수로 두 벡터 사이의 거리가 멀 경우의 결합을 늦춰주는 역할을 한다. ϕ는 음이 아닌 값을 가지며, $\phi = 0$ 일 경우는 균일한 가중치를 가지게 된다.

3. 알고리즘

본 장에서는 볼록 최적화(convex optimization)를 포함한 다양한 모델을 최적화 할 수 있는 일반적인 알고리즘인 alternating direction method of multipliers (ADMM) (Boyd 등, 2011)에 대해 소개한 후, 제안하는 볼록 최적화(convex optimization)를 풀고자 한다. ADMM은 두 단계의 원시(primal) 문제와 쌍대(dual) 문제를 교차 최적화로 해를 업데이트하는 과정으로 구성된다.

3.1. ADMM 알고리즘

ADMM은 그래디언트 상승(dual ascent)과 증강 라그랑지안(augmented Lagrangian) 방법의 장점을 가진 방법이다. 그래디언트 상승이란, 쌍대 문제의 목적식을 최대화하기 위해 시작점 $u^{(0)}$에서 시작하여 $t = 1, 2, 3, \ldots,$에 대해 2개의 단계를 반복하는 방법이다. 먼저 선형 제약 식을 가지는 다음의 원시 문제를 고려해 보자. 여기서 x와 u는 p차원의 열 벡터이다.

$$\min_x f(x) \quad \text{subject to} \quad Ax = b. \tag{3.1}$$

식(3.1)의 라그랑지안 함수(Lagrangian function)는 다음과 같다.

$$L(x, u) = f(x) + u^T(Ax - b).$$

여기서, u는 라그랑지안 승수(Lagrangian multipliers) 벡터이다. $t = 1, 2, 3, \ldots,$ 가 주어질 때 그래디언트 상승은 다음을 반복한다.

$$x^{(t)} = \arg\min_x L\left(x, u^{(t-1)}\right).$$
$$u^{(t)} = u^{(k-1)} + k_t\left(Ax^{(t)} - b\right).$$

여기서 k_t는 스텝 크기(step size)이다. 그래디언트 상승 방법은 그래디언트 수렴을 보장하기 위해 f가 엄격한 볼록(strongly convex)조건을 만족해야 하는 단점을 가지고 있다. 이런 단점은 증강 라그랑지안 방법에 의해 개선이 가능하다. 원시 문제 (3.1)을 다음과 같이 고려해보자.

$$\min_x f(x) + \frac{\rho}{2}\|Ax - b\|_2^2 \quad \text{subject to} \quad Ax = b. \tag{3.2}$$

여기서 $\rho > 0$는 스텝 크기 역할을 하며 $k_t = \rho$이다. 식 (3.2)의 수정된 라그랑지안 함수는 다음과 같다.

$$L_\rho(x, u) = f(x) + u^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2.$$

그래디언트 상승방법과 마찬가지로 $t = 1, 2, 3, \ldots,$ 가 주어질 때 다음을 반복한다.

$$x^{(t)} = \arg\min_x L_\rho\left(x, u^{(t-1)}\right).$$
$$u^{(t)} = u^{(t-1)} + \rho\left(Ax^{(t)} - b\right).$$

증강 라그랑지안 방법은 그래디언트 상승 방법보다 훨씬 좋은 수렴성을 가지고 있지만 문제를 분해할 수 있는 특징을 잃는다는 단점을 가지고 있다. ADMM은 수렴성과 함께 문제를 분해할 수 있는 특징을 잃지 않는 방법으로 먼저 다음의 식을 고려한다.

$$\min_x f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = c. \tag{3.3}$$

$\rho > 0$에 대해서, 식 (3.3)의 증강 라그랑지안 함수는 다음과 같다.

$$L_\rho(x, z, u) = f(x) + g(z) + u^T (Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|_2^2 .$$

$t = 1, 2, 3, \ldots,$ 에 대해서, ADMM은 다음의 단계를 반복적으로 수행한다.

$$\begin{aligned} x^{(t)} &= \arg\min_x L_\rho \left(x, z^{(t-1)}, u^{(t-1)}\right). \\ z^{(t)} &= \arg\min_z L_\rho \left(x^{(t)}, z, u^{(t-1)}\right). \\ u^{(t)} &= u^{(t-1)} + \rho \left(Ax^{(t)} + Bz^{(t)} - c\right). \end{aligned}$$

ADMM은 ρ값이 결과에 민감하고, 어떤 값으로 선택해야 하는지에 대해 정해진 바가 없다. 그러나 일반적으로 어려운 문제들을 ADMM 알고리즘을 통해 풀 수 있기 때문에 가장 많이 사용되는 알고리즘이다.

3.2. 제안 알고리즘

Chi와 Lange (2015)는 쌍별 규합벌점함수(pairwise fused penalty)를 활용한 다음의 볼록 군집 문제를 ADMM을 활용하여 최적화하는 방법을 제안하였다. 행렬 $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p \times n}$가 주어질 때, 다음의 U에 대한 식을 최소화하는 최적화를 고려한다.

$$F_\gamma(U) = \frac{1}{2} \sum_{i=1}^{n} \|X_i - U_i\|_2^2 + \gamma \sum_{i<j} w_{ij} \left\|U_i - U_j\right\|.$$

여기서 U의 i번째 열 벡터 U_i는 X_i에 연결된 군집 중심이다. 보조변수 벡터를 $U_i - U_j$로 두면 ADMM 알고리즘을 적용할 수 있다. 본 논문에서도 제안하는 볼록 최적화(convex optimization)를 ADMM 알고리즘을 통해 풀고자 한다. 먼저, 식 (2.6)은 다음의 식과 같이 구성할 수 있다.

$$\min_{X,\widetilde{Y}} \frac{1}{2} \sum_{i=1}^{n} \left\|\widetilde{Y}_i - ZX_i\right\|_2^2 + \gamma_1 \sum_{i=1}^{n} \left\|(Y_1)_i - \widetilde{Y}_i\right\|_2^2 + \gamma_2 \sum_{l \in \mathcal{E}} w_l \|V_l\| \quad \text{subject to} \quad X_{l_1} - X_{l_2} - V_l = \mathbf{0}. \tag{3.4}$$

여기서 $l = (l_1, l_2)$은 $l_1 < l_2$를 만족하는 인덱스 쌍으로 정의하고, 집합 $\mathcal{E} = \{l = (l_1, l_2) : w_l > \mathbf{0}\}$ 으로 정의한다. 원래의 목적 식 (2.6)을 보조변수 $V_l \in \mathbb{R}^k$을 사용하여 최적화 문제를 분할하였고, ε을 인덱스 쌍 $l = (l_1, l_2)$의 개수라고 정의하면, $V = (V_1, \ldots, V_\varepsilon) \in \mathbb{R}^{k \times \varepsilon}$로 정의된다. 계산의 편의를 위해 프로베니우스 노름을 유클리디안 노름(Euclidean norm)형태로 전환하였고, $\|(Y_1 - \widetilde{Y})_\Omega\|_F^2$를 $\sum_{i=1}^{n} \|(Y_1)_i - \widetilde{Y}_i\|_2^2$로 표현하였다. 식 (3.4)의 증강 라그랑지안 함수는 다음과 같이 정의된다.

$$\begin{aligned} L_\rho\left(\widetilde{Y}, X, V, \Lambda\right) = & \frac{1}{2} \sum_{i=1}^{n} \|Y_i - ZX_i\|_2^2 + \gamma_1 \sum_{i=1}^{n} \left\|(Y_1)_i - \widetilde{Y}_i\right\|_2^2 + \gamma_2 \sum_{l \in \mathcal{E}} w_l \|V_l\| \\ & + \sum_{l \in \mathcal{E}} < \lambda_l, V_l - X_{l_1} + X_{l_2} > + \frac{\rho}{2} \sum_{l \in \mathcal{E}} \left\|V_l - X_{l_1} + X_{l_2}\right\|_2^2. \end{aligned}$$

$\boldsymbol{\lambda}_l \in \mathbb{R}^k$은 라그랑지안 승수(Lagrangian multipliers)이고, $\Lambda = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_\varepsilon) \in \mathbb{R}^{k \times \varepsilon}$로 정의된다. 다음의 알고리즘을 통해 증강 라그랑지안 함수를 최소화하는 최적해를 구하였으며, 각 단계별 설명은 부록에서 확인할 수 있다.

Algorithm 1: Finding the optimal solution to minimize the augmented Lagrangian function using alternating direction method of multipliers (ADMM)

Require: $Y, Z, \rho, \gamma_1, \gamma_2, \epsilon_{tol}$

1: **Initialized** $\widetilde{Y}^{(0)} = Y_1, X^{(0)} = (X_1^{(0)}, \ldots, X_n^{(0)})$where$X_i^{(0)} = (Z^T Z + \epsilon \mathbf{I}_k)^{-1} \mathbf{Z}^T \widetilde{Y}_i^{(0)}$, $V^{(0)} = V^{(0)}, \Lambda^{(0)} = \Lambda^{(0)}$.

2: **Repeat**

- Update $\widetilde{Y}^{(t+1)}$ by solving $\arg\min_{\widetilde{Y}} L_\rho(\widetilde{Y}, X^{(t)}, V^{(t)}, \Lambda^{(t)})$
- Update $X^{(t+1)}$ by solving $\arg\min_X L_\rho(\widetilde{Y}^{(t+1)}, X, V^{(t)}, \Lambda^{(t)})$
- Update $V_l^{(t+1)}$ by solving $\arg\min_V L_\rho(\widetilde{Y}^{(t+1)}, X^{(t+1)}, V, \Lambda^{(t)})$
- Update $\lambda_l^{(t+1)}$ with $\lambda_l^{(t)} + \rho(V_l^{(t+1)} - X_{l_1}^{(t+1)} + X_{l_2}^{(t+1)})$

Until convergence

3.3. 수렴 기준

t번째 반복문에서 $\widetilde{Y}^{(t)} \in \mathbb{R}^{p \times n}$, $X^{(t)} \in \mathbb{R}^{k \times n}$, $V_l^{(t)} \in \mathbb{R}^k$, $\boldsymbol{\lambda}_l^{(t)} \in \mathbb{R}^k$는 각 반복문에서 업데이트된다. 알고리즘을 멈추기 위해서는 각각의 결과가 일정 값으로 수렴해야 한다. t번째와 $t-1$의 예측 값이 거의 차이가 없을 때, t번째 예측 값을 선택하기 위해 수렴 기준을 다음과 같이 설정한다.

$$\left\|\widetilde{Y}^{(t)} - \widetilde{Y}^{(t-1)}\right\|_F^2 + \left\|X^{(t)} - X^{(t-1)}\right\|_F^2 + \left\|V^{(t)} - V^{(t-1)}\right\|_F^2 + \left\|\Lambda^{(t)} - \Lambda^{(t-1)}\right\|_F^2 \leq \epsilon.$$

알고리즘에 사용한 수렴 기준값은 $\epsilon = 10^{-5}$이고, 수렴 기준을 충족하지 않을 경우 20번째 반복에서 알고리즘이 멈추도록 설정한다.

3.4. 군집 개수의 추정

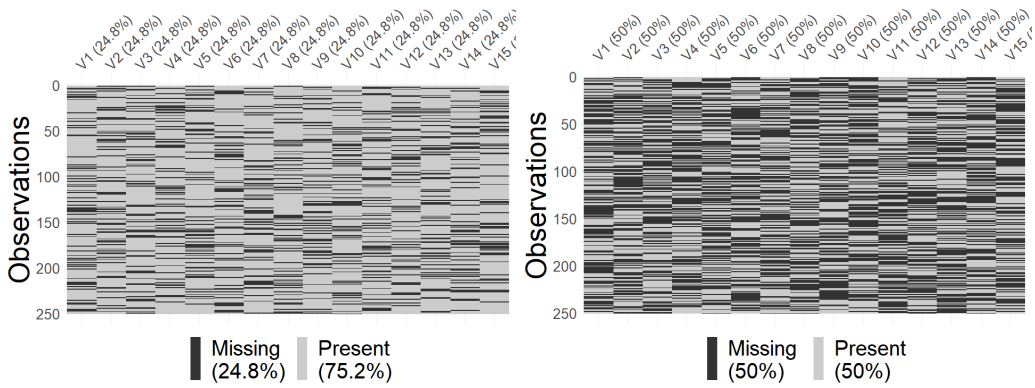
군집의 개수를 추정하기 위해 Park 등 (2021)에서 제안한 방법을 이용하였다. 본 논문에서 사용되는 실데이터는 실제 label이 없기 때문에, 원 데이터에 노이즈를 추가한 데이터를 생성해서 원 데이터와의 군집화 결과 비교를 통해 군집 수를 선택하는 데 사용한다. 군집화의 비교로는 Kvalseth (1987)의 normalized mutual information (NMI)를 사용하였다. 노이즈가 추가된 데이터는 $T = Y + E$, $E_{ij} \sim N(0, \mathrm{var}(Y))$로 정의한다. 여기서 $Y \in \mathbb{R}^{p \times n}$는 영화 평점 데이터이다. 원 데이터 Y와 노이즈가 추가된 데이터 T를 통해 식 (2.6)에 대한 최적해인 $\widehat{X}, \widehat{X}_T$를 구하고 각각 군집 수 c를 적용하여 비교한 NMI 값 중 가장 큰 값을 가지는 군집의 개수를 선택한다.

$$\widetilde{C} = \arg\max_{c \geq 2} \mathrm{NMI}\left(\widehat{X}(c), \widehat{X}_T(c)\right).$$

여기서 $\widehat{X}(c)$와 $\widehat{X_T}(c)$는 c개의 군집 수를 이용한 군집화 결과 label을 의미한다. 자세한 추정 방법은 Park 등 (2021)에서 확인할 수 있다.

4. 시뮬레이션 및 실데이터 적용

본 장에서는 제안하는 알고리즘의 모수를 선택하고 기존 군집화 기법과 제안하는 군집화 기법의 성능을 비교한다. 군집화의 성능 평가 척도로는 NMI (Kvalseth, 1987), purity, ARI (Hubert와 Arabie, 1985)를 사용한다.



(a) missing rate : 25% (b) missing rate : 50%

Figure 4: *Graph for the MCAR pattern according to each missing rate.*

4.1. 시뮬레이션

시뮬레이션 연구에서는 높은 결측 비율을 갖는 영화 평점 데이터의 특성을 반영한 가상 데이터를 만들어 진행하였다. Park과 Zhao (2018)의 과정을 참조한 희소 가우시안 혼합 모델(sparse Gaussian mixture model)을 기반으로 하는 시뮬레이션 모델을 사용하였으며, 다음 네 가지 단계로 진행하여 생성하였다.

1. *i.i.d*를 만족하는 표준 가우시안 랜덤 행렬의 좌특이행렬(left singular matrix)로 $\widetilde{B} \in \mathbb{R}^{C\times q}$의 요소를 생성한다. $B \in \mathbb{R}^{C\times k}$를 $B = [\sigma\widetilde{B}, 0_{C\times(k-q)}]$로 얻는다. 여기서 C는 군집 개수를 의미한다.
2. 하나의 그룹에 무작위로 할당하여 $i = 1, 2, \ldots, n$ 번째 샘플의 군집 레이블 $m_i \in \{1, \ldots, C\}$ 를 생성한 후, $m_{ij} = 1(m_i = j)$ 를 사용하여 멤버십 행렬(membership matrix)인 $M \in \mathbb{R}^{n\times C}$를 생성한다. n개의 샘플을 C개의 그룹 중 하나로 무작위로 할당하는 경우, 멤버십 행렬은 아래의 행렬처럼 표현된다.

$$M = \begin{array}{c} \\ \text{sample1} \\ \text{sample2} \\ \vdots \\ \text{sample}n \end{array} \begin{array}{c} \begin{array}{cccc} \text{cluster1} & \text{cluster2} & \cdots & \text{cluster}C \end{array} \\ \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

3. 표준 가우시안 노이즈 행렬(standard Gaussian noise matrix)인 W 를 추가한 데이터 행렬 $X = MB + W, W \in \mathbb{R}^{n\times k}$를 생성한 후, 식 (2.5)의 가정을 만족하는 $Y = ZX^T + W, W \in \mathbb{R}^{p\times n}$ 를 생성한다. X와 Y에 사용된 노이즈 W는 표준 정규 분포(standard normal distribution)를 따른다. Z는 실데이터에 사용되는 영화별 장르 가중치 행렬이다.
4. 마지막으로 데이터 Y 의 결측치가 임의의 확률로 완전한 무작위 패턴(missing completely at random, MCAR)을 갖도록 정의한다. MCAR은 결측치와 데이터 사이에 어떠한 상관관계도 없는 경우를 뜻한다.

1단계와 2단계는 희소 가우시안 혼합 모델(sparse Gaussian mixture model)을 기반으로 하였고, σ값은 신호 대 잡음비(signal to noise ratio)를 제어한다. 다른 사용자들의 유형을 구별하기 위한 정보를 제공하는 장르의 개수는 희소하고, 노이즈가 많은 장르는 사용자 식별의 어려움을 증가시킬 수 있다. 본 논문에서는 $n = 500, q =$

Table 1: RWCSS results

Missing rate	γ_1	γ_2	
		$10^{2.24}$	$10^{3.14}$
25%	$10^{1.79}$	0.1932	0.2662
	$10^{2.24}$	0.1915	0.2724
50%	$10^{1.79}$	0.2407	0.2359
	$10^{2.24}$	0.2554	0.2323
Average RWCSS for each parameter pair (γ_1, γ_2)			
$(10^{1.79}, 10^{2.24})$	$(10^{1.79}, 10^{3.14})$	$(10^{2.24}, 10^{2.24})$	$(10^{2.24}, 10^{3.14})$
0.2169	0.2510	0.2234	0.2523

Table 2: Results of imputation of missing values by parameter pair

Missing rate	$(10^{1.79}, 10^{2.24})$	$(10^{1.79}, 10^{3.14})$	$(10^{2.24}, 10^{2.24})$	$(10^{2.24}, 10^{3.14})$
25%	0.00943	0.00941	0.00941	0.00942
50%	0.00664	0.00664	0.00665	0.00665

$7, C = 5, \sigma = 5$를 사용하였고, 결측 비율은 각각 25%, 50%로 고려하였다. 시뮬레이션에 사용될 데이터의 결측치는 Figure 4에서 확인할 수 있듯이, 완전한 무작위 패턴을 가진다. 제안하는 알고리즘의 모수를 결정하기 위해 시뮬레이션에서 ratio of within-cluster sum of squares (RWCSS)를 사용하였고, 가장 작은 RWCSS 값을 제공하는 γ_1, γ_2를 선택하였다.

$$\text{RWCSS} := \frac{\sum_{c=1}^{C} \sum_{j=1}^{n_c} \left\| X_j^{(c)} - \overline{X}^{(c)} \right\|^2}{\sum_{j=1}^{n} \left\| X_j - \overline{X} \right\|^2}.$$

시뮬레이션 데이터를 $X \in \mathbb{R}^{k \times n}$라 할 때, n_c는 c번째 군집에 포함되는 사용자의 수를 나타내며 $\sum_{c=1}^{C} n_c = n$을 만족한다. X_j와 $\bar{X}$는 j번째 사용자 데이터와 전체 사용자의 평균을 나타내고 각각 k차원 벡터이다. 그리고 $X_j^{(c)}$와 $\bar{X}^{(c)}$는 c번째 군집에 속해있는 j번째 사용자 데이터와 c번째 군집에 속해있는 사용자의 평균을 의미한다. 위에 정의된 RWCSS 값은 군집 내에 있는 데이터의 제곱합(within-cluster sum of squares)의 총합을 전체 데이터의 제곱합 (total sum of squares)으로 나눠준 값이며 항상 0과 1 사이의 값을 가진다. 작은 RWCSS 결과값은 해당 군집화 결과가 같은 군집에 있는 데이터의 동질성과 다른 군집에 있는 데이터의 이질성을 잘 설명한다는 것을 의미한다. 본 시뮬레이션에 사용한 모수는 $\gamma_1 = \{10^{1.79}, 10^{2.24}\}$와 $\gamma_2 = \{10^{2.24}, 10^{3.14}\}$이다. 시뮬레이션 분석 결과는 Tabel 1에서 확인할 수 있으며, 각 모수쌍의 조합에 대한 RWCSS 값은 100번 반복한 결과의 평균값이다. 결측 비율이 25%일 때 구한 RWCSS 값과 50%일 때 구한 RWCSS 값의 평균이 가장 작은 모수 쌍은 $\gamma_1 = 10^{1.79}, \gamma_2 = 10^{2.24}$인 것을 확인할 수 있다.

Table 2는 모수 쌍별로 $\widetilde{Y}^{(t)}$의 결측치 대체 값이 Y의 실제값과 얼마나 유사한지를 확인하기 위해 아래의 식을 통해 오차 값을 측정한 결과이다.

$$\frac{1}{n_\Omega} \sum_{(i,j) \in \Omega} \left(y_{ij} - \hat{y}_{ij} \right)^2.$$

여기서 n_Ω는 결측치의 개수이다. 결측 비율이 25%일 때 모수 쌍별로 모두 오차 값이 약 0.0094이고, 50%일 때는 약 0.0066이므로 모수 쌍별로 실제 값과 결측치 대체 값의 차이에 대한 오차값이 작은 것을 확인할 수 있다. 모수 쌍은 최종적으로 RWCSS결과를 바탕으로 $\gamma_1 = 10^{1.79}, \gamma_2 = 10^{2.24}$를 선택하였다.

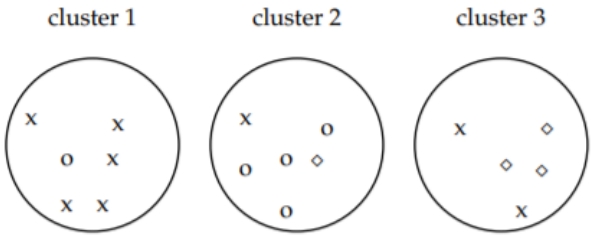


Figure 5: *Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5(cluster1); o,4(cluster2); and ◇,3(cluster3). Purity is* (1/17) ×(5+4+3)≈ 0.71 *(Christopher et al., 2008).*

Table 3: Contingency table

	Y_1	$\cdots$	Y_s	sums
X_1	n_{11}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	$\cdots$	n_{rs}	a_r
sums	b_1	$\cdots$	b_s	

4.2. 군집 성능 평가 척도

군집 성능 평가 척도로 NMI, purity, ARI가 사용되었다.

NMI와 purity, ARI는 0과 1 사이의 값을 가지며, 실제 군집 label이 존재할 경우 1에 가까울수록 군집화가 잘 되었다는 것을 의미한다. purity는 제대로 분류된 샘플의 비율이며, Figure 5에서 purity 예시를 확인할 수 있다.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_k \max_j \left|w_k \cap c_j\right|.$$

여기서 $\Omega = \{w_1, \ldots, w_K\}$는 군집 집합이고, $\mathbb{C} = \{c_1, \ldots, c_J\}$는 클래스(class)의 집합이다. K개의 군집별로 각 군집 내에 속한 클래스 중 가장 많은 빈도 수를 가진 클래스의 수를 더한 후 전체 관측 값 개수 N으로 나눈 값으로 측정된다. Adjusted rand index (ARI)는 Table 3과 같은 분할표(Contingency table)가 있을 때 아래의 식으로 계산된다.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}.$$

$X = \{X_1, X_2, \ldots, X_r\}$와 $Y = \{Y_1, Y_2, \ldots, Y_s\}$는 각각 군집 집합을 의미하며 n_{ij}는 X_i와 Y_j의 공통된 객체 수를 의미한다.

Table 4는 시뮬레이션 분석을 통해 선택된 모수쌍 $\gamma_1 = 10^{1.79}, \gamma_2 = 10^{2.24}$으로 제안하는 방법과 기존 군집 방법인 K-means, 계층적 군집화(hierarchical clustering)방법의 NMI, Purity, ARI 값을 비교한 결과이다. 시뮬레이션에서 사용한 군집 수는 $C = 5$로 실제 군집 label과 각 군집화 방법을 통해 구한 label의 NMI, purity, ARI 값을 비교하였다. 각 값들은 100번 반복한 결과의 평균값이다. 시뮬레이션 모델은 실제 label이 존재 하므로 1에 가까울 수록 군집화가 잘 되었다고 해석할 수 있다. 결측 비율이 각각 25%, 50%일 때 제안하는 방법이

Table 4: NMI, purity, ARI results

Missing Rate		NMI	purity	ARI
25%	new method	0.3583	0.5138	0.2883
	hierarchical	0.0455	0.2875	0.0248
	K-means	0.0274	0.2646	0.0145
50%	new method	0.2608	0.4435	0.2093
	hierarchical	0.0411	0.2862	0.0213
	K-means	0.0393	0.2590	0.0258

기존 군집 방법에 비해 NMI, purity, ARI 값이 상대적으로 큰 값을 가지는 것을 확인할 수 있다. 실데이터 분석에서는 실제 군집 label이 존재하지 않기 때문에 NMI, purity, ARI 값을 노이즈 및 이상치에 얼마나 민감한지 확인하는 척도로 사용한다. 각 값이 1에 가까울수록 상대적으로 오차에 민감하지 않다고 해석한다.

4.3. 실데이터 적용

본 장에서는 시뮬레이션에서 선택된 모수들을 실데이터에 적용하여 제안하는 군집화 결과를 해석하고, 기존 군집화 방법들과 성능을 비교한다. 데이터는 2.2장에서 설명한 MovieLens에서 제공하는 데이터로, 영화 평점 데이터 $Y \in \mathbb{R}^{p\times n}$와 영화 장르 가중치 데이터 $Z \in \mathbb{R}^{p\times k}$를 사용한다. 실데이터 분석에서는 영화 평점 정보를 가장 많이 가지고 있는 상위 500명의 사용자를 기준으로 50%의 결측 비율을 가지는 250개의 영화를 선택하였다. 250개의 영화 중에서 Documentary장르가 포함된 영화가 존재하지 않으므로 실데이터 분석에서는 해당 장르는 제거하고 총 18개의 장르를 사용하였다. γ_1은 Y의 결측치 대체 범위에 영향을 주는데, Figure 6에서 볼 수 있듯이 $\gamma_1 = \{0, 0.25, 0.5, 0.75, 1\}$일 때 결측치 대체 값의 범위가 γ_1값이 증가할수록 감소하는 것을 확인할 수 있다. 영화 평점 데이터는 0.5점과 5점 사이의 값을 가지므로 결측치도 동일한 범위 내에서 대체되어야 한다. 따라서, 최소값과 최대값이 각각 0.5와 5점에 가까운 적절한 γ_1을 선택해 주어야 한다. 마찬가지로 γ_2는 군집화에 영향을 주는 값으로, 값이 증가할수록 군집이 합쳐지게 된다. Figure 7은 $\gamma_2 = \{0, 10^{-1}, 10^{2.24}, 10^{2.35}\}$일 때의 heatmap 결과로, γ_2가 증가할수록 하나의 군집으로 합쳐지는 것을 확인할 수 있다. 실데이터 분석에서는 시뮬레이션 분석에서 선택된 모수를 사용하였고, 군집 개수는 3.4장에서 설명한 방법으로 추정하였다. Figure 8은 군집 수별로 100번 반복한 NMI 결과를 오차막대(error bar)로 표현한 그래프이며, 그래프에서 확인할 수 있듯이 군집의 수가 11개일 때 가장 큰 평균 NMI 값을 가지므로 군집의 수를 11개로 추정하였다. 다음으로 제안하는 군집 방법과 기존 군집 방법의 노이즈 및 이상치에 민감한 정도를 확인하기 위해 다음 세 단계로 진행하여 각 군집 방법별 NMI, purity, ARI 값을 구하고 비교하였다.

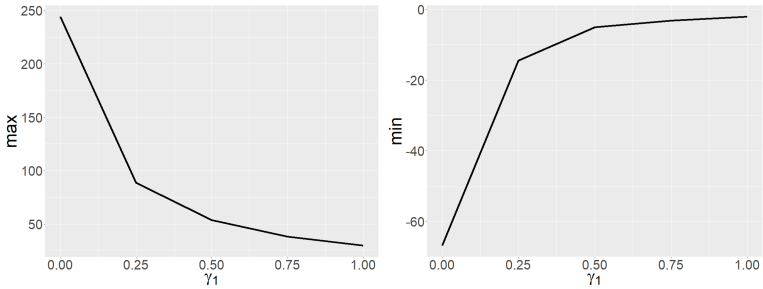
1. 결측치가 존재하는 영화 평점 데이터 Y를 사용하여 노이즈를 추가한 데이터 $T = (T_1, \ldots, T_n)$를 생성한다.

$$T = Y + E \quad E_{ij} \sim N(0, \mathrm{var}(Y)).$$

2. 다음의 식을 만족하는 $X^{(0)}$을 통해 $X^{(0)}$과 $X_T^{(0)}$을 생성한다.

$$x_i^{(0)} = \left(Z^{TZ} + \epsilon I\right)^{-1} Z^T \widetilde{Y}_i^{(0)}, \quad i = 1, \ldots, n.$$
$$x_{Ti}^{(0)} = \left(Z^{TZ} + \epsilon I\right)^{-1} Z^T \widetilde{T}_i^{(0)}, \quad i = 1, \ldots, n.$$

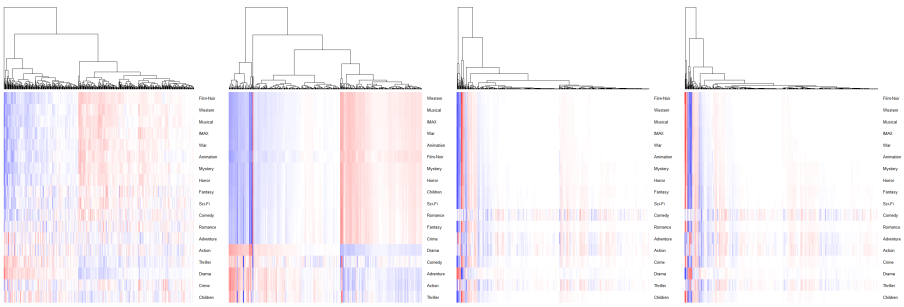
여기서 $\widetilde{T} = (\widetilde{T}_1, \ldots, \widetilde{T}_n)$는 다음의 식을 만족하며, $\widetilde{T}_{ij}^{(0)}$은 T의 i번째 행에서 결측치를 제외한 평균으로 정의



(a) The maximum value of the missing value replacement.

(b) The minimum value of the missing value replacement.

Figure 6: *Line plot for replacing missing values in Y with a smaller range of values as γ_1 increases.*



(a) $\gamma_2 = 0$ (b) $\gamma_2 = 10^{-1}$ (c) $\gamma_2 = 10^{2.24}$ (d) $\gamma_2 = 10^{2.35}$

Figure 7: *Heatmap graph for changing as γ_2 increases. x-axis and y-axis indicates 500 users and 18 movie genres, respectively.*

한다.

$$\widetilde{T} = \begin{cases} t_{ij}, & \text{if} \quad (i,j) \notin \Omega \\ \\ \widetilde{T}^{(0)}_{ij}, & \text{if} \quad (i,j) \in \Omega \end{cases}$$

3. 군집 수 $k = 2, \ldots, 20$ 를 적용하여 $X^{(0)}(k)$과 $X^{(0)}_T(k)$의 NMI, purity, ARI를 구한다.

제안하는 군집 방법은 $X^{(0)}$와 $X^{(0)}_T$가 아닌 $X^{(t)}$와 $X^{(t)}_T$를 사용하여 위의 세 단계와 동일한 방법으로 구하였으며, NMI, purity, ARI 값은 군집 수별로 100번 반복한 결과의 평균값으로 구하였다. Figure 9는 NMI 값으로 군집 결과를 비교한 그래프로 제안하는 방법의 평균 NMI와 군집 수별 NMI 값이 기존 군집화 방법에 비해 상대적으로 높은 값을 가지는 것을 확인할 수 있다. 이는 제안된 방법이 기존 군집화 방법에 비해 상대적으로 노이즈 및 이상치에 민감하지 않다는 것을 의미한다. purity, ARI 결과는 Figure 10, Figure 11에서 확인할 수 있으며, 마찬가지로 제안하는 방법이 기존 방법보다 높은 값을 가지는 것을 확인할 수 있다. 다음으로 가중치 분석을 통해 변수의 영향력을 확인한다. 실데이터에서 중요한 장르일수록 군집의 성격을 잘 반영한다. 같은 군집 안에 있는 사용자끼리 유사한 값을, 다른 군집 안에 있는 사용자와 구별되는 값을 지닐수록 사용자 간의 이질성에 영향력 있는 장르로 분류할 수 있다. 이런 점을 이용해 중요한 장르 변수를 확인할 수 있는 측도를 식 (4.1)과 같이 정의한다.

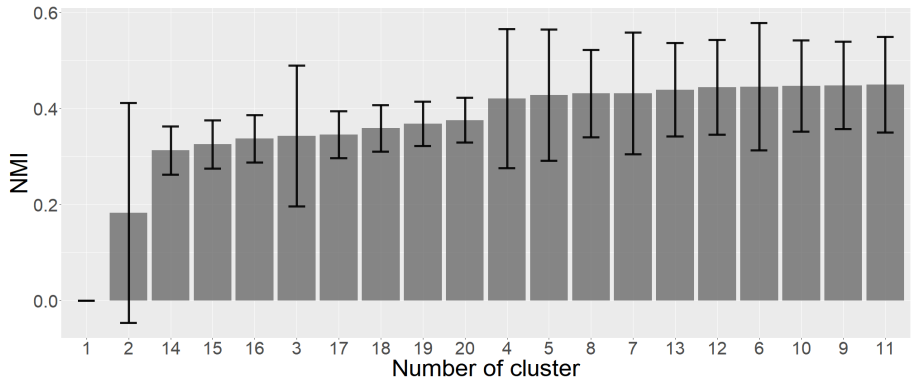
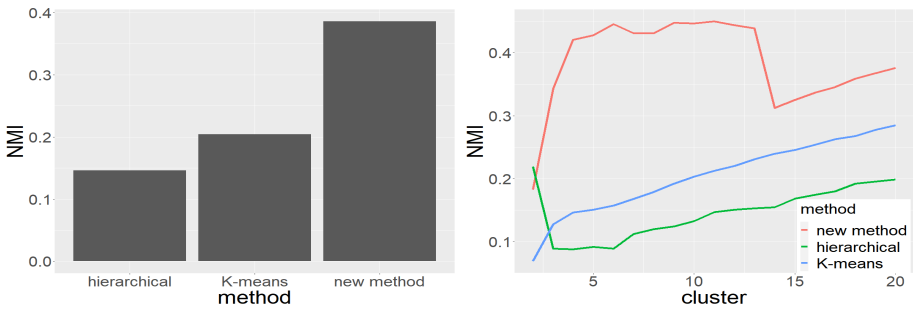
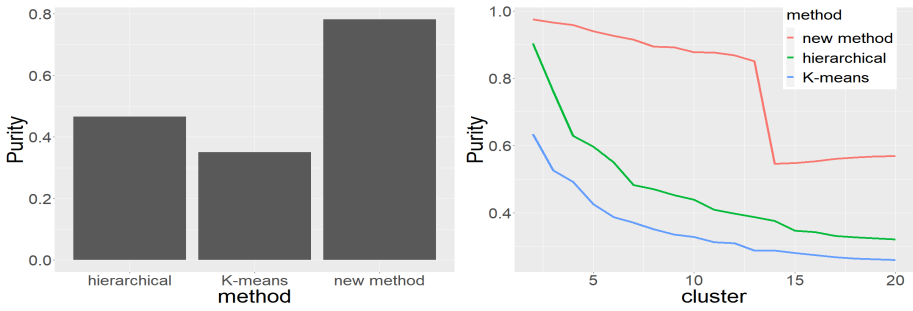


Figure 8: *Choosing the optimal number of cluster.*



(a) Mean of NMI by cluster method (b) Line plot of NMI by clustering method

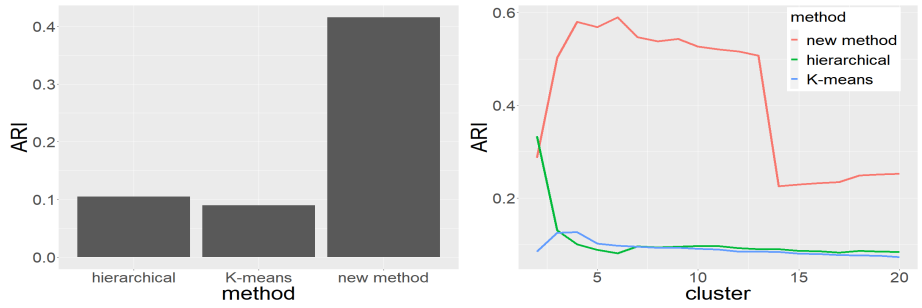
Figure 9: *NMI Comparison with other clustering methods. NMI=normalized mutual information.*



(a) Mean of purity by cluster method (b) Line plot of purity by clustering method

Figure 10: *purity Comparison with other clustering methods.*

여기서 i라는 임의의 장르에 대해서 x_{ij}를 i번째 장르의 j번째 사용자 값, $\bar{X}_{i\cdot} = (1/n)\sum_{j=1}^{n} x_{ij}$를 i번째 장르의 전체 사용자의 평균값, $x_{ij}^{(c)}$를 c번째 군집에서 $1 \le j \le n_c$번째 사용자의 장르 i값으로, $\bar{X}_{i\cdot}^{(c)} = (1/n_c)\sum_{j=1}^{n_c} x_{ij}^{(c)}$를 i번째 장르의 c번째 군집 내의 사용자 평균값으로 정의한다. 군집 내 제곱합과 전체 제곱합을 이용해 다음과 같이 i번째 장르의 RWCSS 값 (RWCSS-i)을 정의하여 장르 i의 중요도를 측정한다. 같은 군집 내에 있는 사용자들이 유사한 장르 정보를 가질수록 군집 내 제곱합은 작아지고, 군집 간 거리가 멀어질수록 전체 제곱합이



(a) Mean of ARI by cluster method (b) Line plot of ARI by clustering method

Figure 11: *ARI Comparison with other clustering methods. ARI=adjusted rand index.*

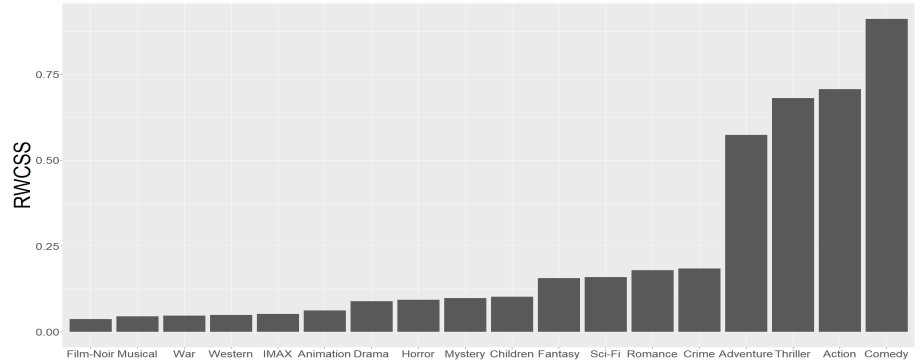
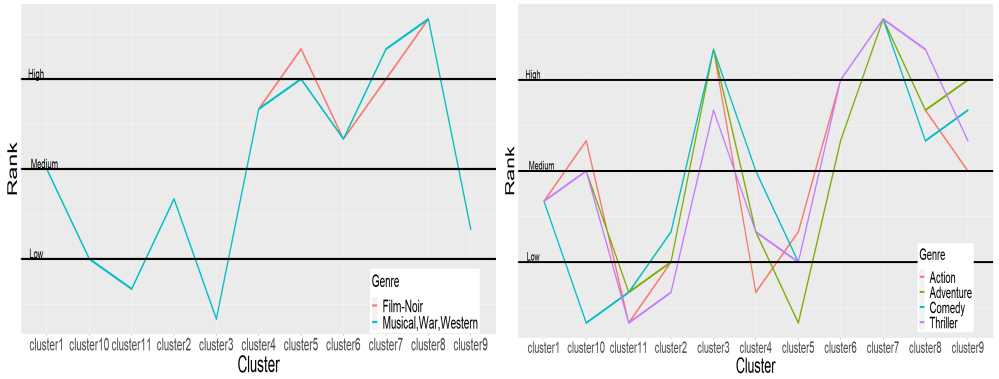


Figure 12: *Graph for ratio of within-cluster sum of squares for i^{th} genres.*



(a) Line plot of genres with low RWCSS-i (b) Line plot of genres with high RWCSS-i

Figure 13: *Graph of clusters distinguished by important genre variables.*

군집 내 제곱합보다 큰 경향을 보인다. RWCSS-i는 항상 0과 1 사이의 값을 가지고, RWCSS-i가 작을수록 장르 i가 군집화에 중요한 정보를 지닌다고 해석할 수 있다.

실데이터에 대한 RWCSS-i를 Figure 12에서 살펴보면, 18개의 장르 중 Film-Noir, Musical, War, Western 등의 장르들이 추정된 장르 선호도를 군집화 할 때 가장 영향력 있는 변수라고 해석할 수 있다. 이는 특정 사용자 군집에서만 선호되는 장르로 해석할 수 있다. 마찬가지로, Adventure, Thriller, Action, Comedy는 장르 군집화를 진행할 때 영향을 크게 주지 않는 변수라고 해석할 수 있다.

$$\text{RWCSS} - i := \frac{\sum_{c=1}^{C}\sum_{j=1}^{n_c}\left(x_{ij}^{(c)} - \overline{X}_{i\cdot}^{(c)}\right)^2}{\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_{i\cdot}\right)^2}. \tag{4.1}$$

Figure 13의 (a)는 군집별 Film-Noir, Musical, War, Western 장르의 평균 선호 점수를 사용하여 군집별로 상대적인 순위를 매긴 그래프이고, (b)는 RWCSS-i값이 큰 Adventure, Thriller, Action, Comedy 장르를 사용하여 동일한 방법으로 구한 그래프이다. (a) 그래프를 살펴보면, 4개의 장르는 군집별로 비슷한 순위를 가지고 있으며, 11개의 군집 중에서 여덟 번째 군집이 Film-Noir, Musical,War, Western 장르를 가장 많이 선호하고, 세 번째 군집은 상대적으로 낮은 선호 점수를 가지는 것을 알 수 있다. (b)의 경우 Film-Noir, Musical, War, Western 장르에 비해 경향이 비슷하지 않으므로 RWCSS-i의 결과처럼 군집을 구별 짓는 데 영향력이 상대적으로 적은 변수들로 해석할 수 있다.

5. 결론

본 논문에서는 MovieLens 데이터의 특성을 반영한 볼록 군집화(convex clustering) 기반의 방법을 제안하였고, 반복 알고리즘인 ADMM을 통해 제안하는 최적화 문제를 풀어 진행하였다. MovieLens 데이터는 결측 비율이 높은 데이터이므로 결측치 대체와 군집화를 동시에 진행하여 기존의 군집화 분석과 차별화를 두었다. 본 논문에서는 시뮬레이션을 통해 평균 RWCSS가 가장 작은 적절한 모수쌍 γ_1, γ_2를 찾고, 이를 실데이터 분석에 사용하였다. 실데이터 분석에서는 원 데이터에 노이즈를 추가한 데이터를 생성하여 원 데이터와의 군집화 결과를 통해 적절한 군집 개수를 선택하였다. 또한, 제안하는 방법과 기존 군집화 방법의 군집 성능을 비교하기 위해 NMI, purity, 그리고 ARI 의 평균값을 구하였으며, 제안하는 방법이 상대적으로 높은 값을 보여 기존 방법보다 노이즈나 이상치에 상대적으로 민감하지 않다는 것을 확인할 수 있었다. 마지막으로 가중치 분석을 통해 군집화에 중요한 장르를 확인하였다. 본 논문에서는 MovieLens 데이터인 영화 평점 데이터와 영화 장르 데이터를 통해 사용자별 장르 선호도를 추정하고 군집화하여 새로운 정보를 얻는 데 의의를 둔다. 새로운 사용자가 부여한 평점을 통해 추정한 고유한 선호 장르 점수가 특정 선호도 군집에 속하게 될 경우, 해당 군집이 가지고 있는 장르 정보를 이용하여 장르 가중치가 적절하게 부여된 영화들을 추천하는 연구는 향후 연구과제로 제시한다. 또한 다중대체(multiple imputation) 방법을 군집화 최적화 식에 동시에 고려하는 방법은 추후 연구에서 고려해 볼 수 있는 중요한 주제이다.

Appendix A: $\widetilde{Y}$ step

반복 과정의 첫 번째 단계로 $\widetilde{Y}$를 업데이트하기 위해서는 다음의 식을 풀어야 한다.

$$\widetilde{Y}^{(t+1)} = \arg\min_{\widetilde{Y}} L_\rho\left(\widetilde{Y}, X^{(t)}, V^{(t)}, \Lambda^{(t)}\right). \tag{A.1}$$

$\widetilde{Y}^{(t+1)}$은 다음의 값으로 업데이트된다.

$$\widetilde{Y}^{(t+1)} = \frac{1}{1+2\gamma_1}\left(ZX^{(t)} + 2\gamma_1 Y_1\right). \tag{A.2}$$

Appendix B: X step

반복 과정의 두 번째 단계는 다음의 식을 풀어야 한다.

$$X^{(t+1)} = \arg\min_{X} L_\rho\left(\widetilde{Y}^{(t+1)}, X, V^{(t)}, \Lambda^{(t)}\right). \tag{B.1}$$

X를 업데이트하기 위해서는 아래의 $f(X)$를 최소화하는 해를 구해야 한다.

$$\begin{aligned} f(X) &= \frac{1}{2}\sum_{i=1}^{n}\left\|\widetilde{Y}_i - ZX_i\right\|_2^2 + \sum_{l\in\mathcal{E}} < \boldsymbol{\lambda}_l, V_l - X_{l_1} + X_{l_2} > + \frac{\rho}{2}\sum_{l\in\mathcal{E}}\left\|V_l - X_{l_1} + X_{l_2}\right\|_2^2. \\ &= \frac{1}{2}\sum_{i=1}^{n}\left\|\widetilde{Y}_i - ZX_i\right\|_2^2 + \frac{\rho}{2}\sum_{l\in\mathcal{E}}\left\|\tilde{V}_l - X_{l_1} + X_{l_2}\right\|_2^2. \\ &= \frac{1}{2}\left\|\mathrm{vec}\left(\widetilde{Y}\right) - \mathrm{vec}\left(ZX\right)\right\|_2^2 + \frac{\rho}{2}\left\|\mathrm{A}\mathrm{vec}X - \mathrm{vec}\left(\tilde{V}\right)\right\|_2^2. \end{aligned} \tag{B.2}$$

여기서 $\widetilde{V}_l = V_l + \rho^{-1}\boldsymbol{\lambda}_l \in \mathbb{R}^k$이고, $\widetilde{V} = (\widetilde{V}_1, \ldots, \widetilde{V}_\varepsilon) \in \mathbb{R}^{k\times\varepsilon}$이다. 또한, $A^T = (A_1^T, \ldots, A_\varepsilon^T) \in \mathbb{R}^{nk\times\varepsilon k}$이고, $A_l^T = (e_{l_1} - e_{l_2})\otimes I_k \in \mathbb{R}^{nk\times k}$으로 정의된다. 여기서, $e_{l_1}, e_{l_2} \in \mathbb{R}^n$는 각각 l_1번째, l_2번째 원소가 1인 단위벡터(unit vector)를 의미한다. $\otimes$는 크로네커 곱(kronecker product)이다. $X^{(t+1)}$의 값은 아래와 같이 업데이트된다.

$$\mathrm{vec}(X)^{(t+1)} = \left(\rho A^T A + \left[I_n \otimes Z^T Z\right]\right)^{-1}\left(\mathrm{vec}\left(Z^T\widetilde{Y}^{(t+1)}\right) + \rho A^T \mathrm{vec}\left(\widetilde{V}^{(t)}\right)\right).$$

Appendix C: V step

반복 과정의 세 번째 단계로 V를 업데이트하기 위해서는 다음의 식을 풀어야 한다.

$$V^{(t+1)} = \arg\min_{V} L_\rho\left(\widetilde{Y}^{(t+1)}, X^{(t+1)}, V, \Lambda^{(t)}\right). \tag{C.1}$$

해당 단계에서는 V의 열 벡터 V_l를 차례대로 업데이트하여 $t+1$번째 V인 $V^{(t+1)} = (V_1^{(t+1)}, \ldots, V_\varepsilon^{(t+1)})$를 구한다. KKT 조건(Karush–Kuhn–Tucker condition ; Kuhn 과 Tucker, 2014)에 의해 V_l는 다음과 같이 업데이트된다.

$$V_l^{(t+1)} = \begin{cases} \left(\frac{\gamma_2 w_l}{\left\|\rho\left(X_{l_1}^{(t+1)} - X_{l_2}^{(t+1)}\right) - \lambda_l^{(t)}\right\|} + \rho\right)^{-1}\left(\rho\left(X_{l_1}^{(t+1)} - X_{l_2}^{(t+1)}\right) - \lambda_l^{(t)}\right), & \text{if} \quad \left\|\rho\left(X_{l_1}^{(t+1)} - X_{l_2}^{(t+1)}\right) - \lambda_l^{(t)}\right\| \geq \gamma_2 w_l. \\ \mathbf{0}, \quad \text{if} \quad \left\|\rho\left(X_{l_1}^{(t+1)} - X_{l_2}^{(t+1)}\right) - \lambda_l^{(t)}\right\| < \gamma_2 w_l. \end{cases} \tag{C.2}$$

Appendix D: Λ step

반복 과정의 마지막 단계로 Λ를 업데이트하기 위해서는 다음의 식을 풀어야 한다. Λ의 열 벡터 $\boldsymbol{\lambda}_l$를 업데이트하여 $\Lambda^{(t+1)} = (\boldsymbol{\lambda}_1^{(t+1)}, \ldots, \boldsymbol{\lambda}_\varepsilon^{(t+1)})$를 구한다.

$$\boldsymbol{\lambda}_l^{(t+1)} = \boldsymbol{\lambda}_l^{(t)} + \rho\left(V_l^{(t+1)} - X_{l_1}^{(t+1)} + X_{l_2}^{(t+1)}\right). \tag{D.1}$$

References

Boyd S, Parikh N, and Chu E (2011). *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*, Now Publishers Inc, **3**, 1—122.

Chi EC and Lange K (2015). Splitting methods for convex clustering, *Journal of Computational and Graphical Statistics*, **24**, 994–1013.

Christopher DM, Prabhakar R, and Hinrich S (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Friedman N and Russell S (2013). *Image Segmentation in Video Sequences: A Probabilistic Approach*, arXiv preprint arXiv:1302.1539

Hocking TD, Joulin A, Bach F, and Vert JP (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning*, 745–752.

Hubert L and Arabie P (1985). Comparing partitions. *Journal of classification*, **2**, 193–218.

Harper FM and Konstan JA (2015). The movielens datasets: History and context. *Acm Transactions on Interactive Intelligent Systems (tiis)*, **5**, 1–19.

Hartigan JA and Wong MA (1979). Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 100–108.

Kvalseth TO (1987). Entropy and correlation: Some comments, *IEEE Transactions on Systems, Man, and Cybernetics*, **17**, 517–519.

Kuhn HW and Tucker AW (2014). Nonlinear programming, *Traces and Emergence of Nonlinear Programming*, 247–258.

Lindsten F, Ohlsson H, and Ljung L (2011). Clustering using sum-of-norms regularization: With application to particle filter output computation, *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 201–204.

Ng AY, Jordan MI, and Weiss Y (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 849–856.

O'Connor M and Herlocker J (1999). Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems*, **128**, UC Berkeley.

Park S and Zhao H (2018). Spectral clustering based on learning similarity matrix, *Bioinformatics*, **34**, 2069–2076.

Park S and Zhao H (2019). Sparse principal component analysis with missing observations, *Annals of Applied Statistics*, **13**, 1016–1042.

Park S, Xu H, and Zhao H (2021). Integrating multidimensional data for clustering analysis with applications to cancer patient data, *Journal of the American Statistical Association*, **116**, 14–26.

Sibson R (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, **16**, 30–34.

Tibshirani R, Saunders M, Rosset S, Zhu J, and Knight K (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 91–108.

Received December 10, 2021; Revised January 14, 2022; Accepted February 5, 2022

영화 데이터를 위한 쌍별 규합 접근방식의 군집화 기법

김희진[a,b], 박세영[1,a]

[a]성균관대학교 통계학과; [b]성균관대학교 핀테크융합전공

요 약

사용자들의 영화정보를 기록한 MovieLens 데이터는 추천 시스템 연구에서 아이디어를 탐색하고 검증하는데 상당한 가치가 있는 데이터로, 기존 데이터 분할 및 군집화 알고리즘을 사용하여 사용자 평점 데이터를 기반으로 항목 집합을 분할하는 연구 등에 사용되는 데이터이다. 본 논문에서는 기존 연구에서 대표적으로 사용되었던 영화 평점 데이터와 영화 장르 데이터를 통해 사용자의 장르 선호도를 예측하여 선호도 패턴을 기반으로 사용자를 군집화(clustering)하고, 유의미한 정보를 얻는 연구를 진행하였다. MovieLens 데이터는 영화의 전체 개수에 비해 사용자별 평균 영화 평점 수가 낮아 결측 비율이 높다. 이러한 이유로 기존의 군집화 방법을 적용하는 데 한계가 존재한다. 본 논문에서는 MovieLens 데이터 특성에 모티브를 얻어 쌍별 규합 벌점함수(pairwise fused penalty)를 활용한 볼록 군집화(convex clustering) 기반의 방법을 제안한다. 특히 결측치 대체(missing imputation)도 동시에 해결하는 최적화 문제를 통해 기존의 군집화 분석과 차별화하였다. 군집화는 반복 알고리즘인 ADMM을 통해 제안하는 최적화 문제를 풀어 진행한다. 또한 시뮬레이션과 MovieLens 데이터 적용을 통해 제안하는 군집화 방법이 기존의 방법보다 노이즈 및 이상치에 상대적으로 민감하지 않은 것으로 보인다.

주요용어: MovieLens, 볼록 군집화, alternating direction method of multipliers (ADMM), 결측치 대체, 최적화, 쌍별 규합 벌점함수

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: ishspsy@skku.edu