Journal of the Korean Data & Information Science Society
2012, **23**(2), 235–245

한국데이터정보과학회지

공통요인분석자혼합모형의 요인점수를 이용한 일반화가법모형 기반 신용평가[†]

임수열[1] · 백장선[2]

[12]전남대학교 통계학과

접수 2012년 1월 11일, 수정 2012년 1월 31일, 게재확정 2012년 2월 15일

요 약

로지스틱판별분석은 금융 분야에서 유용하게 사용되고 있는 통계적 기법으로 신용평가 시 해석이 쉽고 우수한 분별력으로 많이 활용되고 있지만 종속변수에 대한 설명변수들의 비선형적인 관계를 설명하는 부분에는 한계점이 있다. 일반화가법모형은 로지스틱판별모형의 장점과 함께 종속변수와 설명변수 사이의 비선형적인 관계도 설명할 수 있다. 그러나 연속형 설명변수의 수가 대단히 많은 경우 이 두 방법은 모형에 유의한 변수를 선택해야하는 문제점이 있다. 따라서 본 연구에서는 다수의 연속형 설명변수들을 공통요인분석자혼합모형에 의한 차원축소를 통해 변환된 소수의 요인점수들을 일반화가법모형의 새로운 연속형 설명변수로 사용하여 신용분류를 하는 방법을 제시한다. 실제 금융자료를 이용하여 로지스틱판별모형과 일반화가법모형, 그리고 본 연구에서 제안한 방법에 의한 정분류율을 비교한 결과 본 연구에서 제안한 방법의 분류 성능이 더 우수하였다.

주요용어: 공통요인분석자, 로지스틱판별분석, 신용분류, 일반화가법모형.

1. 서론

신용평점제도란 고객들의 신용상태를 점수로 산출하여 대출여부와 대출금액을 결정하는 방법으로 금융관련 문제에 널리 사용되고 있다. 하지만 현재 국내의 신용평가는 전문 심사인력의 부족과 주관적이고 낙후된 심사기법 등으로 인하여 개인의 신용상태에 따라 대출규모, 대출기간 및 이자율이 결정되는 것이 아니라 주로 담보대출 위주로 이루어지고 있다. 따라서 신용평점제도를 도입하여 고객들에 대한 과학적이고 효율적인 신용관리가 이루어져야 한다 (구자용 등, 2005). 개인 차원에서는 신용관리의 중요성에 대한 인식의 제고가 필요하며, 제도적인 차원에서는 개인 및 기업의 신용을 엄밀하게 평가하는 기관 및 제도적 뒷받침이 활발하게 이루어져야 한다. 최근 여러 금융기관에서는 개인의 신용위험 관리에 대한 필요성을 인지하고 이를 관리할 수 있는 선진화된 신용평가 방법들을 도입하여 금리 및 이자의 적용에 차등을 두어 실행을 하고 있지만, 우리나라의 금융시장 환경을 고려한 더욱 적합하고 다양한 신용평가 모형들을 개발하여 현실적인 수준의 신용평가가 이루어 질 수 있도록 노력하여야 한다 (한성실과 정기문, 2004).

개인의 신용상태 (우량/불량)를 평가하는 방법으로서 다양한 판별분석 방법들이 사용된다. 가장 일반적인 통계적 판별분석 방법인 선형판별분석과 이차판별분석의 경우 설명변수들이 다변량정규분포를 따

[†] 이 논문은 2008년도 전남대학교 학술연구비 지원에 의하여 연구되었음.
[1] (500-757) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과, 박사과정.
[2] 교신저자 : (500-757) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과, 교수.
E-mail: jbaek@jnu.ac.kr.

른다고 가정을 하고 있지만 실제 자료의 경우 설명변수가 연속형 변수와 이산형 변수의 혼합으로 되어 있는 등 다변량정규분포의 가정이 적합하지 않는 경우가 흔히 있다. 로지스틱판별분석의 경우 설명변수들의 분포 형태에 대한 가정을 요구하지 않기 때문에 다변량정규분포의 가정에 대한 문제점을 해결할 수 있는 하나의 방법이 된다 (Press와 Wilson, 1978; Berkson, 1951). 또한 종속변수와 설명변수들이 선형결합으로 이루어져 있으며 해석이 쉽고 분별력이 우수하다는 장점을 가지고 있다. 로지스틱판별분석을 이용한 국내 신용평가 연구로는 홍종선과 정민섭 (2011)이 있다. 하지만 종속변수와 설명변수 사이에 비선형 관계가 존재하는 경우, 일반화가법모형 (Generalized Additive Model; GAM)을 이용하게 되면 로지스틱판별분석보다 우수한 예측력을 갖게 된다. GAM은 로지스틱판별모형이 갖는 장점과 함께 종속변수에 대한 설명변수들의 비선형적인 관계까지도 설명이 가능하지만, 로지스틱판별분석과 마찬가지로 설명변수의 수가 많을 경우 실제 모형에 유의한 변수를 선택해야하는 문제를 해결하지 못할 경우 과대적합 (overfitting)의 문제가 발생할 수 있다 (기승도와 강기훈, 2010).

따라서 본 연구에서는 모형에 적합한 변수를 선택하는 방법 대신 다수의 연속형 설명변수의 정보를 소수의 특징변수로 축약할 수 있는 방법으로서 Baek 등 (2010)이 제안한 공통요인분석자혼합모형 (Mixtures of Common Factor Analyzers; MCFA)을 이용하여 고차원의 연속형 설명변수 대신에 새로운 저차원의 요인점수를 GAM의 새로운 특징변수로 사용함으로써 각 판별모형에 따른 분류 성능을 비교하였다.

2. 신용평가 모형에 대한 고찰과 제안된 방법

집단이 두 개이고 p-차원 특징변수 $\boldsymbol{x} = (x_{1,}x_2, \cdots, x_p)'$를 갖는 경우 로지스틱판별모형은 다음과 같다. 만약 첫 번째 집단에 대한 밀도함수가 $\pi(\boldsymbol{x})$이고, p_i가 i번째 집단의 사전확률이라고 하면 로그오즈는 다음과 같은 선형모형을 따른다 ($i = 1, 2$).

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

이 때 새로운 관측치 $\boldsymbol{x}$에 대하여

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) > \ln\left(\frac{p_1}{p_2}\right).$$

을 만족하면 이 새로운 관측치 $\boldsymbol{x}$는 집단 g_1으로 분류하게 된다.

다음은 집단이 수가 두 개이고 설명변수들이 l개의 선형관계의 설명변수들과 m개의 비선형관계의 설명변수들로 이루어져 있고, 비선형적인 관계가 각각 매끄러운 함수 $S_{l+j}(x_{l+j})$로 설명될 수 있다면 GAM의 로그오즈는 다음과 같이 표현된다 ($j = 1, \cdots, m$).

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_l x_l + S_{l+1}(x_{l+1}) + \cdots + S_{l+m}(x_{l+m}).$$

로지스틱판별분석과 비교할 때 GAM의 특징은 로그오즈 모형에서 선형관계인 설명변수는 로지스틱판별분석과 같이 선형함수 그대로 사용하지만, 비선형관계인 설명변수에 대해서는 평활함수추정방법 (smoothing function estimation)을 이용하여 비선형관계식을 추정할 수 있게 함으로써 종속변수의 값을 예측하는데 있어서 효율적인 분류 방법이 된다는 점이다. 또한, 만약에 두 개의 설명변수 x_i와 x_j가 강한 상관관계를 갖고 있다면 로그오즈 모형에 $S(x_i, x_j)$항을 추가할 수 있으며 이를 2차원 평활함수추정방법을 이용하여 추정할 수 있다.

집단이 두 개인 경우 Press와 Wilson (1978)은 정규성의 가정이 위반된 자료에 대하여 로지스틱판별모형이 선형판별모형보다 설명력이 높은 것으로 결론을 내렸다. 또한, Brooks 등 (1988)은 설명변수들이 비정규분포를 갖는 경우 로지스틱모형과 선형판별모형을 비교한 결과 설명변수들의 분포를 정규분포라고 가정해야 할 특별한 이유가 없거나, 설명변수들과 종속변수 사이에 강한 통계적인 관계가 성립되면 로지스틱판별모형이 선형판별모형보다 더 효율적이라고 제안하였다. 따라서 GAM은 설명변수들이 연속형이고 정규성의 가정을 크게 위반한 경우나 설명변수들이 이산형과 연속형으로 되어 있는 경우, 그리고 설명변수 중 비선형인 설명변수가 존재하는 경우 널리 사용될 수 있는 판별분석모형이다.

2.1. 제안된 방법

로지스틱판별모형과 GAM은 설명변수가 많은 경우 실제로 어느 변수가 모형에 유의한지에 대한 변수 선택의 문제점이 발생하게 되고, 이런 문제점을 피할 수 있는 방법으로서 요인분석을 사용할 수 있다. 요인분석은 단일 모집단 내 서로 상관되어 있는 변수들 사이의 복잡한 구조를 잠재적인 공통인자를 이용하여 설명할 수 있는 통계적 기법이지만, 만약 집단의 수가 많고 각 변수들 사이의 상호작용 구조가 다른 경우 이를 반영할 수 없는 문제점을 갖게 된다. 이에 비하여 Baek 등 (2010)이 제안한 MCFA는 집단별로 상이한 분포를 갖는 인자와 공통 요인적재행렬을 이용하여 집단별로 상이한 요인분석 모형을 구축한 것이다.

MCFA는 각 성분 (집단)별로 공통의 요인적재행렬을 가정하며, 요인벡터는 서로 다른 평균과 공분산을 갖는 요인모형을 가정한다. 즉, g개의 집단이 존재하는 경우 고차원 (p-차원) 변수벡터 $\boldsymbol{X}_j$가 i번째 그룹에 속해 있다고 한다면, 그것이 다음과 같이 p-차원 보다 훨씬 작은 저차원 (q-차원)의 요인벡터 $\boldsymbol{U}_{ij}$에 의한 요인분석 모형을 따른다고 가정한다.

$$\boldsymbol{X_j} = \boldsymbol{AU_{ij}} + \boldsymbol{e_{ij}},\ i = 1, \cdots, g,\ j = 1, \cdots, n. \tag{2.1}$$

이 때 $\boldsymbol{U}_{ij}$는 평균이 $\boldsymbol{\xi}_i$이고 공분산행렬이 $\boldsymbol{\Omega}_i$인 q-차원 다변량정규분포 $\boldsymbol{N}_q(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i)$를 따른다. 즉 요인벡터 $\boldsymbol{U}_{ij}$는 일반적으로 $\boldsymbol{N_q}(\boldsymbol{0}, \boldsymbol{I})$를 따르는 직교요인과 다르게 집단별로 평균벡터도 다르고 요인들 간 서로 상관관계가 존재하는 구조를 가지고 있다. $\boldsymbol{A}$는 $p \times q$차원의 공통요인적재행렬이다. 오차벡터 $\boldsymbol{e}_{ij}$는 $\boldsymbol{U}_{ij}$와 독립이며 대각행렬 $\boldsymbol{D}$를 공분산 행렬로 가지는 $\boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{D})$를 따른다. MCFA에 따르면 성분공분산행렬은 $\boldsymbol{\Sigma}_i = \boldsymbol{A\Omega}_i\boldsymbol{A} + \boldsymbol{D}$로 표현되며, $\boldsymbol{A}$는 $\boldsymbol{A'A} = \boldsymbol{I}$를 만족한다. 훈련자료의 수가 매우 적거나 그룹의 수가 매우 많은 경우, 서로 다른 요인적재행렬을 가지고 있다면 추정해야하는 모수의 수가 매우 많아지기 때문에 추정의 정확도가 떨어질 수 있으므로 일반적인 요인분석자혼합모형 (Ghahramani와 Hinton, 1996)과는 다르게 동일한 요인적재행렬을 가정하고 있다. 따라서 $\boldsymbol{x}_j$가 식 (2.1)을 만족하는 경우 $\boldsymbol{x}_j$의 분포는 식 (2.2)의 혼합정규분포모형을 따르게 된다.

$$f(\boldsymbol{x_j}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \boldsymbol{\pi}_i \boldsymbol{\Phi}(\boldsymbol{x}_j; \boldsymbol{A\xi}_i, \boldsymbol{A\Sigma}_i\boldsymbol{A'} + \boldsymbol{D}). \tag{2.2}$$

이 경우 모수는 EM (Expectation-Maximization) 알고리즘에 의하여 추정을 하고, 이렇게 모수의 추정이 끝나고 얻은 저차원의 $\boldsymbol{U}_{ij}$들은 각 그룹별로 해당 $\boldsymbol{x}_j$에 대하여 $\widehat{\boldsymbol{U}}_{ij} = \widehat{\boldsymbol{\xi}}_i + \widehat{\boldsymbol{\gamma}}_i'(\boldsymbol{x}_j + \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\xi}}_i)$로서 추정이 가능하게 된다. 이 때 $\boldsymbol{\gamma}_i = (\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Omega}}_i\widehat{\boldsymbol{A}}' + \widehat{\boldsymbol{D}})^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Omega}}_i$이다 (Baek 등, 2010). 이렇게 추정된 새로운 저차원의 특징벡터 $\widehat{\boldsymbol{U}}_{ij}$을 이용하여 판별분석을 실시하게 된다. 즉, MCFA를 이용하여 실제 신용평가 자료에 대하여 고차원의 연속형 설명변수에 대하여 저차원의 요인점수를 계산하여 (주로 1차원 또는 2차원) 새로운 특징변수로 변환함으로써 로지스틱판별분석과 GAM의 경우 변수 선택을 수행해야 하는 어려움을 피하고 새로운 저차원의 특징변수를 GAM의 연속형 설명변수로 사용한다. 신용평가를 위하여 본 연구에서 제안된 방법을 정리하면 다음과 같다.

1. 전체자료를 훈련자료와 검증자료로 무작위로 나눈 후, 훈련자료에 대하여 $q = 1$부터 시작하여 $q = 2, 3$등으로 한 단계씩 증가하여 MCFA를 적용하여 모수들을 추정한다.

2. 각각의 q에 대응하는 MCFA 추정모수를 이용하여 검증자료에 대한 요인점수벡터 추정치 $\widehat{\boldsymbol{U}}_{ij} = \widehat{\boldsymbol{\xi}_i} + \widehat{\boldsymbol{\gamma}_i}'(\boldsymbol{x_j} + \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\xi}_i})$를 계산한다.

3. 각각의 q에 대응하는 추정된 요인점수벡터들을 GAM에 투입하여 검증자료를 분류하고 정분류율을 계산한다.

4. 가장 높은 정분류율을 도출한 q값에 대응한 MCFA 모수추정치들을 저장하고, 향후 분류하려는 새로운 개체에 대한 요인점수벡터를 추정한 후 이를 GAM에 적용하여 분류한다.

3. 실제 금융회사의 신용평가 자료를 이용한 실험

본 절에서는 실제 금융자료에 대하여 일반적인 로지스틱판별분석과 GAM, 그리고 MCFA를 통해 구한 저차원의 요인점수를 이용한 GAM의 신용분류에 따른 정분류율의 차이를 비교하여 성능의 차이를 검증한다. 실험에서 사용된 자료는 Baesens 등 (2003)에서 사용된 호주 금융회사와 독일 금융회사의 실제 신용평가 자료이다. 특히 호주 금융자료의 경우 총 37개의 결측값이 존재하고 있으며 연속형 설명변수의 결측값은 각 변수의 평균, 범주형 설명변수의 결측값은 각 변수의 최빈값으로 대체하여 사용하였다. 또한 호주 금융자료의 경우 원자료 자체가 정보보호를 위하여 변수에 대한 이름 및 속성 등이 포함되어 있지 않지만, 독일 금융자료의 경우 변수에 대한 정보가 공개되어 있다. 본 연구에서는 $q = 1, 2$일 때의 MCFA 요인점수를 GAM에 이용하여 성능을 비교하였는데, 그 이유는 $q \geq 3$에 대한 요인점수들을 사용한 분류성능보다 더 우수하였기 때문이다.

본 연구의 목적은 원자료를 이용한 로지스틱판별모형과 일반화가법모형 (GAM), 그리고 본 연구에서 제안한 MCFA를 통해 얻은 q-차원의 요인점수를 이용한 일반화가법모형 (GAM)의 분류 성능을 비교하여 제안된 방법의 우수성을 밝히는데 있다. 이를 위하여 호주 금융자료의 경우 총 690개의 자료 중에서 무작위로 각각 490개와 200개를 훈련자료와 검증자료로서, 독일 금융자료의 경우 총 1,000개의 자료 중 무작위로 각각 700개와 300개를 훈련자료와 검증자료로 나누었으며, 이렇게 나뉜 훈련 자료와 검증 자료에 대하여 로지스틱판별분석, GAM, 그리고 MCFA를 이용한 GAM 모형에 적용하여 정분류율을 구하였다.

각 실험마다 동일한 수의 훈련자료와 검증자료의 추출을 100번 반복 시행하였으며, 100번 반복 추출 후 각 실험마다 얻게 되는 정분류된 신용분류 (우량/불량)의 평균을 정분류율로 사용하였다. 또한 짝비교 $t-$검정을 통해 각각의 두 방법들 사이의 분류성능을 비교하였다.

3.1. 호주 금융자료에 대한 고찰

호주 금융자료는 총 690명의 개인에 대한 6개의 연속형 설명변수 $(X1, X2, \cdots, X6)$와 2개에서 14개까지의 범주를 갖는 8개의 범주형 설명변수 $(X7, X8, \cdots, X14)$로 이루어져 있다. 연속형 설명변수들에 대하여 커널밀도함수를 추정해보면 모든 연속형 설명변수들의 분포들이 오른쪽으로 기울어져 있다는 특징을 알 수 있었다. 또한 표 3.1은 연속형 설명변수에 대한 상관계수행렬로서 $(X1, X6)$, $(X3, X6)$, $(X4, X6)$, $(X5, X6)$을 제외하고는 서로 상관관계가 존재하기 때문에 적절한 연속형 설명변수에 대한 선택이 필요함을 알 수 있다.

로지스틱회귀분석의 변수선택 방법 중 단계적 선택법 (stepwise)을 이용하여 모형을 선택한 결과 연속형 설명변수 중 $X4$, $X5$, $X6$, 범주형 설명변수 중 $X8$, $X9$, $X11$, $X12$, $X14$가 모형에 유의하였다. 따라서 로지스틱판별분석의 경우 로지스틱회귀분석을 통해 모형에 유의했던 연속형 설명변수 3개와 범주형 설명변수 5개를 분석에 사용하였다. GAM은 로지스틱판별분석의 설명변수 중 비선형관계의 설명

표 3.1 호주 금융자료의 연속형 설명변수에 대한 상관계수행렬

	X1	X2	X3	X4	X5	X6
X1	1.00000	0.20127***	0.39279***	0.18557***	-0.07716*	0.01854
X2		1.00000	0.29887***	0.27118***	-0.22232***	0.12310**
X3			1.00000	0.32232***	-0.07639*	0.05134
X4				1.00000	-0.11981**	0.06369
X5					1.00000	0.06561
X6						1.00000

*: 유의확률 < 0.05 , **: 유의확률 < 0.01 , ***: 유의확률 < 0.001

변수가 존재할 경우 매끄러운 함수로 관계식을 표현하고 적절한 평활함수추정방법을 이용하여 비선형 관계식을 추정함으로써 로지스틱판별모형에서는 고려할 수 없는 비선형 설명변수의 특성까지 고려하는 통계적 분류기법이다. 모형에 선택된 설명변수들이 일반화선형모형을 만족하는지 검정하는 하나의 방법으로서 잔차의 누적합에 근거한 GENMOD 방법이 있다 (Lin 등, 2002). 만약 설명변수가 종속변수와 선형의 관계에 있다면 잔차의 누적합은 특별한 경향을 보이지 않으며 0에 가까운 값들을 갖는 형태가 된다.

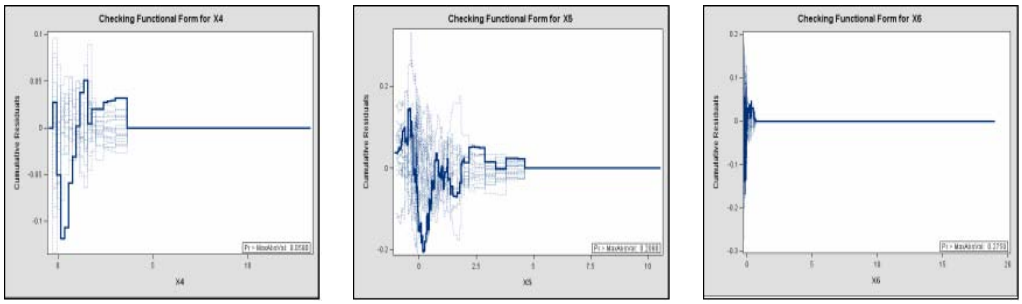

그림 3.1 호주 금융자료에 대한 GENMOD-절차 결과

그림 3.1은 GENMOD 절차에 의하여 로지스틱회귀분석에서 모형에 유의했던 연속형 설명변수 3개에 대하여 각 1,000번의 모의실험을 한 결과이며, 굵은 실선은 잔차누적합의 경로를 나타내고 밝은 점선은 예측값들의 경로를 나타내고 있다. 여기서 3개의 연속형 설명변수 $X4$, $X5$, $X6$모두 잔차의 누적합이 0에서 크게 벗어나고 있으므로 비선형 변수로 고려할 수 있다. 따라서 원자료를 이용한 GAM의 경우 로지스틱판별분석에서 모형에 유의하였던 연속형 설명변수 3개는 평활함수를 이용하여 비모수적 방법을 사용하였으며, 범주형 설명변수 5개는 모수적 방법을 이용하였다.

MCFA는 로지스틱판별분석과 GAM을 사용할 때 발생하는 변수 선택의 문제점을 피할 수 있는 방법이다. 또한, 고차원의 설명변수를 새로운 저차원의 특징변수로 변환할 수 있어 차원축소의 장점도 가지고 있다. 따라서 6개의 연속형 설명변수 전체를 MCFA에 적용하여 새로운 1차원 ($q = 1$) 또는 2차원 ($q = 2$)의 요인점수를 구할 수 있으며 각각의 모수 추정치들은 표 3.2와 같다.

표 3.2에서 $q = 1$에서의 모수추정치에서 원래의 변수 $X1 \sim X6$에 대응하는 공통요인적재행렬 $\widehat{\boldsymbol{A}}$의 열을 검토해보면 $X5$는 음의 계수, 나머지 변수는 양의 계수가 대응됨을 확인할 수 있다. 식 (2.1)의 양변에 $\widehat{\boldsymbol{A}}'$을 곱하면 $\widehat{\boldsymbol{A}}'\boldsymbol{X}_j \approx \widehat{\boldsymbol{U}}_j$이므로, 요인점수는 공통요인적재행렬 계수 값에 의한 원 변수들의 선형결합이므로 종합적인 특성을 나타낸다고 할 수 있다. 따라서 $q = 1$인 경우 1차원 요인점수 $\widehat{U}$는 $X5$변수와 나머지 변수그룹의 대비 특성을 나타낸다고 해석할 수 있다. 또한 $q = 2$일 때 모수추정치의 공통

요인적재행렬 $\widehat{\boldsymbol{A}}$의 첫 번째 요소 열을 검토해보면 $X1 \sim X6$에 양의 계수가 대응됨을 확인할 수 있다. 그러므로 2차원 요인벡터 $\boldsymbol{U} = (U_1, U_2)'$의 첫 번째 요인점수 $\widehat{U}_1$은 $X1 \sim X6$의 종합적 특성을 나타낸다고 해석할 수 있으며, $\widehat{\boldsymbol{A}}$의 두 번째 열을 살펴보면 $X1 \sim X4$는 음의 계수가, $X5, X6$은 양의 계수가 대응되므로 두 번째 요인접수 $\widehat{U}_2$는 $X1 \sim X4$ 변수그룹과 $X5, X6$ 변수그룹의 대비 특성을 나타낸다고 해석할 수 있다. 이렇게 추정된 요인점수를 설명변수로 사용한 GAM 모형이 자료에 적합된 후 추정된 요인의 기울기가 양/음 이면 그 요인은 해당 개체가 우량/불량 집단으로 분류되는데 공헌하는 것으로 해석된다. 그러므로 유의한 요인점수는 신용평점을 위한 평점표로서 활용될 수 있겠다.

표 3.2 호주 금융자료에 대한 MCFA의 모수 추정치

1차원 ($q = 1$)모수추정치
$\widehat{\boldsymbol{\xi}}_1' = (-0.5553, 1.0895)$,
$\widehat{\boldsymbol{\Omega}}_1 = \begin{pmatrix} 0.0011 \\ 2.1441 \end{pmatrix}$,
$\widehat{\boldsymbol{D}} = diag(0.8269, 0.9064, 0.04039, 0.8990, 0.99143, 0.9957)$,
$\widehat{\boldsymbol{A}}' = \begin{pmatrix} 0.3590 & 0.2631 & 0.8483 & 0.2734 & -0.0731 & 0.0461 \end{pmatrix}$
2차원 ($q = 2$) 모수추정치
$\widehat{\boldsymbol{\xi}}_1' = (-0.33, 0.42)$, $\widehat{\boldsymbol{\xi}}_2' = (0.34, -0.43)$,
$\widehat{\boldsymbol{\Omega}}_1 = \begin{pmatrix} 0.009 & 0.0007 \\ 0.0007 & 0.0006 \end{pmatrix}, \widehat{\boldsymbol{\Omega}}_2 = \begin{pmatrix} 2.18 & 0.03 \\ 0.03 & 2.05 \end{pmatrix}$,
$\widehat{\boldsymbol{D}} = diag(0.49, 0.89, 0.83, 0.16, 0.97, 0.0007)$,
$\widehat{\boldsymbol{A}}' = \begin{pmatrix} 0.09 & 0.20 & 0.17 & 0.33 & 0.02 & 0.90 \\ -0.20 & -0.22 & -0.33 & -0.78 & 0.15 & 0.42 \end{pmatrix}$

원래의 6차원 설명변수 벡터를 $\boldsymbol{X}_j$라고 할 때, GAM에 사용될 새로운 요인점수 벡터는 $\widehat{\boldsymbol{U}}_{ij} = \widehat{\boldsymbol{\xi}}_i + \widehat{\gamma}_i'(\boldsymbol{x}_j + \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\xi}}_i)$로서 추정이 가능하게 되며 $\boldsymbol{\gamma}_i = (\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Omega}}_i\widehat{\boldsymbol{A}}' + \widehat{\boldsymbol{D}})^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\Omega}}_i$이다.

3.2. 독일 금융자료에 대한 고찰

독일 금융자료는 총 1,000명의 개인에 대한 20개의 설명변수들로 이루어져 있는데, 설명변수들은 7개의 연속형 변수와 2개에서 10개까지의 범주를 갖는 13개의 범주형 변수들로 이루어져있다. 연속형 설명변수들에 대한 커널밀도함수를 추정해보면 연속형 설명변수 중 $X1, X2, X5$의 분포는 오른쪽으로 기울어져 있으며 $X3, X4, X6$은 조사된 자료 값들이 세 개 이상으로 연속형 설명변수로 간주할 수 있지만, $X7$의 경우 조사된 자료 값들이 두 개로서 연속형 설명변수가 아닌 범주형 설명변수로 고려되었다.

표 3.3은 연속형 설명변수에 대한 상관계수행렬로서 $(X1, X2)$, $(X1, X3)$, $(X2, X3)$, $(X4, X5)$, $(X4, X6)$, $(X5, X6)$의 경우 서로 상관관계가 존재하고 있기 때문에 독일 금융자료 역시 적절한 연속형 변수들에 대한 선택이 필요함을 알 수 있다.

표 3.3 독일 금융자료의 연속형 설명변수에 대한 상관계수행렬

	X1	X2	X3	X4	X5	X6
X1	1.00000	0.62484***	0.07483*	0.03410	-0.03621	-0.01124
X2		1.00000	-0.27138***	0.02891	0.03260	0.02068
X3			1.00000	0.04925	0.05839	0.02165
X4				1.00000	0.26660***	0.08954**
X5					1.00000	0.14935***
X6						1.00000

*: 유의확률 < 0.05 , **: 유의확률 < 0.01 , ***: 유의확률 < 0.001

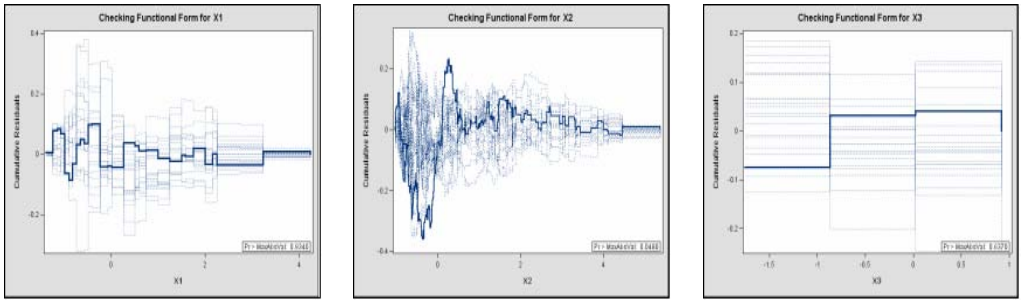

그림 3.2 독일 금융자료에 대한 GENMOD-절차 결과

로지스틱회귀분석의 변수선택 방법 중 단계적 선택법을 이용하여 모형을 선택한 결과 연속형 설명변수 중 $X1 \sim X3$의 변수가 모형에 유의하였으며 범주형 설명변수 중 $X8 \sim X11, X13, X14, X16, X20$의 변수가 로지스틱판별분석 모형에 유의한 변수임을 알 수 있었다. 따라서 로지스틱판별분석의 경우 연속형 설명변수 3개와 범주형 설명변수 8개를 사용하여 분석에 사용하였다. 또한 그림 3.2는 GENMOD 절차에 의한 결과로서 3개의 연속형 설명변수 중 $X2$는 비선형 변수로 판단되며 적절한 평활함수가 필요함을 알 수 있다. 즉, 원자료를 이용한 GAM의 경우 로지스틱판별분석에서 모형에 유의하였던 연속형 설명변수 3개 중에서 $X2$에 대해서만 평활함수를 이용한 비모수적 추정방법을 사용하여 해당 회귀함수를 추정하였으며, 범주형 설명변수 8개는 모수적 방법을 이용하였다.

다음의 표 3.4는 독일 금융자료를 MCFA에 $q = 1$과 $q = 2$일 때 각각 적용하여 구한 모수의 추정치이다. $q = 1$일 때 원래의 연속형 설명변수 $X1 \sim X6$에 대응하는 6×1차원 공통요인적재행렬 추정치 $\widehat{\boldsymbol{A}}$의 열을 검토해보면 $X3$은 음의 계수, 나머지 변수는 양의 계수가 대응됨을 알 수 있다. 그러므로 요인점수 U는 개인의 부채 상환능력과 기타 금융거래 조건과의 대비라고 볼 수 있다. $q = 2$일 때 $\widehat{\boldsymbol{A}}$의 첫 번째 요소 열을 검토해보면 $X2, X5, X6$은 양의 계수, $X1, X3, X4$는 음의 계수가 대응된다. 그러므로 첫 번째 요인점수 U_1은 개인의 부채상환 의지와 개인의 금융거래 신용도와의 대비라고 볼 수 있다. $\widehat{\boldsymbol{A}}$의 두 번째 열을 살펴보면 $X5, X6$은 양의 계수, $X1 \sim X4$는 음의 계수가 대응되므로 두 번째 요인점수 U_2는 개인의 금융거래 신용도와 기타 금융거래 환경과의 대비라고 볼 수 있다.

표 3.4 독일 금융자료에 대한 MCFA의 모수 추정치

1차원 ($q = 1$)모수추정치
$\widehat{\boldsymbol{\xi}}_1' = (-0.1243, 0.29)$,
$\widehat{\boldsymbol{\Omega}}_1 = \begin{pmatrix} 1.0468 \\ 2.2729 \end{pmatrix}$,
$\widehat{\boldsymbol{D}} = diag(0.604, 0.0177, 0.9269, 0.9981, 0.9980, 0.9986)$,
$\widehat{\boldsymbol{A}}' = \begin{pmatrix} 0.5218 & 0.8225 & -0.2229 & 0.0244 & 0.0265 & 0.017 \end{pmatrix}$
2차원 ($q = 2$) 모수추정치
$\widehat{\boldsymbol{\xi}}_1' = (0.0123, -0.0286,\ \widehat{\boldsymbol{\xi}}_2' = (0.1745, - - 0.1071)$,
$\widehat{\boldsymbol{\Omega}}_1 = \begin{pmatrix} 0.3523 & -0.1026 \\ -0.102 & 1.2655 \end{pmatrix}, \widehat{\boldsymbol{\Omega}}_2 = \begin{pmatrix} `1.0114 & -0.4520 \\ -0.4520 & 2.0294 \end{pmatrix}$,
$\widehat{\boldsymbol{D}} = diag(0.0637, 0.0087, 0.818, `0.9976, 0.9924, `0.9975)$,
$\widehat{\boldsymbol{A}}' = \begin{pmatrix} -0.5012 & 0.6351 & -0.5745 & -0.0094 & 0.1120 & 0.0535 \\ -0.7812 & -0.6232 & -0.0028 & -0.0301 & 0.0189 & 0.0043 \end{pmatrix}$

4. 실제 금융회사의 신용평가 자료를 이용한 실험 결과 비교

지금부터는 금융 원자료를 이용한 로지스틱판별분석, GAM, 그리고 MCFA를 이용하여 구한 요인점수 (1차원 또는 2차원)를 이용한 GAM 등 3가지 판별모형에 대한 분류성능 결과를 알아본다.

4.1. 호주 금융자료에 대한 실험 결과

표 4.1은 호주금융자료에 대한 검증자료의 신용분류 (우량/불량)에 대한 정분류율의 결과로서 로지스틱판별분석과 GAM의 경우 정분류율의 차이가 거의 없다. 또한 1차원과 2차원 요인 점수를 이용한 GAM 역시 정분류율의 차이는 거의 없다. 하지만 1차원 요인점수와 2차원 요인점수를 이용한 GAM의 경우 원자료를 이용한 로지스틱판별분석과 GAM의 정분류율 보다 더 높은 분류 성능을 보인다.

표 4.1 호주 금융자료에 대한 정분류율

방법	정분류율
Logistic	85.96 %
GAM	85.43 %
GAM(1 factor score)	90.96 %
GAM(2 factor score)	89.05 %

표 4.2는 호주금융자료에 대한 각 방법별 정분류율의 분류성능에 대한 짝 비교 $t-$검정 결과이다.

표 4.2 호주금융자료에 대한 $t-$검정 결과

	t값 (표준오차)	유의확률
GAM - Logistic	-2.83 (0.395)	0.0057
GAM(1 factor) - Logistic	7.99 (1.23)	<.0001
GAM(2 factor) - Logistic	2.33 (2.47)	0.1219
GAM(1 factor) - GAM	9.04 (1.23)	<.0001
GAM(2 factor) - GAM	2.75 (2.44)	0.00072
GAM(2 factor) - GAM(1 factor)	-1.49 (2.69)	0.1391

표 4.2의 결과를 보면 MCFA를 이용하여 구한 요인점수를 이용한 GAM의 경우 원자료를 이용한 로지스틱판별모형과 GAM보다 우수한 분류성능을 보이고 있음을 알 수 있다. 하지만 1차원 요인점수와 2차원 요인점수를 이용한 GAM의 경우 평균 정분류율의 차이는 없다. 즉 MCFA를 이용하여 새로운 저차원의 요인점수를 이용할 경우의 분류성능이 원자료를 이용한 모형보다 더 우수한 분류성능을 갖는다고 판단된다.

4.2. 독일 금융자료에 대한 실험 결과

표 4.3은 독일 금융자료에 대한 검증자료의 신용분류 (우량/불량)에 대한 정분류율의 결과이다. 로지스틱판별분석과 GAM을 비교한 경우 GAM의 분류율이 로지스틱판별분석에 비하여 모두 더 높은 분류율을 보인다. 하지만 1차원 요인점수를 이용한 GAM의 경우 호주 금융자료와는 다르게 원자료를 이용한 GAM보다 오히려 낮은 분류율을 보이고 있다. 하지만 2차원 요인점수를 이용한 GAM의 경우 다른 세 가지 분류모형보다 매우 우수한 분류 성능을 나타낸다. 이 때, 2차원 요인점수 GAM 모형은 두 요인점수추정치들 사이에 상관관계가 존재하였으므로 2차원 평활스플라인을 적용하여 비선형 회귀함수를 추정하였다.

독일 금융자료에 대한 각 방법별 정분류율의 분류성능에 대한 짝 비교 $t-$검정 결과는 표 4.4와 같다.

표 4.3 독일 금융자료에 대한 정분류율

방법	정분류율
Logistic	62.68 %
GAM	74.43 %
GAM(1 factor score)	73.33 %
GAM(2 factor score)	91.34 %

표 4.4 독일금융자료에 대한 $t-$검정 결과

	t값 (표준오차)	유의확률
GAM - Logistic	39.66 (0.88)	<.0001
GAM(1 factor) - Logistic	29.57 (1.07)	<.0001
GAM(2 factor) - Logistic	105.35 (0.81)	<.0001
GAM(1 factor) - GAM	-4.76 (0.69)	<.0001
GAM(2 factor) - GAM	77.07 (0.65)	<.0001
GAM(2 factor) - GAM(1 factor)	62.35 (0.86)	<.0001

표 4.4에서 확인할 수 있듯이 2차원 요인점수를 이용한 GAM의 경우 다른 모형들에 비하여 유의적으로 매우 우수한 분류 성능을 보이고 있음을 알 수 있다. 1차원의 요인점수를 이용한 GAM의 분류율이 낮은 이유는 연속형 설명변수 중 범주형의 특성을 갖는 변수들의 영향이 있는 것으로 판단된다.

5. 결론 및 토의

실제 신용평가를 하는데 사용되는 자료는 선형판별분석과 이차판별분석에서 가정하는 다변량정규분포를 따르지 않는 경우가 대부분이다. 반면에 로지스틱판별분석은 설명변수들의 선형 결합을 가정하지만, 비선형 설명변수가 존재하는 경우 종속변수를 예측하는데 무리가 따르게 된다. 이때 GAM은 비선형 설명변수도 고려할 수 있는 모형으로 로지스틱판별분석이 갖는 문제점을 해결할 수 있는 방법이 될 수 있지만 로지스틱판별모형과 GAM은 설명변수의 수가 많을 경우 변수 선택의 문제점을 갖게 된다. 따라서 본 연구에서는 이러한 문제점을 해결함과 동시에 연속형 설명변수의 차원 축소도 가능한 공통요인분석자혼합모형 (MCFA)을 이용하여 구한 새로운 저차원의 요인점수를 연속형 설명변수 대신 GAM에 적용하는 새로운 신용평가 모형을 제안하고, 실제 자료를 이용하여 기존의 방법들과 분류 성능을 비교 하였다.

실제 자료인 호주와 독일 금융자료를 이용하여 연속형 설명변수 대신 MCFA를 적용하여 연속형 설명변수를 1차원 요인점수 또는 2차원 요인점수로 변환하여 GAM을 시행하였을 때 개인고객의 신용상태에 대한 정분류율이 원래의 변수 자료를 사용한 로지스틱판별분석과 GAM 보다 모두 우수한 분류 성능을 보였다. 따라서 MCFA를 이용한 실험방법의 분류율이 그렇지 않은 실험방법보다 우수한 분류 성능을 갖는 신용평가 모형임을 알 수 있다.

Baesens 등 (2003)에서 제시된 다양한 모형들과의 정분류율을 비교해 보면 원자료를 이용한 로지스틱판별분석과 GAM보다는 MCFA를 통해 구한 요인점수를 사용할 경우 호주 금융자료는 1차원 요인점수와 2차원 요인점수를 이용한 GAM에서 기존의 분석방법들보다 우수한 분류 성능을 보였다. 또한 독일금융자료의 경우 2차원 요인점수를 이용하여 2차원 평활스플라인을 적용한 GAM에서 매우 우수한 분류 성능을 나타내고 있다. 즉, MCFA를 이용하여 구한 요인점수를 사용할 경우 원자료를 이용하는 판별모형들보다 우수한 분류 성능을 나타낸다고 할 수 있다.

MCFA 모형의 요인의 차원 q는 해당 자료에 대하여 가장 높은 분류성능을 도출하는 값을 선택하여 사용한다. 즉, 본 연구에서와 같이 전체자료를 훈련자료와 검증자료로 무작위로 나눈 후, 훈련자료에 대

하여 $q = 1$부터 시작하여 $q = 2, 3$등으로 한 단계씩 증가하여 MCFA를 적용하여 모수들을 추정한 다음 그것들을 이용하여 검증자료에 대한 각각의 q에 대응하는 요인점수들을 추정한다. 검증자료에 대하여 각각의 q에 대응하는 추정된 요인점수들을 GAM에 투입하여 분류한 후 정분류율을 계산하여 가장 높은 정분류율을 도출한 q값을 선택한다.

마지막으로 MCFA는 고차원 소표본 자료의 군집 및 판별분석에 더욱 유용한 방법으로 제시된 것이지만, 그 자체가 모형기반 (model-based)인 방법으로서 중간차원 혹은 대단위 자료에도 언제나 적용이 가능하다. 본 연구에서 분석한 신용평가 자료는 중간차원의 연속형 설명변수만을 가지고 있었지만 향후 연구에서는 고차원의 신용평가 자료를 적용하여 MCFA의 유용성을 검증하고자 한다.

참고문헌

기승도, 강기훈 (2010). 일반화가법모형에서 축소방법의 적용연구. <응용통계연구>, **23**, 207-218.

구자용, 최대우, 최민성 (2005). 스플라인을 이용한 신용 평점화. <응용통계연구>, **18**, 543-553.

한성실, 정기문 (2004). 로지스틱 회귀모형을 이용한 채택확률모형. <한국자료분석학회>, **6**, 1153-1161.

홍종선, 정민섭 (2011). 신용평가에서 로지스틱회귀를 이용한 미결정자 추론. <한국데이터정보과학회지>, **22**, 149-157.

Baek, J., McLachlan, G. J. and Flack, L. (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1298-1309.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**, 627-635.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics*, **7**, 327-339.

Brooks, C. A., Clark, R. R., Hadgu, A. and Jones, A. M. (1988). The robustness of the logistic risk functions. *Communication in Statistics, Simulation*, **17**, 1-24.

Ghahramani, Z. and Hinton, G. E. (1996). *The EM algorithm for mixture of factor analyzers*, Technical Report CRG-TR-96-1, **8**, University of Toronto, Canada.

Lin, D. Y., Wei, L. J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, **58**, 1-12.

Press, S. R. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, **73**, 669-705.

Journal of the Korean Data & Information Science Society
2012, **23**(2), 235–245

한국데이터정보과학회지

A credit classification method based on generalized additive models using factor scores of mixtures of common factor analyzers[†]

Suyeol Lim[1] · Jangsun Baek[2]

[12]Department of Statistics, Chonnam National University

Received 11 January 2012, revised 31 January 2012, accepted 15 February 2012

Abstract

Logistic discrimination is an useful statistical technique for quantitative analysis of financial service industry. Especially it is not only easy to be implemented, but also has good classification rate. Generalized additive model is useful for credit scoring since it has the same advantages of logistic discrimination as well as accounting ability for the nonlinear effects of the explanatory variables. It may, however, need too many additive terms in the model when the number of explanatory variables is very large and there may exist dependencies among the variables. Mixtures of factor analyzers can be used for dimension reduction of high-dimensional feature. This study proposes to use the low-dimensional factor scores of mixtures of factor analyzers as the new features in the generalized additive model. Its application is demonstrated in the classification of some real credit scoring data. The comparison of correct classification rates of competing techniques shows the superiority of the generalized additive model using factor scores.

Keywords: Credit classification, generalized additive model, logistic regression, mixtures of common factor analyzers.

[†] This study was financially supported by Chonnam National University, 2008.
[1] Ph. D. candidate, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
[2] Corresponding author: Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail : jbaek@jnu.ac.kr.