2018 빅콘테스트 Innovation 금융분야

고객 금융생활 정보지수 개발

77 x SHINHANBANK

표본의 특성을 활용한 PAM군집에 대한
Bayesian Network 적합과
Gibbs Sampling을 통한 문제해결

서문

제공받은 17076개 데이터는 표본조사 이론에 입각해 층화추출 된 고객들의 정보들로, 금융거래정보 모집단에 대한 대표성을 지닌다. 이 대표성 덕분에, 표본에 군집화를 실시해도 모집단을 군집화했을 때와 비슷한 양상을 관찰할 수 있다. 그렇기에 PAM 기법을 사용해서 주어진 데이터를 먼저 군집화 하였다. 이는 제공받은 금융정보 변수들의 왜도가 커서(밀도가 왼쪽에 치우쳐져 있음), 데이터를 그대로 사용하면 그 부분에 치우쳐진 결과를 얻을 우려가 있었기 때문이었다. 이렇게 적합시킨 군집별로 베이지안 네트워크를 적합시켜 변수들의 군집별 결합확률분포를 모델링하였고, 이를 통해 고객기본정보 유형별로 금융거래정보 항목의 결측치를 추정하였다.

이 과정을 다음의 순서로 설명한다:
1. 데이터 전처리 및 탐색적 분석
2. PAM기법을 통한 군집화와 Bayesian Network
3. Gibbs Sampling을 통한 유형별 금융거래정보 추정
4. Validation및 고객기본정보 variable selection

1 데이터 전처리 및 탐색적 분석(EDA)

1) Handling이 필요한 변수들을 수정·삭제한다.
2) 변수들의 결측값들을 채운다.
3) 전처리 내용들을 요약하고, 이를 바탕으로 용어 및 과제를 정의한다.

제거 · 수정된 변수들

1. 제거(TOT_ASSET, M_FUND_STOCK)

- TOT_ASSET = ASS_FIN + ASS_REAL + ASS_ETC
- M_FUND_STOCK = M_FUND + M_STOCK

다른 변수들의 합이라는 의미밖에 없음 → 삭제

2. 수정(M_TOT_SAVING, D_DAMBO, TOT_DEBT)

- 세 변수 모두 총액이라는 의미를 갖고 있는 변수들이므로,
  그 총액을 구성하는 부분들의 합보다는 항상 크거나 같아야 한다.
- 총액이 부분들의 합보다 작게 추정되는 것을 방지하기 위해
  총액변수들을 다음과 같은 부분들로 나눠서 추정한 후 차후에 합산한다:
  - M_ETC = M_TOT_SAVING – (M_JEOK 등 월 납입액 관련 5개 변수들의 합)
  - D_DAMBO_ETC = D_DAMBO – D_JUTEAKDAMBO
  - D_ETC = TOT_DEBT – (D_DAMBO_ETC 등 대출 잔액 관련 4개 변수들의 합)

제거 · 수정된 변수들

3. 수정(CHUNG_Y, MARRY_Y, DOUBLE_IN, NUMCHILD)

데이터 설명서에 정의된 대로 값들을 수정:

- CHUNG_Y — Null이 미보유, 5가 보유를 의미하므로 각각 N, Y로 수정
- MARRY_Y — '미응답'에 해당하는 level(3) 추가
- DOUBLE_IN — 결측값들을 '미응답'에 해당하는 값인 3으로 변경
- NUMCHILD — 결측값들을 '미응답'에 해당하는 값인 4로 변경 및 '없음'에 해당하는 level(0) 추가

금융상품 잔액 관련 변수들(TOT_YEA 등 5개 변수)

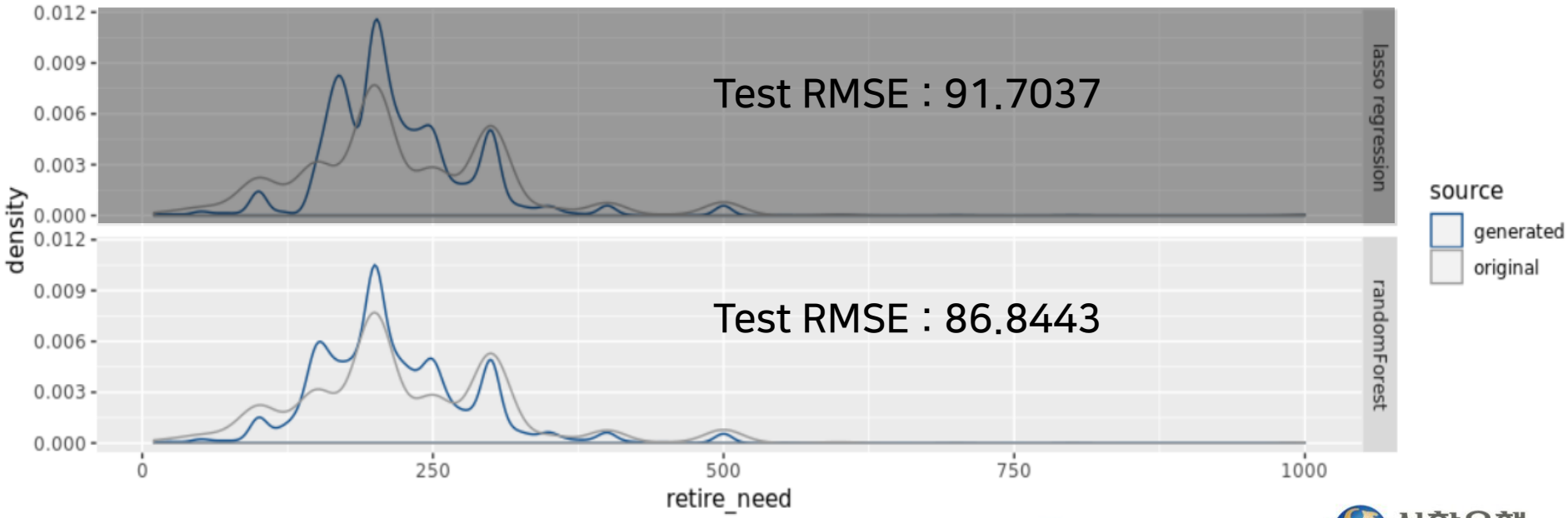
- 자신에게 해당되는 항목이 아니면 입력하지 않는 경향 존재
  (ex. 모든 미혼인 응답자들은 맞벌이 여부를 입력하지 않음)

- 이와 같이 금융상품 잔액이 없어서 해당 항목을 결측 처리했다면, 월 납입/투자액이 0이 아닌 응답자는 금융상품 잔액도 결측 처리하지 않았을 것이다(최소한 그 금액만큼의 잔액이 있을 것이므로).
  → 금융상품 잔액이 결측이면 월 납입/투자액이 0임을 확인
  → 금융상품 잔액 변수들의 NA를 모두 0으로 바꿈

- Randomforest imputation을 시행해도 결측값들을 0으로 대체했을 때와 매우 유사한 결과 얻음(앞선 논리 뒷받침)

은퇴 후 필요자금(RETIRE_NEED) imputation

- 하지만 은퇴 후 필요자금의 결측치는 응답자가 은퇴 후 생활비가 필요하지 않다고 생각해서 발생했다기보다는, 얼마가 필요할 거라고 구체적으로 생각해본 적이 없기에 결측치로 남겨뒀을 가능성이 크다.
  → RETIRE_NEED를 반응변수(Y)로 갖는 모형들(회귀모형, 랜덤포레스트)을 결측치가 없는 데이터로 train시켜 결측치들을 예측 (랜덤포레스트 선택)



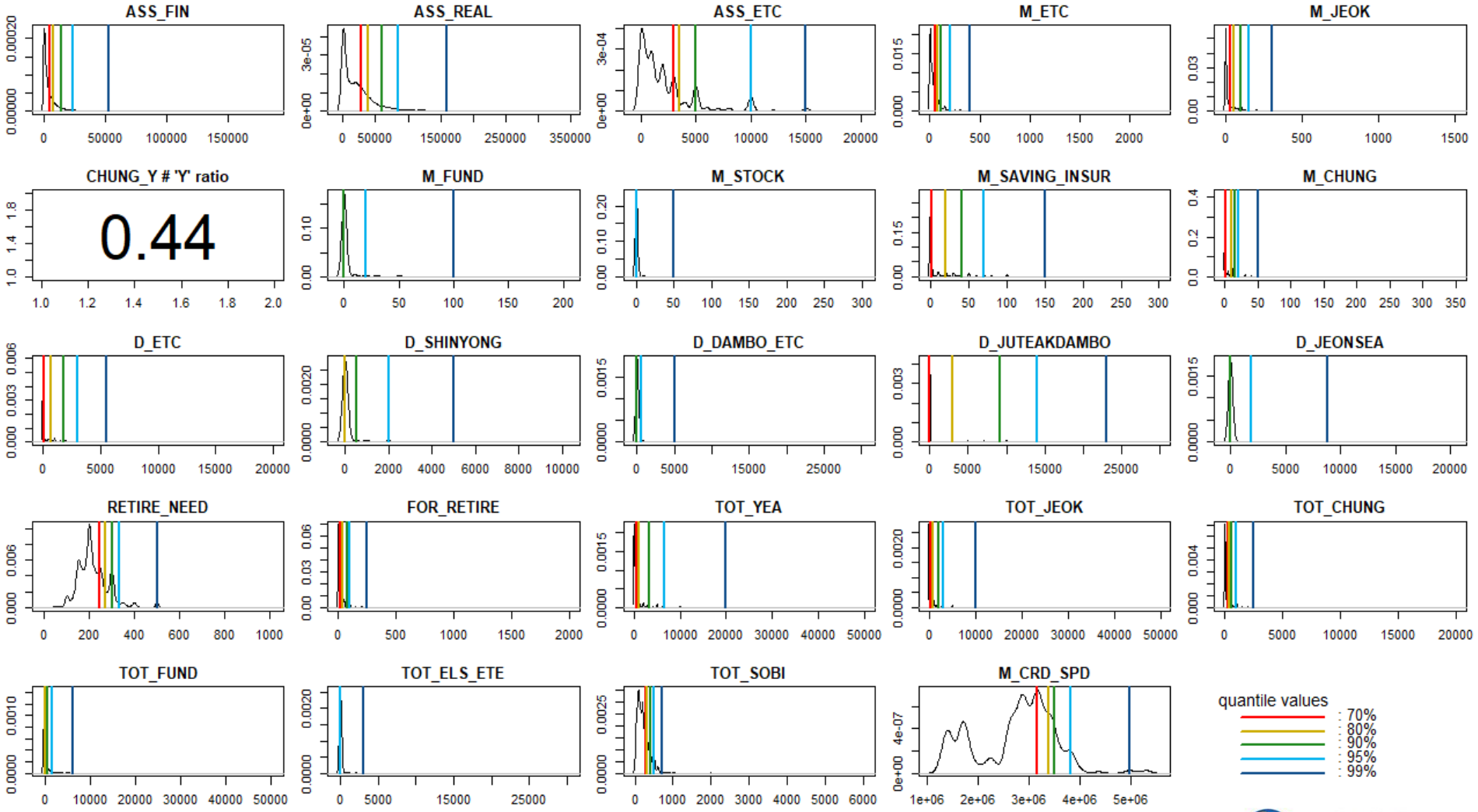
신한은행 SHINHAN BANK

1 Data handling
- 변수 제거 및 수정
- 결측값 처리
- EDA

2 분석의 논리적 배경
3 PAM · BN
4 Gibbs Sampling
5 Validation
6 잉여 고객기본정보
7 결론 및 의의

탐색적 분석(EDA)

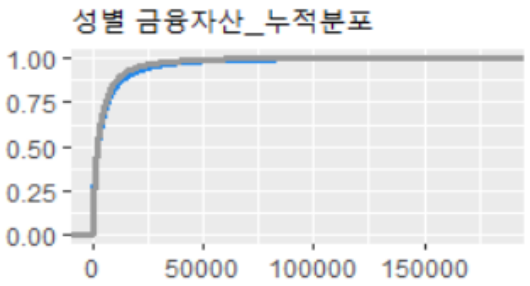
- **왜도가 높고**(밀도가 왼쪽으로 치우쳐져 있음) **이상치의 영향력이 큰** 변수들 많음

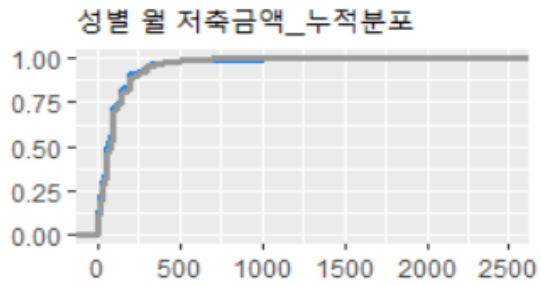


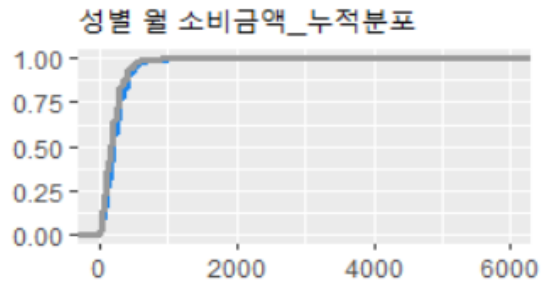
신한은행 SHINHAN BANK

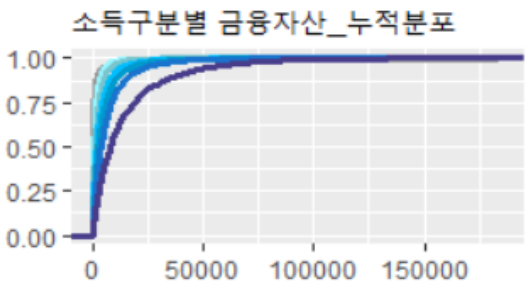
탐색적 분석(EDA)

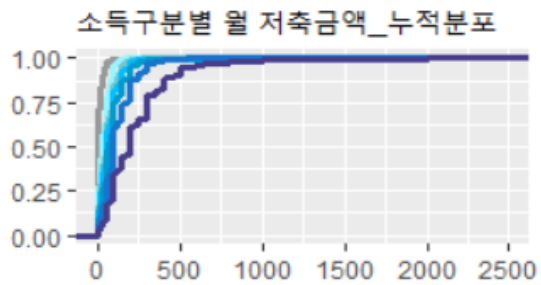
- **모든 고객기본정보들이 금융생활정보들에 영향을 미치지는 않음**
  ex. 금융생활정보 양상이 소득구분별 차이는 있지만, 성별별로는 큰 차이 없음

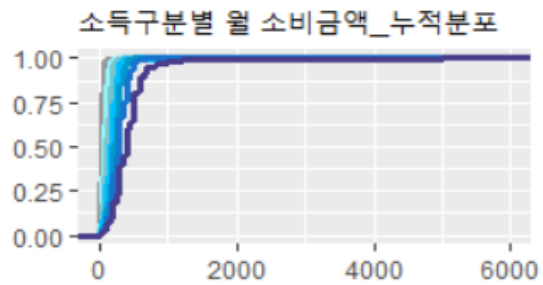












전처리 결과

- 표본 데이터: 기본정보 8개 포함 34개 변수로 이루어진 고객 17076명의 정보

  결측치 추정 : TOT_YEA 등 5개 변수(NA → 0), RETIRE_NEED(랜덤포레스트 사용)

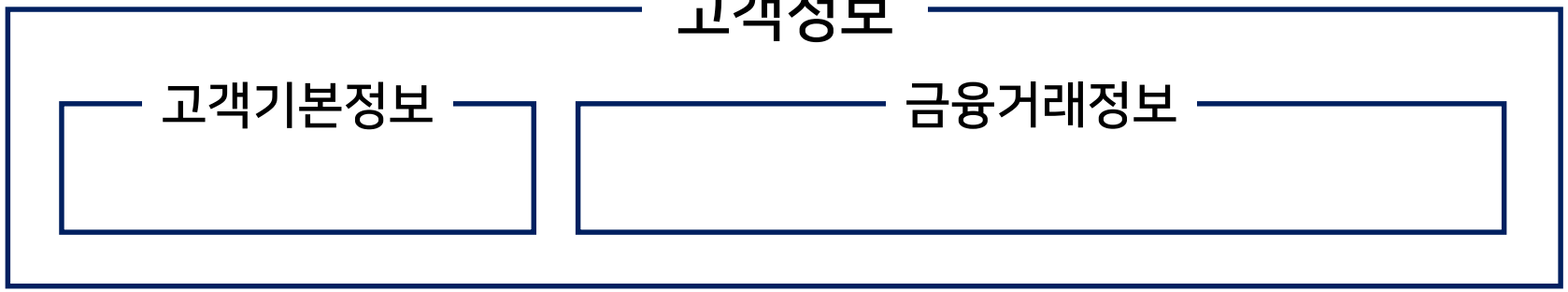
  수정 : MARRY_Y, DOUBLE_IN, NUMCHILD (데이터 정의서에 의거)
  M_TOT_SAVING(→ M_ETC), D_DAMBO(→ D_DAMBO_ETC),
  TOT_DEBT(→ D_ETC)

  제거 : TOT_ASSET, M_FUND_STOCK

- 이에 따라 데이터의 최종 형태는 다음과 같다:

고객기본정보 (SEX_GBN 등 8개)	금융거래정보 (ASS_FIN 등 24개)

용어 및 과제 정의



- **용어 정의**

  고객정보 고객의 고객기본정보, 금융거래정보를 아우르는 용어

  고객유형 한 고객기본정보 조합에 대한 24개 금융거래정보가 기록된 record
  (고객기본정보 조합 수가 141750개 이므로 고객정보유형 또한 141750개)

- **따라서, 주어진 과제는:**

  (1) 모든 **고객기본정보** 조합이 주어졌을 때, 각 **고객유형**별 **고객정보** 완성
  (2) **고객유형**들 군집화
  (3) **고객유형** 군집 각각의 금융자산, 총소비, 총저축의 백분위 도출
  (4) 잉여 **고객기본정보** 탐색 및 근거 제시

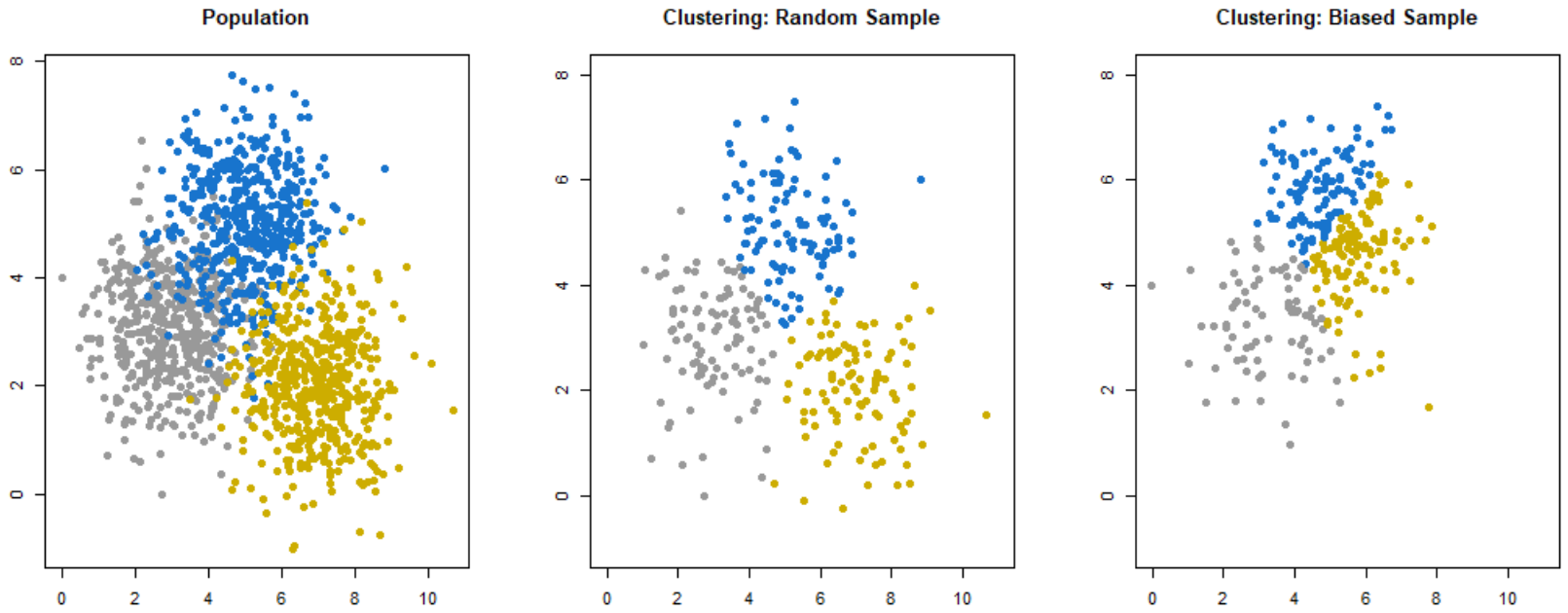
2 분석의 논리적 배경 및 흐름

1) 주어진 표본 데이터의 고객유형들을 군집화한 결과를 이용해서
   141750개 고객유형을 군집화하는 분석 방향의 정당성을 제시한다.
2) 이를 토대로 하는 분석의 흐름을 개괄한다.

논리적 배경: 분석방향 설정

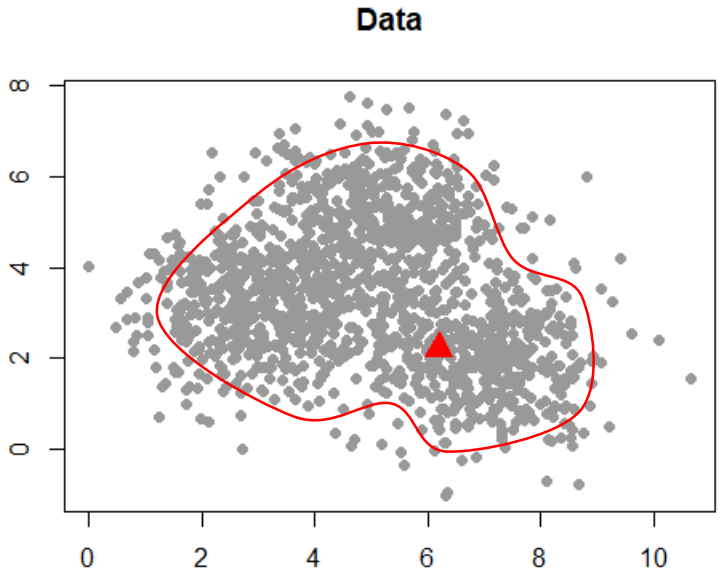
- 고객이 고객기본정보를 입력하면 그 고객이 적절한 peer group에 할당되도록 하는 것이 중요하다.
- 이때, 추정된 금융거래정보로 peer group을 분할하는 것은 추정값들이 정확하지 않을 경우 clustering을 왜곡할 수 있다.
- 추정값에서 발생한 오차가 군집화를 왜곡하는 것을 방지하기 위해, 주어진 표본 데이터로 clustering을 먼저 한다.

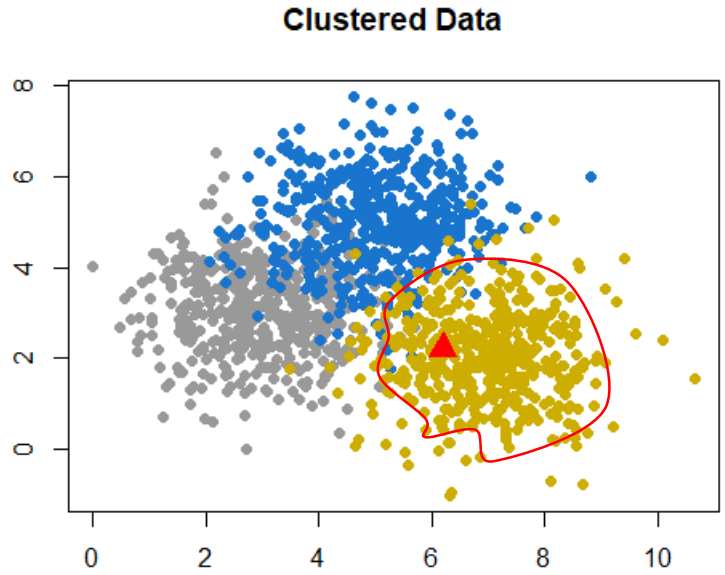
분석 방향의 정당성



- 이는 데이터가 모집단에 대한 대표성을 지닌 표본인 덕분에 가능
  제공받은 데이터는 표본조사 이론에 입각해 추출된 고객들의 정보들
  → **고객유형** 모집단에 대한 대표성을 지닌다
  → 따라서, **고객유형** 모집단을 군집화했을 때의 양상을 표본의 **고객유형**들을 군집화해도 관찰할 수 있다.

분석 방향의 정당성





- 또한, 데이터를 군집으로 나누어 모델링함으로써 정확도 개선
  → 한 **고객유형**이 가질 수 있는 **금융거래정보**들의 분포 범위를 보다 구체화

분석 방향 정리

1) 주어진 표본 데이터 내의 **고객유형**들을 군집화
2) 이 결과를 학습해 전체 **고객기본정보** 조합들 군집화
   - ✓ 이를 통해 141750개의 **고객기본정보** 조합마다 군집이 할당된다.
   - ✓ 한 **고객유형**은 고유한 **고객기본정보** 조합을 가진다.

- 따라서, 141750개의 **고객기본정보** 조합들을 군집화함으로써 141750개의 **고객유형**들을 군집화할 수 있다.
- 또한, **고객기본정보**를 통해 **고객유형**을 먼저 군집화함으로써, **금융거래정보** 추정에 도움을 줄 수 있다.

분석과정 개괄

1 주어진 표본의 **고객유형 군집화**
(PAM Clustering)

2 고객기본정보 조합별로 **군집 할당**
(Bayesian Network)

3 Gibbs Sampling을 위한 **고객정보 내 조건부독립관계 탐색**
(Bayesian Network)

4 고객유형별 고객 샘플 생성
(Gibbs Sampling)

3

PAM기법을 사용한 군집화와 Bayesian Network

1) 주어진 표본데이터를 고객유형들의 데이터로 변형시킨 후 군집화한다.
2) 각 군집 내 고객기본정보 조합들을 이용해 군집별로 Bayesian Network를 적합시키고, 이를 통해 모든 고객기본정보 조합들을 군집에 할당하는 원리를 설명한다.

고객유형 데이터

- 주어진 표본데이터는 고객정보 데이터이다.
  → 이를 고객유형 데이터로 만들기 위해 고객들의 금융거래정보들의 평균을 계산해 한 고객기본정보 조합 당 한 조합의 금융거래정보만 기록되게 함
  (CHUNG_Y의 경우, 해당 조합에서 가장 많이 관측된 값을 사용)

SEX_GBN	AGE_GBN	…	NUMCHILD	ASS_FIN	…	TOT_SOBI
1	2		0	300		110
1	2		0	7000		104
1	2		0	5900		300
1	2		0	16200		400
1	2		0	1500		350
1	2		0	40800		400
1	2		0	1000		50

SEX_GBN	AGE_GBN	…	NUMCHILD	ASS_FIN	…	TOT_SOBI
1	2		0	10385.71		262

- 이를 통해 총 5084개 고객유형 데이터 획득

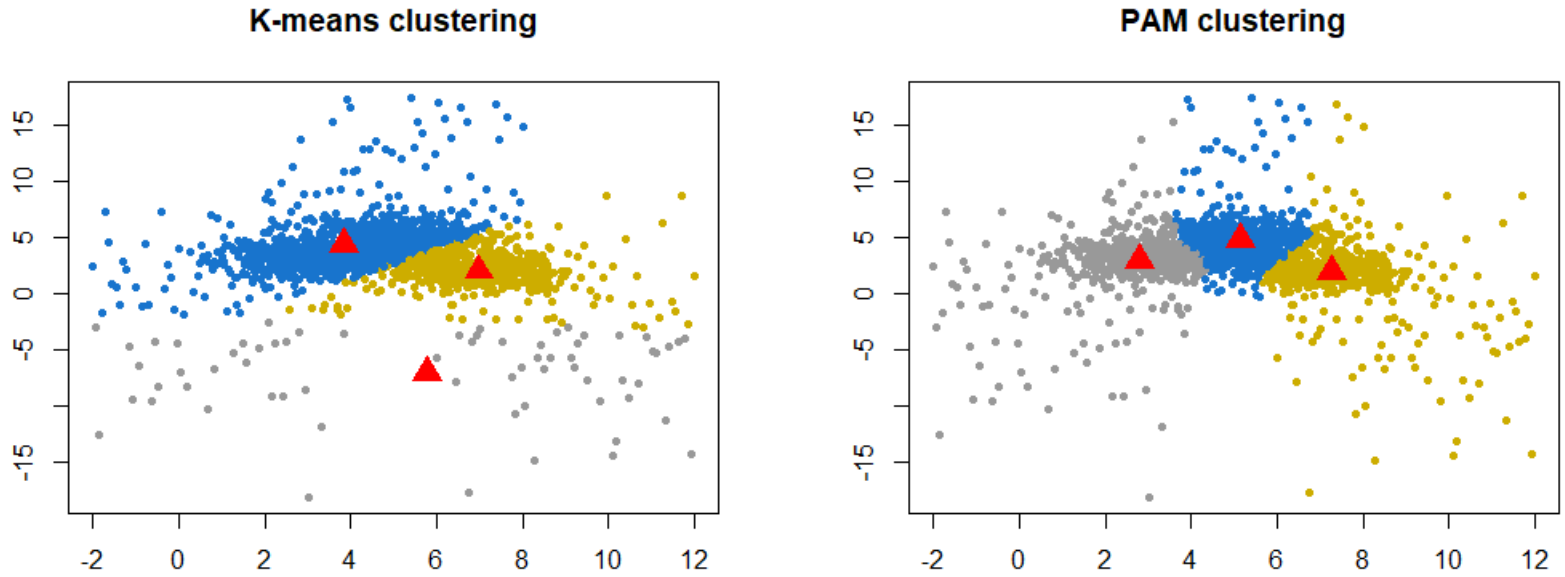
고객유형 데이터

유형	SEX_GBN	…	NUMCHILD	ASS_FIN	…	TOT_SOBI
1	1		1	770		125
2	1		2	550		300
3	1		4	296.667		114.231
…	…		…	…		…
5082	2		2	2500		400
5083	2		3	1320		300
5084	2		4	1028.75		205

- 5084개 고객유형 간 거리를 측정해야 한다.
  → 범주형 정보가 포함되어도 거리를 얻을 수 있는 측도 필요
  혼합형 자료의 거리가 0에서 1사이의 값으로 계산되는 gower의 유사도 사용

PAM(Partition Around Medoids)

- 금융거래정보 변수들의 특징: 이상치가 많음
  → 이상치의 영향력에 덜 민감한 군집화 방법 필요

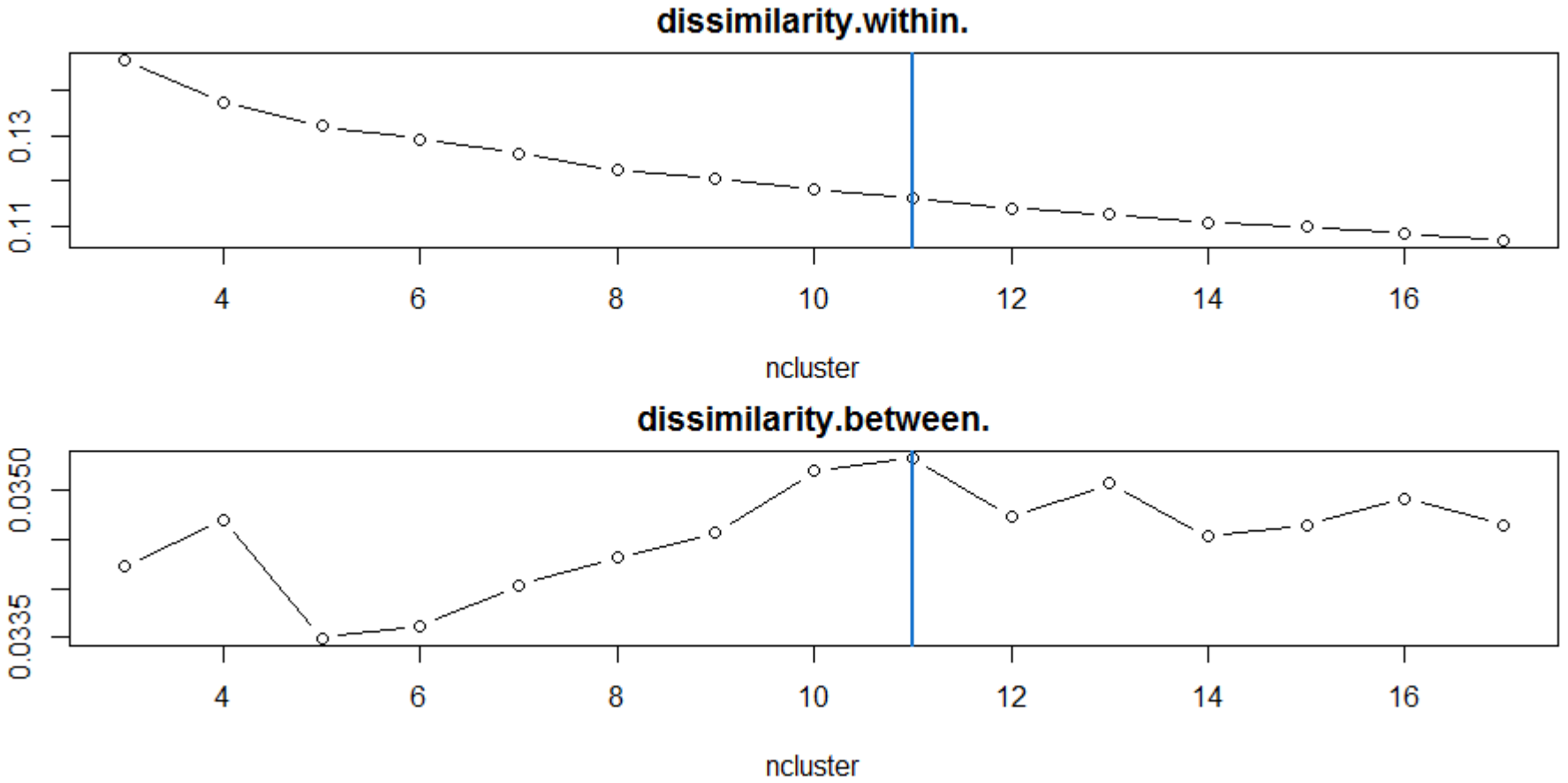


< 앞선 데이터에 이상치를 추가했을 때, 군집화 방법 별 군집화 결과 >

- K-means : 군집 내 데이터들의 평균으로 군집의 중심을 계산
  → 이상치가 군집의 중심을 계산하는 데 영향을 줘서 군집 중심의 위치를 왜곡시킴
- PAM : 군집 내 데이터를 군집의 중심으로 사용
  → 이상치가 군집의 중심을 계산하는 데 주는 영향 감소
  → PAM 기법을 통한 군집화 진행

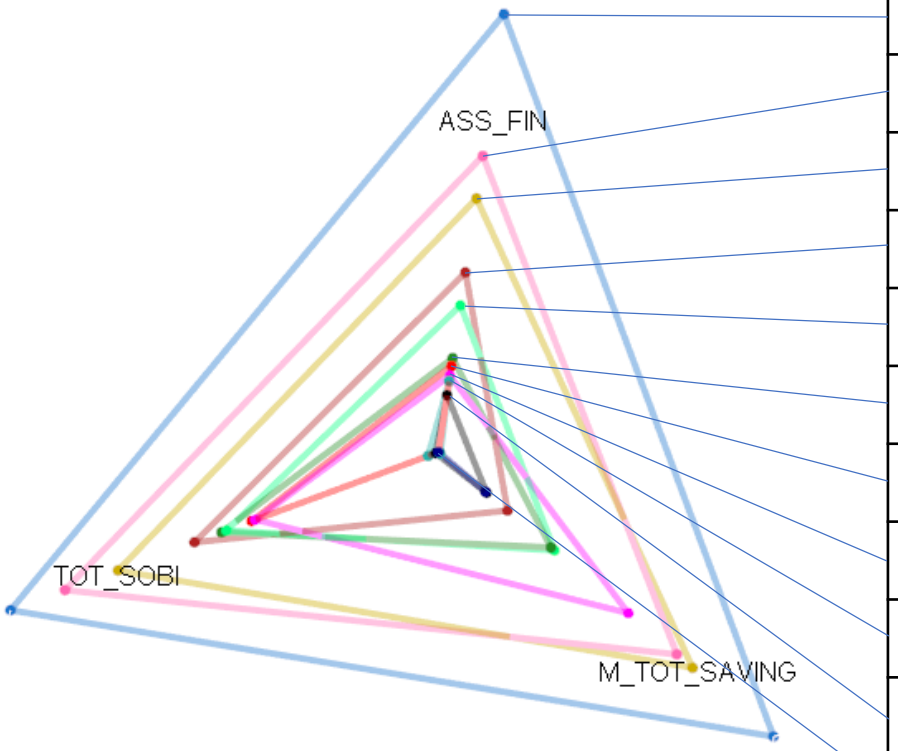
PAM(Partition Around Medoids)

- 다양한 군집 수를 검토해본 결과, 군집 수가 11개일 때가 가장 적절하다고 판단



PAM(Partition Around Medoids)

- 시각화 :11개 군집별 금융자산, 월 총 저축액, 월 소비액 평균의 minmax scaling 결과



Cluster	금융자산	월 저축	월 소비
6	12163	173	368
2	8852	138	332
3	7925	144	299
4	6382	84	253
10	5719	99	234
7	4690	97	237
8	4541	65	220
5	4358	122	218
9	4256	65	126
11	3991	78	123
1	2952	78	121

(단위 : 만 원)

고객기본정보 조합별로 군집별 할당확률 계산

- 고객기본정보 조합이 X=(x1, ⋯, x8)로 주어졌을 때,
  가장 할당될 확률이 높은 군집으로 그 조합을 할당

ex. 고객기본정보 조합이 X일 때, 그 조합이 i번째 군집으로 할당될 확률: P(K = i | X)

P(K = 1 | X) = 0.2
P(K = 2 | X) = 0.7
P(K = 3 | X) = 0.1

두번째 군집에 할당될 확률 최대(70%)
→ 조합 X를 두번째 군집으로 할당

고객기본정보 조합별로 군집별 할당확률 계산

- 고객기본정보 조합이 X=(x1, ⋯, x8)로 주어졌을 때,
  가장 할당될 확률이 높은 군집으로 그 조합을 할당

- 가장 할당될 확률이 높은 군집: $P(K = i|X)$가 최대화되는 i

$$\underset{i}{argmax}\, P(K = i|X) \;\; (i = 1, 2, \dots, 11)$$

- Bayes Theorem:

$$P(K = i|X) = \frac{P(X|K = i)P(K = i)}{P(X)}$$

- 따라서, 고객기본정보 조합별로 다음을 만족하는 군집 i를 찾아 할당한다:

$$\underset{i}{argmax}\, P(X|K = i)P(K = i)$$

($\because P(X)$는 $\underset{i}{argmax}$를 구하는 데는 영향을 미치지 않음)

신한은행 SHINHAN BANK

고객기본정보 조합별로 군집별 할당확률 계산

- **1.** $\boldsymbol{P(K = i)}$
  (i 번째 군집에 할당된 고객유형 수) / (총 고객유형 수: 5084)

- **2.** $\boldsymbol{P(X|K = i) = P(X1, \ldots, X8|K = i)}$
  (X1 : 성별, X2 : 나이, … ,X8 : 자녀 수)
  고객기본정보 조합의 가능한 경우의 수는 141750개
  → 한 군집별로 추정해야 할 모수의 개수 : (141750-1)개

SEX_GBN	AGE_GBN	…	NUMCHILD	P(X1, X2, …, X8)
1	2		0	θ_1
1	2		1	θ_2
…	…		…	…
2	6		3	θ_{141749}
2	6		4	$1 - (\theta_1 + \cdots + \theta_{141749})$

(17076개 데이터로 추정 불가)

→ 각 고객기본정보 분포를 조건부분포들의 곱으로 분할해 추정해야 할 모수의 수를 줄여야 한다.

신한은행 SHINHAN BANK

$P(X|K=i)$ 구하기: 조건부독립관계 탐색

- 고객기본정보의 군집별 결합분포를 조건부분포들의 곱으로 분할

$$P(X1, X2, \ldots X8 \mid K)$$
$$= P(X1 \mid K)P(X2 \mid X1, K) \cdots P(X8 \mid X1, \ldots, X7, K)\ (Chain\ rule)$$
$$= P(X1 \mid K)P(X2 \mid X1, K) \cdots P(X8 \mid X1, X2, K)\ (Conditional\ Independence)$$

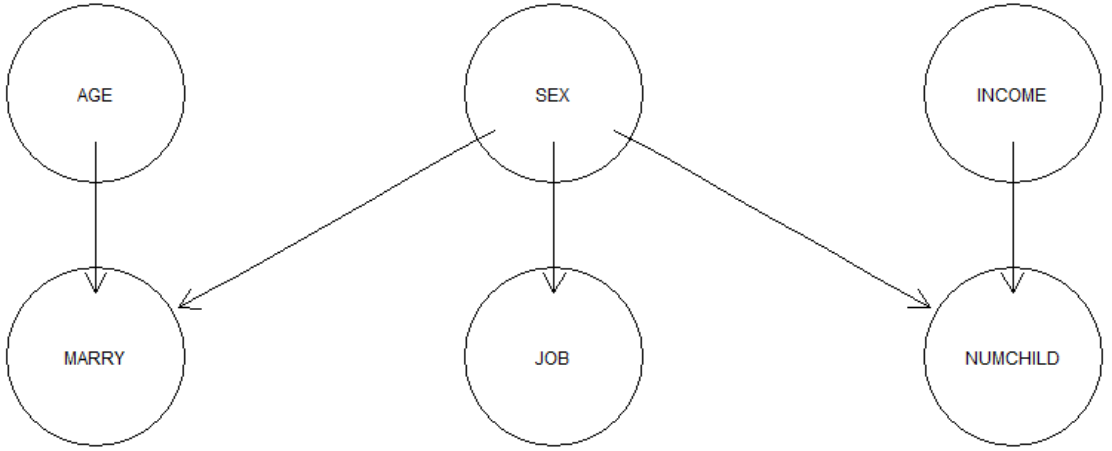
(: X8은 X1, X2가 주어졌을 때 X3, …, X7과 조건부독립)

- 이렇게 분할하기 위해선 X1, …, X8간 조건부 독립 관계를 찾아야 함
  → 군집 별 Bayesian Network 적합을 통한 군집 별 고객기본정보간 조건부독립관계 탐색

$P(X|K=i)$ 구하기: 조건부독립관계 탐색

- Bayesian Network
  : 일방향, 비순환 그래프(DAG: Directed Acyclic Graph)를 통해
  결합확률분포를 여러 조건부 확률분포들의 곱으로 나타낼 수 있게 해주는 그래프 모델

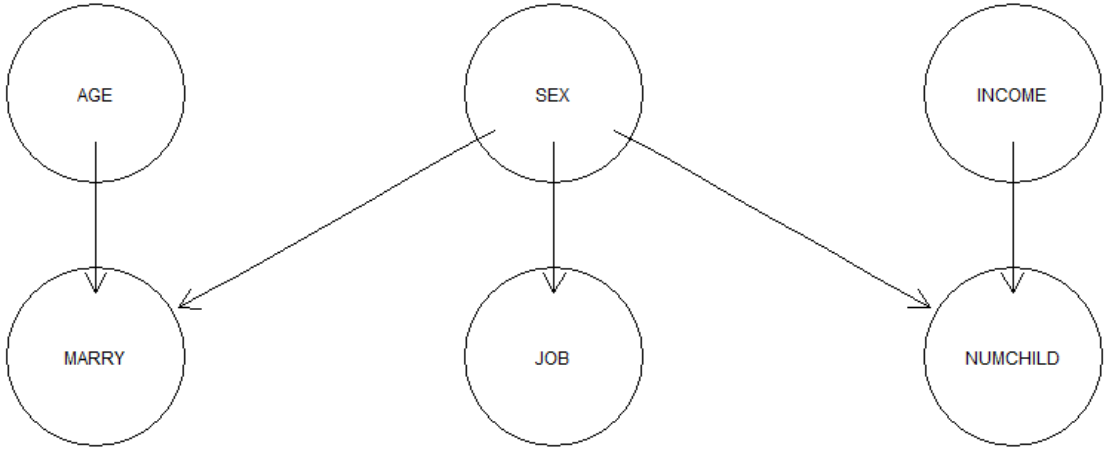
ex) 고객기본정보들로 적합시킨 Bayesian Network의 예시



- 화살표를 보내는 변수를 Parent, 받는 변수를 Child라 한다.
  → Parent가 주어졌을 때, Child는 자신의 Child를 제외한 다른 변수들과 조건부독립이다.
  ex) NUMCHILD의 Parent는 AGE, SEX
  SEX의 Child는 NUMCHILD, JOB, DOUBLE_IN

$P(X|K=i)$ 구하기: 조건부독립관계 탐색

- Bayesian Network
  : 일방향, 비순환 그래프(DAG: Directed Acyclic Graph)를 통해
  결합확률분포를 여러 조건부 확률분포들의 곱으로 나타낼 수 있게 해주는 그래프 모델
- 조건부 확률분포들로 분해된 결합확률분포:



$$P(SEX, AGE, INCOME, NUMCHILD, JOB, MARRY)$$
$$= P(AGE)P(SEX)P(INCOME)P(NUMCHILD|SEX, INCOME)P(JOB|SEX)P(MARRY|SEX, AGE)$$

- 이를 통해 군집별로 추정해야 할 모수의 개수가 141749개에서 최대 87개로 감소

$P(X|K=i)$ 구하기: 조건부독립 분포들의 모수 추정

- **모수의 추정**

  ex) 군집 2의 성별, 소득, 결혼여부의 결합확률분포가 다음과 같이 분해된다 하자:

  P(SEX, AGE, MARRY | K=2)

  = P(SEX | K=2)P(AGE | K=2)P(MARRY | SEX, AGE, K=2)

  군집 2의 고객이 20대 남성이고, 결혼여부를 응답하지 않을 확률

  = P(SEX=1, AGE=2, MARRY=3 | K=2)

  = P(SEX=1 | K=2)P(AGE=2 | K=2)P(MARRY=3 | SEX=1, AGE=2, K=2)

$P(X|K=i)$ 구하기: 조건부독립 분포들의 모수 추정

- 최대가능도 추정법(Maximum Likelihood Estimation)

$$P(MARRY = 3 \mid SEX = 1, AGE = 2, K = 2)$$
$$= \frac{\text{2군집 내 20대 남성 중 결혼여부를 응답하지 않은 사람의 수}}{\text{2군집 내 20대 남성의 수}}$$

- 하지만 17076명의 설문자 중 결혼여부를 응답하지 않은 사람이 없으므로 P(MARRY=3 | SEX=1, AGE=2, K=2) = 0이 되고, 이는 2군집 뿐 아니라 어떤 군집에 대해서도 속할 확률이 0이 되는 결과로 이어진다.

- 즉, 최대가능도 추정법으로는 이런 고객기본정보 조합을 어떤 군집에도 할당할 수 없다.

$P(X|K=i)$ 구하기: 조건부독립 분포들의 모수 추정

- 베이지안 추정법(Maximum a Posteriori)

$$P(MARRY = 3 \mid SEX = 1, AGE = 2, K = 2)$$

$$= \frac{P(MARRY = 3, SEX = 1, AGE = 2 \mid K = 2)}{P(SEX = 1, AGE = 2 \mid K = 2)}$$

$$= \frac{\frac{a}{2\text{군집 인원 수} + a} * \frac{1}{2*5*3} + 0}{\frac{a}{2\text{군집 인원 수} + a} * \frac{1}{2*5} + \frac{2\text{군집 인원 수}}{2\text{군집 인원 수} + a} * \frac{2\text{군집 내 20대 남성 수}}{2\text{군집 인원 수}}}$$

- 다항분포의 모수가 따르는 확률분포인 Dirichlet 분포의 성질을 이용하여,
가상의 sample size(a)를 추가해 확률이 0으로 추정되는 것을 방지($a \neq 0$)
(이때, a의 영향력을 줄이기 위해 일반적으로 a를 15 이내의 값으로 설정)

- 따라서, 이 추정법을 통해 고객유형의 군집별 결합확률분포들의 모수를 추정
→ 모든 고객기본정보 조합들의 $argmax_i P(K = i|X)$ 산출 가능

PAM 군집별 Bayesian Network

1 주어진 표본의 **고객정보유형 군집화**
(PAM Clustering)

2 고객기본정보 조합별로 **군집 할당**
(Bayesian Network)

- **표본의 정확한 고객유형들을 가지고 군집화를 먼저 실행**
  → 이상치에 영향을 덜 받는 PAM clustering으로 PAM군집 11개 적합

- **군집별로 고객기본정보를 가지고 Bayesian Network 적합**
- **이때, 표본에 담기지 않은 조합의 발생 확률이 0으로 추정되는 것을 방지하기 위해 베이지안 추정법 사용**
  → X(고객기본정보 유형)별로 속할 확률이 최대인 군집으로 X를 할당

- **이를 통해 141750 가지 고객정보유형 각각을 11가지 군집 중 하나로 할당**

4 Gibbs Sampling을 통한 고객기본정보 조합별 금융거래정보 추정

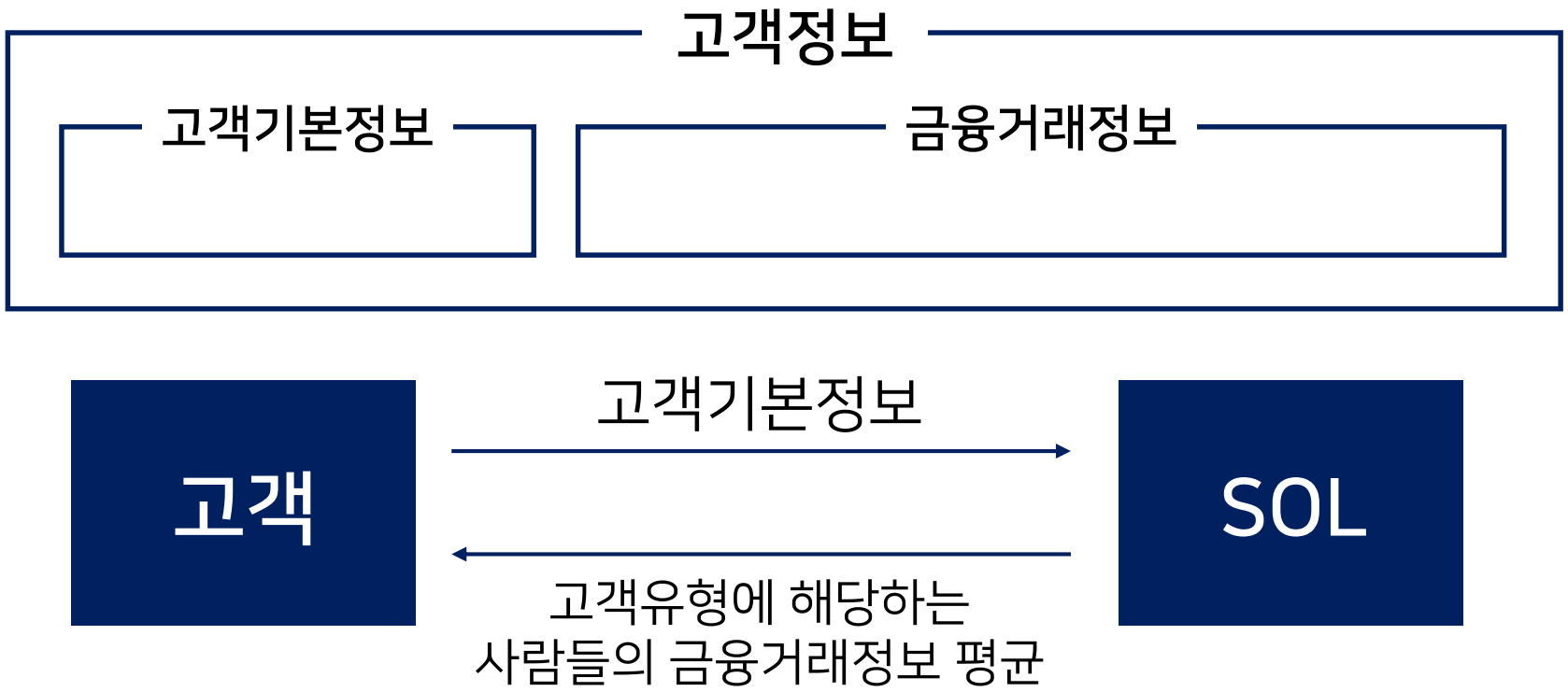
1) 고객기본정보 조합별 금융거래정보들을 추정하기 위한 방법으로 깁스 샘플링을 선택한 배경을 설명한다.
2) 주어진 데이터에 깁스 샘플링을 구현하기 위해 다음과 같은 개념을 설명한다:
   - Acceptance-Rejection 알고리즘
3) 깁스 샘플링을 통해 141750개의 고객기본조합 별 금융거래정보를 추정한다.

군집화 결과가 반영된 고객정보 데이터

idx	SEX_GBN	…	NUMCHILD	ASS_FIN	…	M_CRD_SPD	clust_num
1	1	…	0	300	…	1437017	1
2	1		0	7000		3211358	11
3	2		1	5900		2932871	3
…	…		…	…		…	…
17074	2		2	12300		3106659	3
17075	2		2	2050		3615840	7
17076	1		0	22650		2835021	9

- 5084개 고객유형을 사용해 군집화를 진행했던 앞선 단계와는 달리, 이번 단계에서는 고객정보를 sampling하기 위해 17076개의 고객 데이터를 활용해 Gibbs Sampling을 진행한다.
- 따라서, 군집화 결과를 통해 얻은 군집변수를 추가한 17076개 고객정보 데이터를 사용한다.

Sampling을 통한 문제해결 방향



- 특정 고객기본정보 조합에 해당하는 사람들의 금융거래정보 별 평균 필요
  → 시뮬레이션을 통해 조합별로 가상의 고객 금융거래정보를 생성(Sampling)
  → 생성된 금융거래정보에서 평균 도출

Gibbs Sampling 소개

- 고객유형의 군집별 결합확률분포로부터 고객유형별 금융거래정보 생성

고객기본정보 금융거래정보

$$P(SEX, AGE, \cdots, NUMCHILD, ASS\_FIN, \cdots, M\_CRD\_SPD \mid K=i)$$

i번째 군집의 고객유형 결합확률분포(i=1,2,…,11)

- 위의 결합확률분포를 알고 있지 않기에
  이 분포로부터 직접 유형별 금융거래정보를 sampling하는 것은 불가능하다.

- 따라서, 모든 금융거래정보 변수의 조건부 확률분포에서 생성한 sample을 조합해 결합확률분포에서 얻은 sample과 동질적인 결과를 얻기 위해 Gibbs Sampling 을 사용한다.

Gibbs Sampling 소개

ex) 알려지지 않은 결합확률분포 $P(X_1, X_2, \dots, X_n)$으로부터의 깁스샘플링

1. 임의의 초기값 $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})$ 선택
2. 다음을 반복(i=1,2,⋯,K, K는 반복 수)
   1) $P(X_1 | X_2 = x_2^{(0)}, \dots, X_n = x_n^{(0)})$에서 $x_1^{(i)}$ 생성
   2) $P(X_2 | X_1^{(i)}, \dots, X_n^{(i-1)})$에서 $x_2^{(i)}$ 생성
   ⋮
   n) $P(X_n | X_1^{(i)}, \dots, X_{n-1}^{(i)})$에서 $x_n^{(i)}$ 생성
   n+1) i번째 샘플인 $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})$를 획득
3. 처음 m개의 샘플을 버리고(burn-in) K-m개의 샘플을 획득

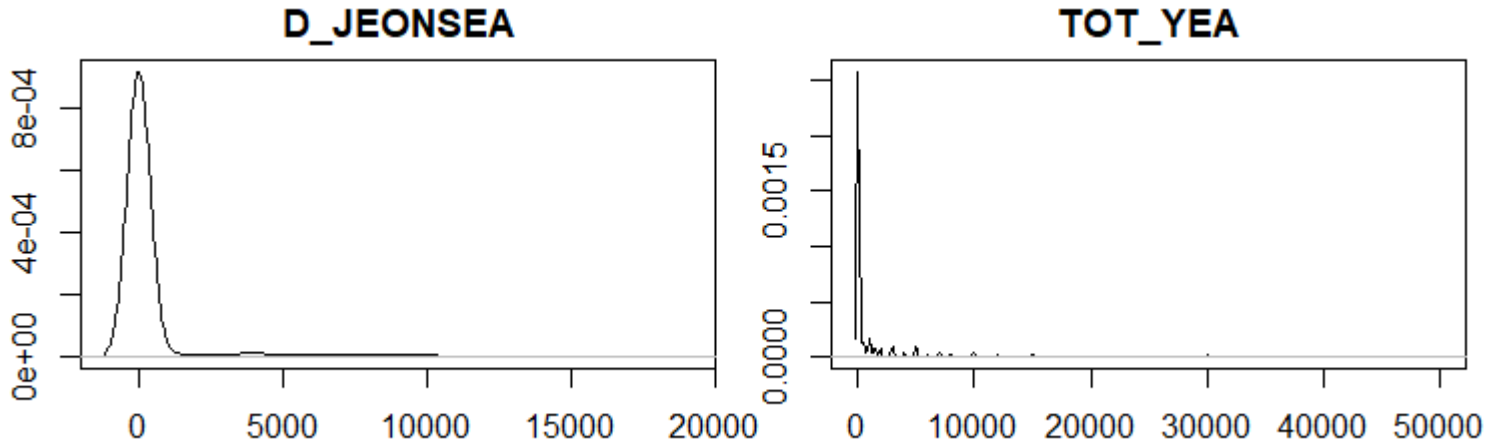
- 이때, 1) ~ n) 단계에서 각 변수들의 조건부 분포를 알고 있다는 조건이 필요하다.
- 하지만 주어진 데이터 내 대부분의 변수는 왜도가 높고 이상치의 영향력이 크므로 조건부 분포를 특정할 수 없다.
  → 데이터로부터 근사시킨 Empirical Conditional Density Function을 사용.

데이터 내 조건부 확률분포 예시

ex) 4군집 내 고객유형 결합확률분포의 일부:

$$P(SEX, AGE, MARRY, D\_JEONSEA, TOT\_YEA | K=4)$$

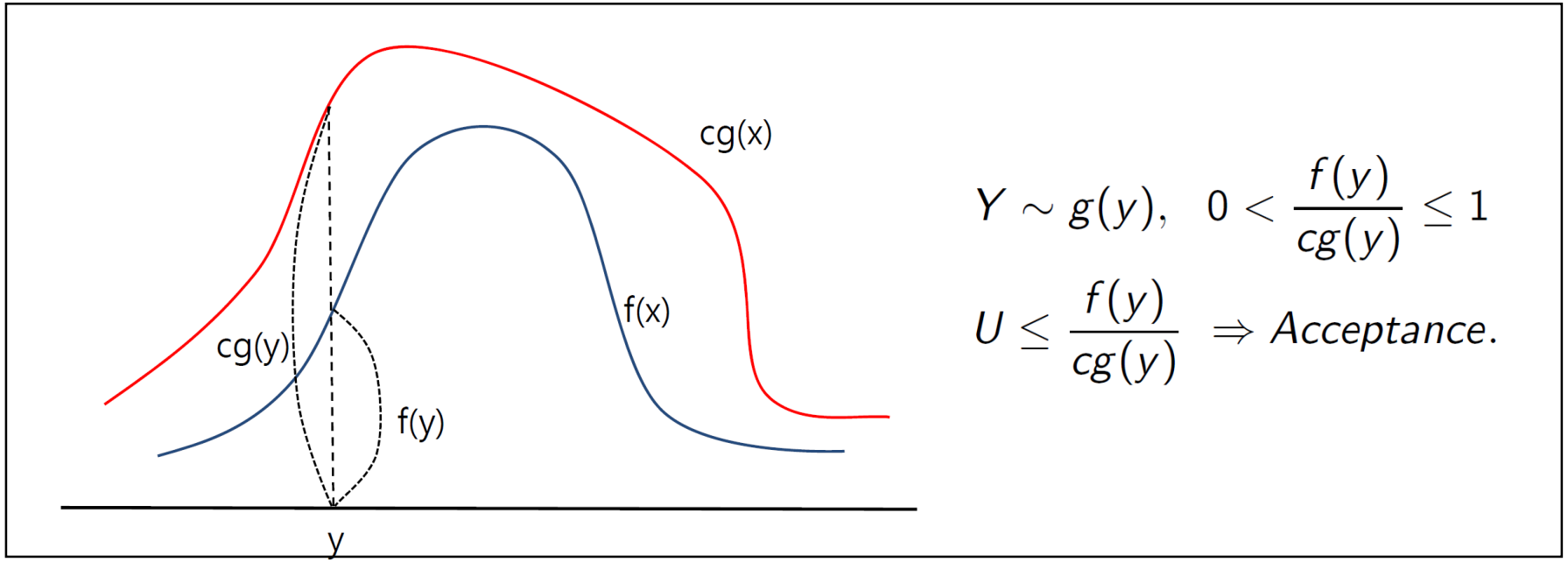
- 이 분포에서 20대 미혼 남성의 전세자금대출 잔액과 정기예금 잔액 정보를 sampling하려면 각 변수들의 분포 형태가 특정되어야 함(ex. 정규분포)



< 4군집 내 고객들의 전세자금대출 잔액과 정기예금 잔액의 분포 >

- 하지만 이는 왜도가 높고 이상치가 많은 데이터의 특성상 사실상 불가능
  → ECDF만으로도 샘플링이 가능한 Acceptance-Rejection알고리즘 선택

Acceptance-Rejection 알고리즘



$$Y \sim g(y), \quad 0 < \frac{f(y)}{cg(y)} \leq 1$$

$$U \leq \frac{f(y)}{cg(y)} \Rightarrow Acceptance.$$

- ECDF f(x)와 비슷하지만 잘 알려져 있는 분포인 g(x)를 통해 샘플링
(단, ecdf가 정의된 모든 구간에서 $f(y) \leq c * g(y)$를 만족하는 상수 c가 존재)

조건부 확률분포와 Bayesian Network

- 하지만 empirical한 조건부 확률분포(ECDF)는 다음과 같은 상황에서는 정의되지 않음:
  1. 연속형 변수의 값이 condition되었을 때
  2. condition variable의 숫자가 너무 많아 해당 조건을 만족하는 observation이 존재하지 않을 때

- 이를 해결하기 위해 다음과 같은 방법을 사용했다:
  1. 연속형 변수를 구간화(discretize)함
  2. Bayesian Network를 통한 조건부독립 관계 탐색
     → 불필요한 condition variable 차단

금융거래정보 추정

- 앞선 과정을 통해 주어진 데이터에 대한 Gibbs Sampling 제약 해소
  1. Sampling에 필요한 변수별 조건부확률 분포를 알 수 없었음
     → ECDF를 사용하여 조건부확률분포 대체
  2. 연속형 변수가 condition 되었을 때 ECDF 도출 불가
     → 연속형 변수의 특정 값이 아니라 구간에 condition되게 함
  3. condition variable이 너무 많은 경우 ECDF 도출 불가
     → 군집별 고객정보에 Bayesian Network를 적합시킴으로써 불필요한 condition variable 차단

- 141750개 고객유형 별 300개의 Gibbs Sample 생성
  → 300개 sample의 평균으로 금융거래정보 추정

5 Validation

지금까지 설명한 방법의 효과를 입증하기 위해 Training/Test 데이터를 분할하여 다음을 평가한다:

1) 고객기본정보 조합의 군집 할당 정확도
2) 금융거래정보 분포 유사도
   - 고객유형별
   - 군집별

Validation 대상

- 분석 방법의 효과를 입증하기 위해서 다음을 측정한다:

1. **군집 할당 정확도**
   한 고객이 자신의 기본정보들을 입력했을 때,
   그 고객이 속해야 할 Peer group으로 그 고객이 정확하게 할당되어야 한다.

2. **고객유형 별 금융거래정보 분포 유사도**
   생성된 Gibbs Sample들이 고객유형의 결합확률분포를 대표함을,
   따라서 이 sample들로 구한 고객유형별 금융거래정보의 평균(1_Data Set)
   이 정당함을 보인다.

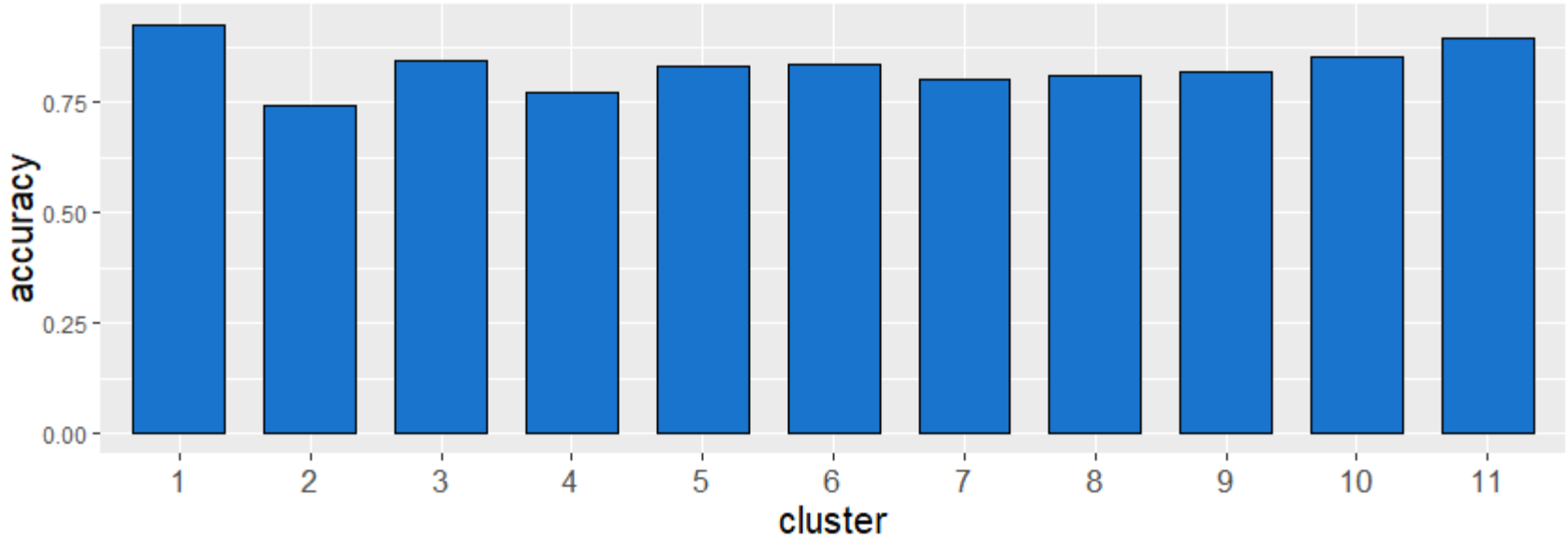
3. **군집별 금융자산, 월 저축금액, 월 소비금액 분포 유사도**
   군집별로 생성한 Gibbs Sample들의 금융자산, 월 저축금액, 월 소비금액의
   분포가 실제 군집별 금융자산, 월 저축금액, 월 소비금액의 분포와 유사함을
   보임으로써 군집별 백분위 테이블(3_Quantile Table)이 정당함을 보인다.

1. 군집 할당 정확도

- Training/Test 데이터를 분할해 군집별 할당확률을 계산하는 방법의 정당성을 확인한다.

- 다음의 절차를 통해 이를 검증한다:
  1. 앞서 획득한 군집화 결과를 17076개 표본 데이터에 저장한다.
  2. 이 데이터를 Training, Test로 나눈 후 Test 데이터의 군집화 결과를 따로 보관한 다음(answer), Test 데이터의 군집정보를 삭제한다.
  3. Training 데이터의 고객기본정보를 사용해 군집 별 Bayesian Network를 적합하고, 이를 기반으로 Test 데이터 내 고객기본정보 조합들에 군집을 할당한다(estimate).
  4. 11개 군집별로 answer와 estimate가 일치하는 개수를 센 후, 군집 별 고객 수로 나누어 군집 별 고객 할당 성공률을 구한다.

군집 할당 정확도 측정



< 군집 별 예측 성공률 >

- 그 결과, 군집별로 다음과 같은 예측 성공률을 얻었다.
  평균 정확도 82.7%

2. 고객유형 별 금융거래정보 분포 유사도

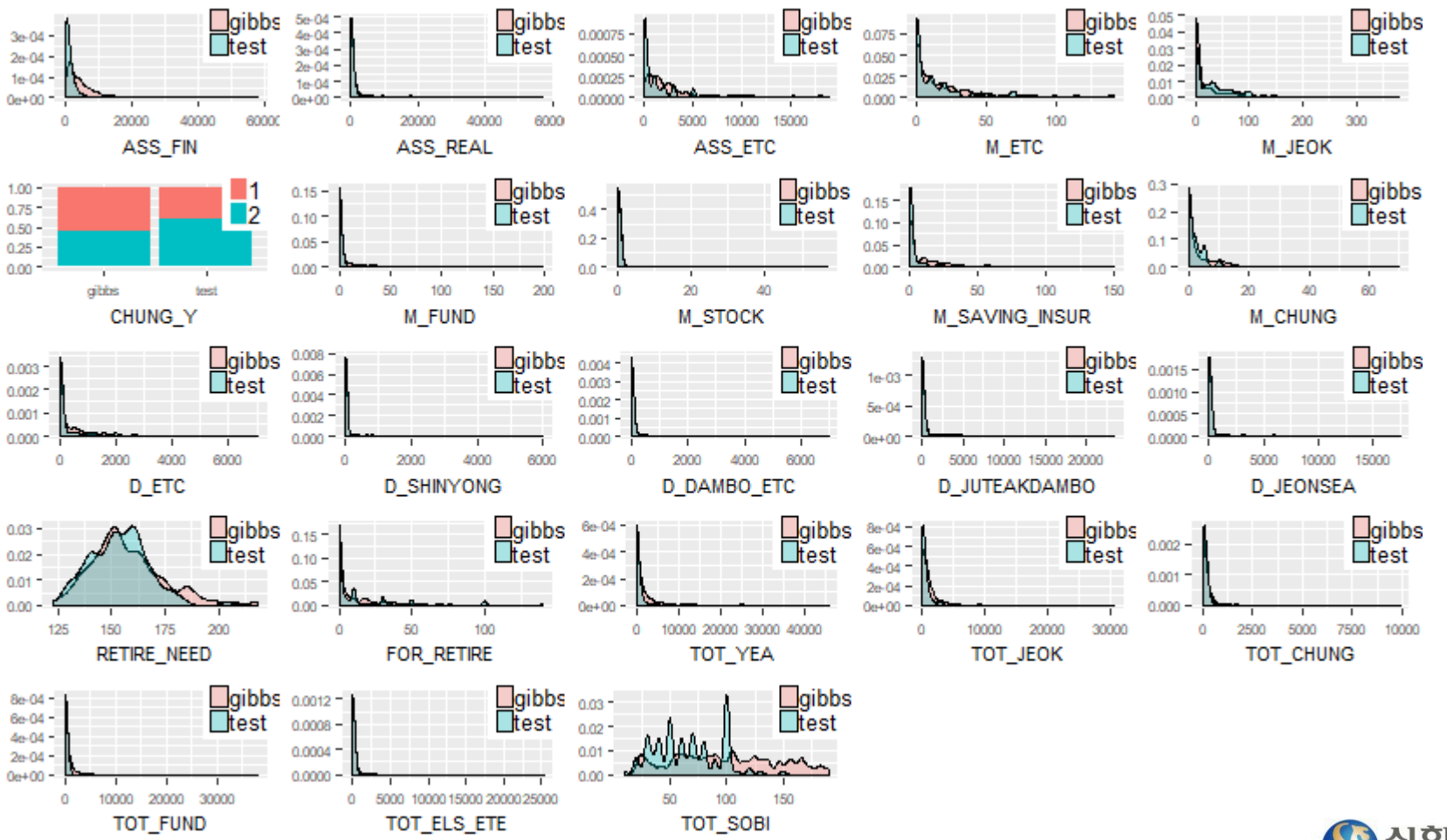
- Gibbs Sampling을 통해 생성된 sample이 고객유형의 결합확률분포에서 추출되었음을 보인다.

- 다음의 절차를 통해 이를 검증한다:

1. 고객유형별로 Train/Test를 나눈다.
2. Train 데이터에 PAM 군집을 적합하고, 3장에서 설명했던 방법을 통해 Test 데이터에 군집을 할당한다.
3. Gibbs Sampling을 통해 군집별 Train 데이터에서 Test 데이터의 고객유형별 금융거래정보 생성
4. Test 데이터 내 고객유형별 금융거래정보와 생성된 고객유형별 금융거래정보를 시각화해 비교

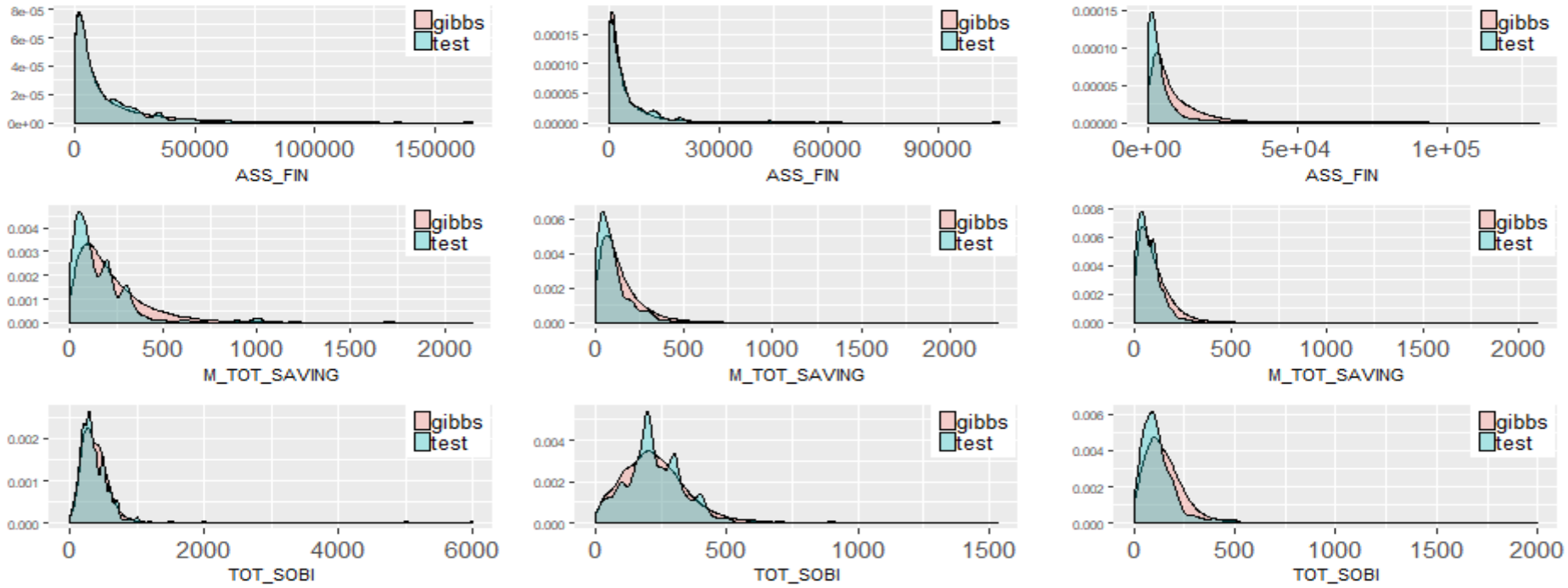
2. 고객유형 별 금융거래정보 분포 유사도

- 효과적인 시각화를 위해, Test 데이터에서 가장 많은 고객에 해당되었던 유형의 금융거래정보와, 이 유형에 대한 Gibbs Sample들의 금융거래정보를 비교했다.



3. 군집별 금융자산, 월 저축금액, 월 소비금액 분포 유사도

- 군집별 금융자산, 월 저축금액, 월 소비금액 분포 유사도 또한 실제 군집과 군집 별 Gibbs Sample들의 시각화를 통해 비교함으로써 평가할 수 있다.
- **이를 통해 Gibbs Sample이 Peer Group의 분포에 맞춰 생성되었음을 확인**



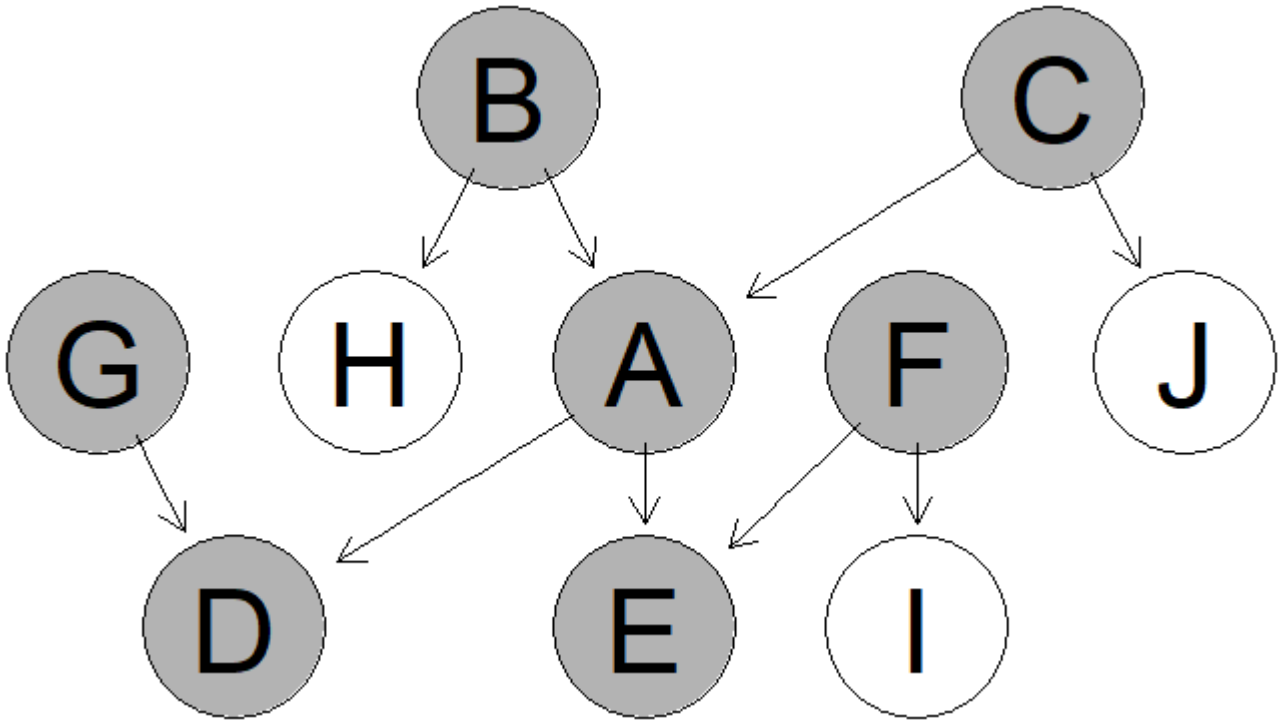
상위(6군집) 중위(7군집) 하위(11군집)

*군집별 금융자산의 평균 기준

6 잉여 고객기본정보 탐색

1) 잉여 고객기본정보를 정의하고, 이를 탐색할 방법으로 Markov Blanket을 사용한다.
2) 찾아낸 잉여 고객기본정보를 제외했을 때의 validation을 성능을 확인한다.

Markov Blanket



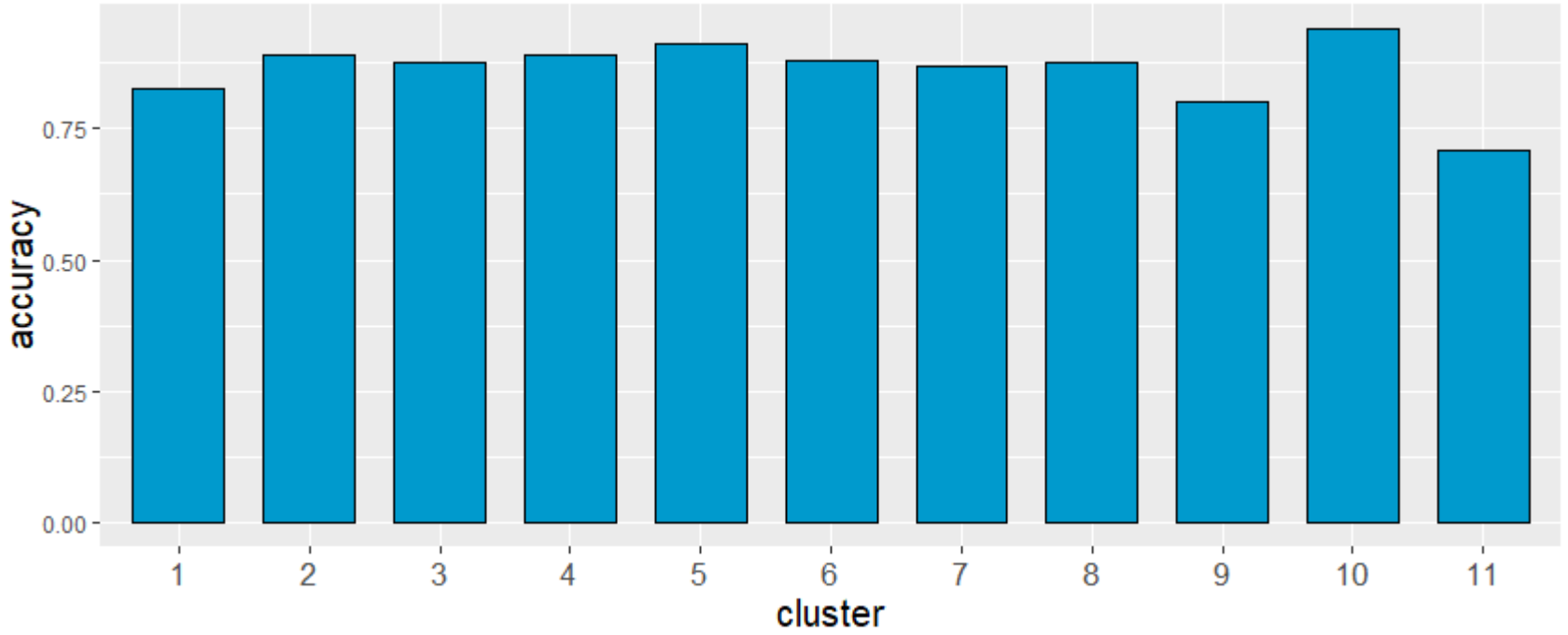
< A의 Markov Blanket. B,C,D,E,F,G가 주어지면 A는 H,I,J와 독립이다 >

- Bayesian Network에서 Markov Blanket이란 어떤 변수의 Parent, Child 그리고 Child의 Parent에 해당하는 변수들을 말한다.
- 어떤 변수의 Markov Blanket이 주어지면 그 변수는 Markov Blanket 밖의 변수들과 독립이다.

Markov Blanket: 잉여정보 정의

- 따라서, 한 군집의 잉여 고객기본정보란, 모든 금융거래정보들의 Markov Blanket 밖에 있는 변수라고 할 수 있다.
- 왜냐하면 각 금융거래정보 변수들의 Markov Blanket이 주어지면 Blanket 밖의 변수는 금융거래정보와 독립이기 때문이다.
- 이 정의에 따라 각 Acceptance-Rejection 알고리즘에 사용한 Bayesian Network를 탐색하여 군집마다의 잉여 고객기본정보를 찾아낼 수 있었고, NUMCHILD는 모든 군집의 잉여 고객기본정보임을 확인했다.
- 따라서, NUMCHILD를 제외하고 앞선 validation을 반복함으로써 잉여정보를 효과적으로 제거할 수 있음을 보인다.

군집 할당 정확도 측정



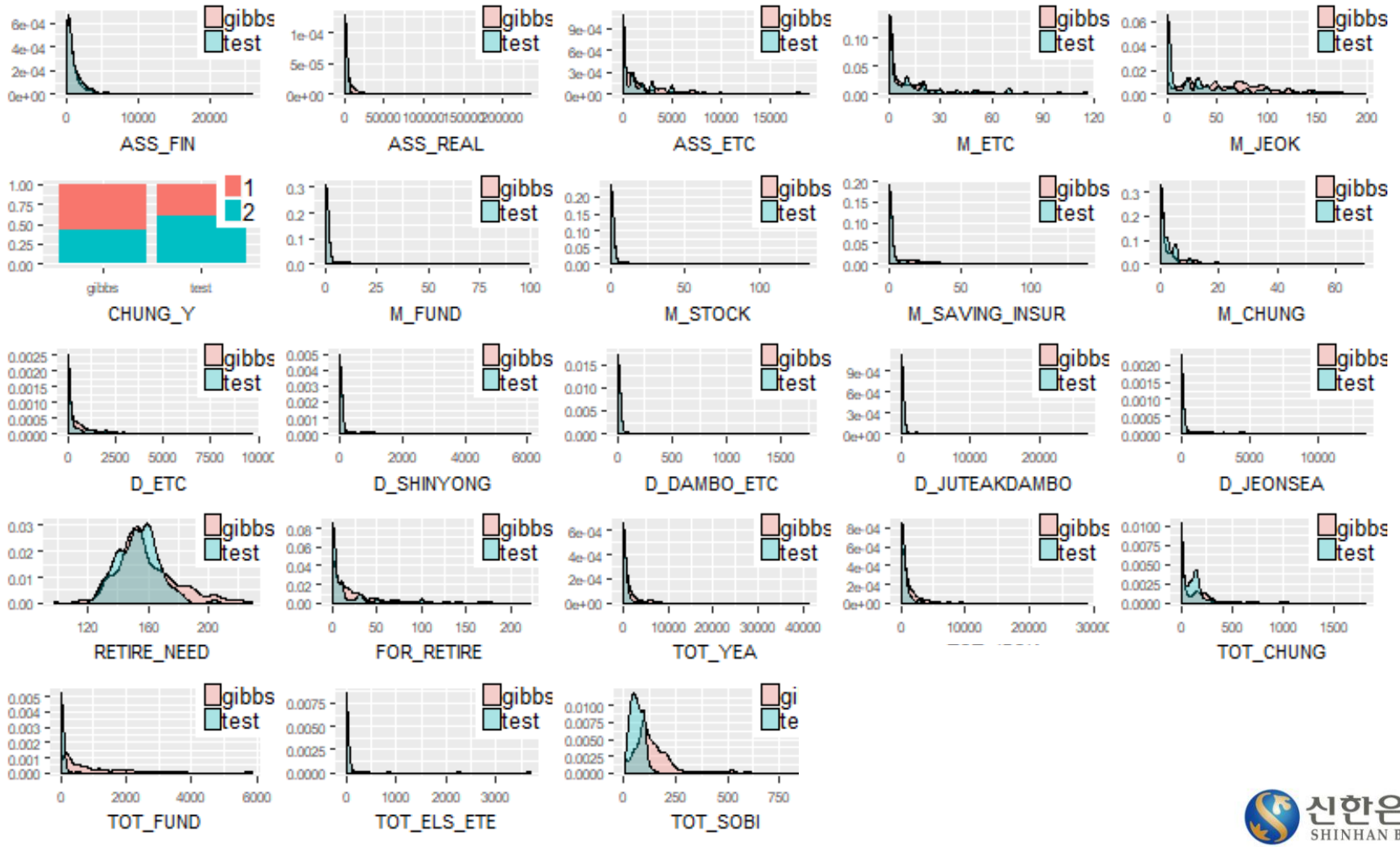
< 군집 별 예측 성공률 >

- 앞선 과정을 반복했을 때, NUMCHILD를 제외했을 때 군집 할당 정확도가 향상함을 확인할 수 있었다.

평균 정확도 86.1%

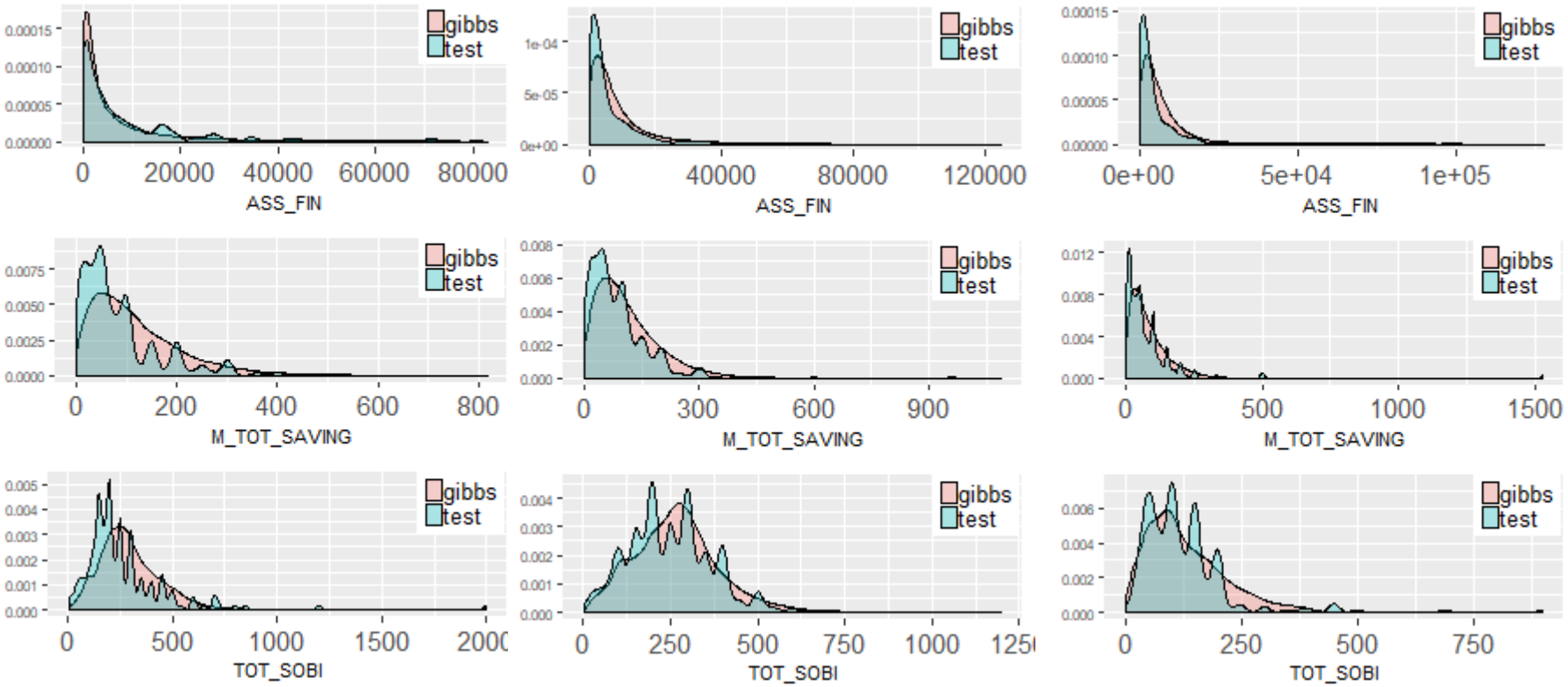
2. 고객유형 별 금융거래정보 분포 유사도

- 또한, 5장과 같은 방식으로 NUMCHILD를 제외하고 Test 데이터 고객유형의 금융거래정보를 추정했을 때, NUMCHILD가 포함되었을 때와 크게 다르지 않은 결과를 얻었다.



3. 군집별 금융자산, 월 저축금액, 월 소비금액 분포 유사도

- 군집별 금융자산, 월 저축금액, 월 소비금액 분포 유사도 또한 NUMCHILD를 제외해도 질적 차이가 없음을 확인할 수 있다.



상위(6군집) 중위(7군집) 하위(11군집)

*군집별 금융자산의 평균 기준

7 결론 및 의의

군집분석 방법으로 PAM 기법을 선택한 이유를 설명하고,
군집별로 적합시킨 Bayesian Network들을 통해
한 고객유형이 각 군집에 속할 확률을 계산해서 고객유형을 군집화한다.

결론 및 의의

- ## 분석의 특징
  1. 표본 데이터를 직접 군집화를 진행함으로써 군집화에 추정오차가 미치는 영향 배제
  2. Bayesian Network 적합을 통해 고객유형이 군집에 할당될 확률을 직접 계산
  3. Gibbs Sampling을 Bayesian Network와 결합하여 부족한 데이터로 인한 제약 해소
  4. Markov Blanket 개념을 활용한 잉여 고객기본정보 탐색

- ## 분석의 의의
  1. 모든 고객유형의 금융거래정보를 추정함으로써 서비스 토대 마련
  2. 잉여 고객기본정보의 탐색을 통한 기본정보 입력에 대한 고객 피로도 감소

감사합니다