노이즈 필터링과 충분차원축소를 이용한 비정형 경제 데이터 활용에 대한 연구

유재근*, 서범석**

본 연구는 노이즈 필터링과 차원축소 등의 방법을 이용하여 텍스트 지표의 정상화에 대해 검토하고 실증분석을 통해 동 지표의 활용 가능성을 제고할 수 있는 후처리 과정을 탐색하고자 하였다. 실증분석에 대한 예측 목표 변수로 월별 선행지수 순환변동치, BSI 전산업 매출실적, BSI 전산업 매출전망 그리고 분기별 실질 GDP 계절조정계열 전기비와 실질 GDP 원계열 전년동기비를 상정하고 계량경제학에서 널리 활용되는 Hodrick and Prescott 필터와 비모수 차원축소 방법론인 충분차원축소를 비정형 텍스트 데이터와 결합하여 분석하였다.

분석 결과 월별과 분기별 변수 모두에서 자료의 수가 많은 경우 텍스트 지표의 노이즈 필터링이 예측 정확도를 높이고, 차원축소를 적용함에 따라 보다 높은 예측력을 확보할 수 있음을 확인하였다. 분석 결과가 시사하는 바는 텍스트 지표의 활용도 제고를 위해서는 노이즈 필터링과 차원축소 등의 후처리 과정이 중요하며 이를 통해 경기 예측의 정도를 높일 수 있다는 것이다.

* 이화여자대학교 통계학과 교수 (e-mail: peter.yoo@ewha.ac.kr, phone: 02-3277-6717)
** 한국은행 경제모형실 거시모형팀 과장 (e-mail: bsseo@bok.or.kr, phone: 02-759-4248)
※ 본 연구의 내용은 집필자들의 개인 의견으로 한국은행의 공식견해를 나타내는 것은 아님.

Ⅰ. 서론

경제 분석에 비정형 데이터를 활용하려는 시도가 크게 증가하고 있다. 그러나 통계 공표 기관에서 작성하는 공식 통계와 달리 비정형 데이터는 노이즈를 포함하고 가공 방식에 따라 정보의 양이 달라질 수 있다는 우려가 제기된다. 이는 공식 통계가 엄격한 품질 관리를 통해 기획되는 조사 방법론(experimental study)에 기반을 두는 반면, 비정형 데이터는 목적 없이 기록된 데이터에서 정보를 추출하는 관측 기법(observational study)에 기반을 두기 때문이다. 특히 경제적으로 관심이 높은 비정형 데이터는 음성, 텍스트, 이동(mobility) 정보 등으로, 특정되지 않은 방대한 양의 정보를 포함하는 빅데이터라는 점에서 정보를 가공하는 방식이 정보의 질에 영향을 미칠 것임은 자명하다.

최근의 연구 서범석(2022, 2023)과 Seo et al.(2022)에 의하면 뉴스 기사, 증권사 리포트 등 텍스트 정보가 경제 분석에 매우 유용함을 알 수 있다. 그러나 해당 연구에서 추출한 텍스트 지표들은 변동성이 다소 높게 나타난다. 이는 텍스트 지표 작성을 위해 뉴스 기사 등에 산재해 있는 정보를 추출하고 합성하는 과정에서 지표에 일부 노이즈가 포함되거나 정보가 효율적으로 가공되지 않았을 가능성이 있음을 시사한다.

따라서 본 연구는 노이즈 필터링 등의 기법과 차원축소 등의 방법을 이용하여 텍스트 지표의 정상화에 대해 검토하고 후처리 과정의 실증분석을 통해 텍스트 지표의 활용가능성을 높이고자 하였다. 특히 Hodrick and Prescott(1997) 필터 등을 이용하여 텍스트 지표의 노이즈를 제거한 경우를 비교하고, 특정 조절 모수 값이 국내총생산(Gross Domestic Product, GDP), 기업경기실사지수(Business Survey Index, BSI) 등 경제지표 예측에 보다 유리한 결과를 제공하는지를 실증분석을 통해 검토하였다. 또한 텍스트는 방대한 양의 정보를 포함하므로 다양한 경제 지표를 산출할 수 있는 점을 고려하여, 해당 정보들의 차원축소를 통해 정보를 합성하는 경우 경제지표 예측에 보다 유리한 결과를 제공하는지를 충분차원축소(Sufficient Dimension Reduction, SDR) 등을 통해 검증하였다. 또한 서범석(2023)이 산출한 산업별 네트워크 데이터와 개별 산업 텍스트 지표의 정보량을 비교하고 텍스트에서 추출한 네트워크 데이터를 경제분석에 활용하는 방안을 함께 검토하였다.

텍스트 지표를 통계 모형에 활용하기 위해서는 텍스트에 포함된 광범위한 정보 중 목적변수와 상관관계가 높은 주요 지표를 식별하는 과정이 중요하다. 이를 위해 본 연구에서는 경제 변수와 관련이 높은 정보는 유지하면서, 고차원의 자료를 저차원의 자료로 변환하는 차원축소 방법론을 주요하게 검토하였다. 특히 시계열 데이터의 특성을 감안하여 텍스트

정보가 거시경제 변수에 반영되어 나타나는 시차를 고려할 수 있도록 시점 정보를 고려한 차원축소 방법론을 제시하였다. 또한 노이즈 필터링 기법은 차원축소를 위한 전처리 과정으로 간주하고 노이즈 필터링과 차원축소 방법을 연결하여 분석하였으며, 경기예측력을 기준으로 각각의 방법론들을 평가하였다.

본 연구는 최근 다양하게 연구되고 있는 텍스트 데이터의 필터링과 차원축소 등을 통하여 비정형 데이터의 활용 가능성을 높이는 방안을 다각도에서 검토했다는 점에서 의의가 있다. 또한 본 연구에서 제시한 필터링 정보와 시차를 고려한 차원축소 등의 방법론은 텍스트 이외의 다양한 비정형 경제 데이터에도 비슷하게 적용될 수 있다는 점에서 향후 관련 연구에도 기여할 수 있을 것으로 기대된다.

본 논고는 다음과 같이 구성하였다. 2장에서 관련 선행연구와 연구 배경에 대해 살펴보고 이어지는 3장에서 실증분석에 활용한 데이터를 소개하였다. 4장에서는 적용 방법론을 자세히 설명하였고, 5장에서 실증분석 결과를 제시하였다. 마지막으로 6장에서 본 연구의 시사점과 발전 방향을 정리하였다.

Ⅱ. 관련 선행연구 리뷰와 연구배경

1. 선행연구

텍스트 지표의 유용성에 대한 연구는 최근 활발히 진행되고 있다. 서범석(2022)은 비정형 뉴스 텍스트의 정량화를 통해 경제지표를 작성하기 위한 새로운 텍스트 마이닝 방법론을 제시하고 있다. 특히 다양한 분야의 경제 뉴스 중에서 관심이 높은 생산, 물가, 고용, 주가, 주택 가격 등 15개 분야에 대해 정량화된 텍스트 지표를 제시하고 동적인자모형(Dynamic Factor Model)과 합성곱 순환신경망(Convolution Recurrent Neural Network, CRNN)을 이용하여 텍스트 지표의 경기예측력을 평가하였다. 서범석(2022)의 분석 결과에 따르면 텍스트 지표와 공식 통계를 함께 사용하여 경기 예측 모형을 구성할 경우 GDP의 평균 예측정확도가 향상되는 것으로 나타난다. 해당 선행연구에서 제시한 15개의 텍스트 지표는 월별 지표로 산출되었다.

또한 서범석(2023)은 증권사 기업평가 보고서의 텍스트 자료를 활용하여 산업별, 지역별 업황 분석을 시도하였다. 해당 선행연구에서는 2019년도 이후 증권사 기업평가 보고서 약 13만 건을 빅데이터로 구축하고 이로부터 산업별 업황지수, 산업간 네트워크 지표 등 정량화된 텍스트 지표들을 산출하였다. 증권사 애널리스트 보고서의 텍스트 정보는 시점별 자료의 양, 시계열의 변동성, 계산 편의 등을 고려하여 분기별로 산출되었다. 증권사 애널리스트 보고서의 통계적 분석을 위해 정량 텍스트정보를 추출하는 자세한 전처리 과정과 변환 방법에 대해서는 서범석(2023)의 3절을 살펴보기를 바란다. 해당 선행연구 분석결과에 따르면 업종별 텍스트 업황은 관련 코스피 산업별 지수 및 거시 경제지표와 비슷한 패턴을 보이고, 1분기 내외의 선행성을 갖는 것으로 나타난다. 또한 동 텍스트 지표들은 관련 코스피 산업별 지수에 대해서도 0.5~0.9의 높은 상관관계를 보인다. 한글이 아닌 다른 언어의 텍스트 분석과 관련된 해외 연구사례는 서범석(2023)의 1절에 자세히 소개되어 있다.

한편 텍스트 데이터의 노이즈 필터링 등 정상화에 대한 연구는 아직 시도된 바가 많지 않다(Chen et al. 2021). 이는 경제 분석에 텍스트 데이터가 활용되기 시작한지가 오래되지 않은 점 때문으로 사료된다. 텍스트 데이터 이외의 다양한 금융경제 데이터에 대한 노이즈 필터링 연구를 살펴보면, 실용적 관점에서는 Hodrick and Prescott(1997)의 방법론이 가장 널리 활용되고 있다. HP 필터는 경제 변수의 추세변동과 순환변동을 구분하기 위해 제시된 비모수적 방법으로 다음의 산식을 이용하여 산출한다.

$$\min_{\tau}\left(\sum_{t=1}^{T}(y_t-\tau_t)^2+\lambda\sum_{t=2}^{T-1}[(\tau_{t+1}-\tau_t)-(\tau_t-\tau_{t-1})]^2\right).$$

여기서 y_t는 관측변수를 나타내며 y_t는 추세 요소 τ_t, 순환성 요소 c_t, 및 잔차 요소 ϵ_t의 합으로 표시되는 일반적인 가법 모형에 의해 분해된다. 여기서 추세 요소는 시계열의 장기적인 움직임을 나타내며 순환성 요소는 추세와 관련이 없는 정기적인 패턴을 그리고 잔차 요소는 추세나 순환성 요소로 설명할 수 없는 무작위한 변동을 나타낸다. HP 필터에서 조절 모수값 λ(lambda)는 추세 요소의 변동성을 결정하는 초매개변수(hyperparameter)로 동 조절 모수값이 커질수록 매끄러운(smooth) 추세 요소를 얻을 수 있다. Hodrick and Prescott(1997)과 Ravn and Uhlig(2002)는 월 자료에 대해 129,600의 λ를, 분기 자료에 대해 1,600의 λ를 경험 법칙(heuristics)으로 제시하였다.

그러나 텍스트 지표의 변동성이 여타 금융 지표보다 높게 나타나며, 또한 본 연구의 목적이 순환성 요소를 찾기보다는 노이즈 제거에 있다는 점에서 HP 필터를 이용할 경우 새로운 λ를 탐색하는 과정이 필요할 것으로 판단된다. HP 필터 이외에 Kalman 필터, Hamilton(2018) 필터 등의 방법론도 경제지표의 필터링 방법으로 널리 활용되고 있는데, 추정과정이 HP 필터에 비해 복잡하고 컴퓨팅 시간이 오래 걸리므로 본 연구에서는 고려하지 않았다.

고차원 자료에서 정보를 가공하는 방식으로는 필터링 외에 차원축소의 방법도 고려할 수 있다. 차원축소는 일반적인 회귀 문제에서 예측력을 높이는 데 이바지할 수 있는 것으로 알려져 있다. Yoo(2019)는 초고차원 자료의 대표인 마이크로 어레이 자료에 대한 생존분석에서 유전체 자료를 차원축소를 통하여 콕스-비례 모형을 적합할 때 위험율(hazard ratio) 대한 예측력을 높일 수 있음을 보여주고 있다. 총 7,399의 유전체 정보를 주성분 분석을 통하여 40개로 일차적으로 차원축소를 하였다. 훈련 자료에 대한 표본의 수가 160개이기에 여전히 40개의 변수는 적지 않은 변수의 수이다. 이를 다시 fused sliced inverse regression을 적용하여 1차원으로 축소하여, 최종적으로는 7,399차원의 유전체 자료를 1차원의 자료로 축소하였다. 이렇게 최종 축소된 자료를 이용하여 생존분석에서 가장 보편적으로 많이 사용되는 콕스-비례 모형에 적합하여 실제 예측력이 기존보다 높아졌음을 검정 자료를 통하여 보여주었다. 차원축소가 자료 분석을 보다 정확하게 해 줄 수 있음은 이러한 의생명 분야뿐만 아니라 화학 분야(Lee et al., 2019), 식품 분야(Kim et al., 2017) 등 다양한 분야에서 실증되고 있다. 이러한 충분차원축소 방법론을 본 연구에서는 비정형 경제 데이터인 텍스트 데이터에 적용하였다. 특히 본 연구에서는 sliced inverse regression(Li, 1991), directional regression(Li and Wang, 2007)과 seeded method(Cook et al., 2007)을 고려하고자 한다.

2. 연구배경

COVID-19, 러우전쟁 등 외생적인 경제 이벤트는 소비, 투자 등 경제주체의 경제활동에 영향을 미친다. 그러나 과거 데이터를 기반으로 한 실증분석을 통해서는 경제 이벤트의 정확한 효과를 파악하는 것이 쉽지 않다. 이는 경제적 영향이 큰 이벤트일수록 과거에 유사한 사례를 찾는 것이 어렵기 때문이다. 따라서 경제가 급변하는 시기에는 뉴스 등 정성적 평가를 반영하는 비정형 데이터의 가치가 크게 증가한다. 뉴스와 같은 텍스트 자료의 경우 실시간으로 다양한 소스를 통해 정보가 생성되며, 전문가 판단 등을 반영하므로 정보의 양과 질에서 경제 이벤트를 설명하는 풍부한 정보를 제공한다고 할 수 있다.

국내에서 경기 판단을 위한 경기선행지수, BSI 및 GDP 등의 예측을 위해 비정형 텍스트 자료를 이용한 분석은 아직 많이 시도되지 않고 있다. 그러나 비정형 텍스트 자료가 가지는 정보의 다양성과 방대함 그리고 속보성을 고려할 때 비정형 데이터를 통계 모형에 반영하여 분석할 필요성은 충분하다.

이러한 가능성을 배경으로 본 연구에서는 경기선행지수, BSI, GDP 등의 목표 변수에 대해 최근 한국은행에서 개발되어 사용되고 있는 뉴스 기반 경제 부문별 텍스트 자료, 애널리스트 리포트 기반 텍스트 자료 그리고 애널리스트 리포트 기반 네트워크 자료 등을 추가한 경기 예측 모형을 개발 및 평가하는 것을 그 목표로 두고자 한다. 구체적으로 비정형 텍스트 자료의 변동성 안정을 위한 자료변환, 차원축소 등 텍스트 지표의 후처리 과정이 실제 분석에서 예측정확도를 높일 수 있는지 실증분석을 통해 검증하고 이를 통해 비정형 텍스트 지표의 활용 가능성을 제고하고자 한다.

Ⅲ. 자료소개

1. 텍스트 지표

가. 뉴스 기반 경제 부문별 텍스트 자료

2005년 이후 뉴스 기반 경제 부문별 텍스트 지표(이하 TFI 자료)로 <표 1> 과 같이 주요 거시 변수 및 일부 산업 관련 미시 변수를 선정하여 15개 부문으로 구성되어 있다. 본 연구에서 사용되는 TFI 자료는 서범석(2022)에 기반한 자료이다. TFI 자료의 경우 월별 자료이고, 분기별 분석을 위해서는 1-3월을 1분기, 4-6월을 2분기, 7-9월을 3분기, 10-12월을 4분기로 합성하여 월별 자료의 평균치를 각 분기 값으로 사용하였다.

나. 애널리스트 리포트 기반 텍스트 업황 자료

2019년 이후 애널리스트 리포트 기반 산업별 텍스트 지표의 업황 자료(이하 MA자료)로 <표 2>와 같이 41개 부문으로 구성되어 있다. MA 자료는 앞서 언급된 선행 연구인 서범석(2023)에 기반한 자료이다. MA 자료의 경우는 월별, 분기별 자료로 구분되어 있다.

다. 네트워크 지표

2019년 이후 애널리스트 리포트 기반 자료에서 추정한 산업별 유사도 네트워크 자료로 <표 3>과 같이 39개 부문으로 구성되어 있고, 이는 MA 자료의 연번 1-39번과 동일하다. MA 자료 연번의 40-41번은 결측치가 과다하게 나타나 동 자료 작성에서는 제외되었다.

2. 예측 목표 변수

월별 자료에 대한 예측 목표 변수는 선행지수 순환변동치, BSI 전산업 매출실적, BSI 전산업 매출 전망이다. 분기별 자료의 예측 목표 변수는 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기비이다.

Ⅳ. 자료의 차원축소

본 연구에서 사용되는 변수는 TFI 자료와 MA 자료인데 두 자료의 변수의 수는 각각 15개와 41개이다. 2005년부터 월별로 만들어진 TFI 자료의 경우 15개의 변수의 수는 그렇게 높은 차원이라고 할 수는 없지만, 2019년부터 분기별로 작성된 MA 자료의 경우 자료의 수에 비해 41개의 변수는 비교적 많은 편이다. 이러한 고차원 분석에서 변수의 차원축소는 모형화를 보다 안정적으로 하기 위해 중요한 역할을 할 수 있음이 이미 다양한 형태의 고차원 자료 분석에서 제시되고 있다. 이를 위해 본 연구에서는 충분차원축소의 접근법을 사용하고자 한다.

1. 충분차원축소 개요

충분차원축소란 회귀분석에서 고차원 데이터의 차원을 줄이기 위해 사용되는 통계적 방법이다. 충분차원축소는 주로 회귀분석에서 사용되는 고차원의 설명변수의 차원을 축소하는 것이 그 주요 목적이지만, 반응변수의 차원이 고차원인 경우 반응변수의 차원 축소도 그 목적에 포함된다. 하지만 본 연구에서는 일차원의 반응변수를 고려하므로 반응변수의 차원축소에 대해서는 언급하지 않을 것이다. 이하 내용에서의 차원축소는 회귀분석에서 설명변수의 차원축소에 대한 것임을 먼저 밝혀 두고자 한다.

여러 가지 특징이 있는 고차원의 데이터를 사용하는 경우 관심 있는 반응변수를 예측하는 데 관련성이 높은 특징을 식별해 내는 것이 어려움에 직면할 수 있다. 충분차원축소의 목표는 회귀분석을 보다 정확히 하기 위해 관련된 설명변수 정보를 유지하면서, 원 설명변수의 저차원의 선형 결합 형태로 차원을 축소하는 방법론이다. 이를 보다 구체적으로 설명하면 원래 p차원의 설명변수 $X \in R^p$ 가 주어졌을 때 이를 회귀의 목적인 조건부 분포 $y \mid X$에 대해 정보의 손실 없이 원래의 설명변수를 저차원의 선형변환인 $C^T X$로 대체하는 것을 의미한다. 여기서 C는 $p \times q$차원의 행렬이고, 이는 다음과 같이 표현할 수 있다.

$$y \perp X \mid C^T X$$

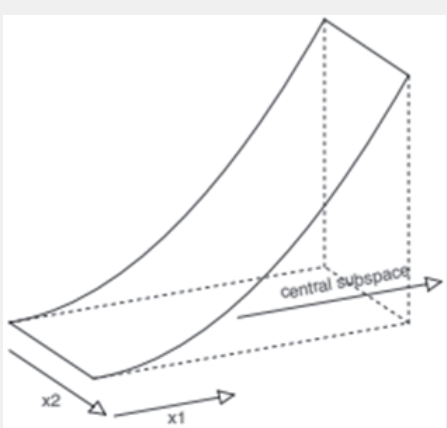
여기서 "$\perp$" 는 통계적 독립을 뜻한다. 만약 $p < q$가 성립한다면, 원래 p차원의 설명변수 X는 회귀에 대한 정보의 손실 없이 q차원의 $C^T X$로 대체될 수 있다. 위의 식을 만족

하는 행렬 C의 열에 의해 생성되는 공간을 차원축소 부분공간(dimension reduction subspace)이라고 한다.

하나의 회귀분석 문제에서 위의 식을 만족하는 C는 여러 개 존재할 수 있다. 그렇다면 차원축소의 목적상 최소부분공간을 찾는 것이 바람직하다. 여러 개의 차원축소 부분공간이 있다는 전제 하에 최소의 차원축소 부분공간은 이들의 교집합을 통해 만들어질 수 있다. 즉, 가능한 모든 차원축소 부분공간들의 교집합이 차원축소 부분공간이 된다면, 당연히 최소 공간일 뿐만 아니라 유일성도 확보될 수 있다. 이 교집합의 차원축소 부분공간을 중심부분공간(central subspace)이라고 하며, $S_{y|X}$로 정의한다. 충분차원축소의 주요 목적은 당연히 이 중심부분공간을 추정하는데 있다. 중심부분공간의 차원을 d라고 표시할 것이며, 정치 기저 행렬을 $\eta \in R^{p\times d}$으로 나타낼 것이다. 또한 저차원의 선형 결합 형태인 설명변수 $\eta^T X$를 충분축소변수(sufficient predictor)라고 할 것이다.

방법론적으로 충분차원축소의 핵심 아이디어는 고차원 특징 공간의 선형 또는 비선형 투영(projection)을 저차원 부분공간에 찾는 것으로 투영은 관련 없는 정보를 제거하면서 가능한 많은 데이터 구조를 보존해야 한다. 이는 <그림 1>(Li, 2018)과 같이 반응변수를 설명하기에 충분한 공간의 방향 집합을 찾아냄으로써 달성할 수 있다. 방향 집합을 식별함으로써 전체 데이터가 아닌 소수의 특징 데이터만을 이용하여, 통계 모델의 정확성과 해석 가능성을 향상시킬 수 있으며, 이는 시각화 측면에도 도움이 될 수 있다. 입력 데이터가 고차원적이고 관측 횟수가 제한적인 상황에서 특히 유용한데, 이는 과적합 위험을 줄이고 통계 추론의 효율성을 향상시키는 데 도움이 될 수 있기 때문이다. 이러한 특징들로 인해 충분차원축소 방법은 많은 분야에서 적용되고 있으며, 특히 고차원 데이터가 주로 쓰이는 유전학, 신경과학, 금융 등의 분야에서 사용되고 있다.

〈그림 1〉 **중심부분공간; X1, X2는 2차원의 설명변수임**



충분차원축소와 중심부분공간의 정의에서 회귀 모형에 대한 특정한 가정을 하지 않고 있다. 즉 충분차원축소는 방법론적으로 모형에 기반한 모수적 방법론보다는 중심부분공간을 추정하기 위한 최소의 가정들만으로 사용될 수 있는 비모수 혹은 반모수적 방법론들이 보편적이다.

보통 중심부분공간을 추정할 때 $y \mid X$에서 직접 추정하기보다는 $y \mid Z$에서 회귀를 고려하여 추정한다. 행렬 Σ 를 X의 공분산 행렬이라고 정의하자. 그러면 Z은 다음과 같이 정의된다: $Z = \Sigma^{-\frac{1}{2}}(X - E(X))$. 즉 Z는 원래의 설명변수를 평균이 0이고 공분산 행렬이 자기 행렬(identity matrix)이 되도록 변환한 설명변수이다. 그렇다면 $y \mid X$의 중심 부분공간과 $y \mid Z$의 중심 부분공간 사이에는 다음이 성립한다:

$$S_{y \mid X} = \Sigma^{-\frac{1}{2}} S_{y \mid Z}$$

표준화한 설명변수 Z를 고려하는 주요 이유는 추정에 있어 발생하는 계산 오차를 줄여 보다 정확하게 중심부분공간을 추정하는 데 있다. 이후 $S_{y \mid Z}$의 중심부분공간에 대한 정치 기저 행렬을 η_Z 라고 정의한다. 이후 소개되는 충분차원축소 방법들은 모두 방법론적 소개를 $y \mid Z$회귀를 고려하고자 한다.

지난 20년 넘게 중심 부분공간을 추정하는 다양한 방법론들이 개발됐지만, 본 연구에서는 다음의 대표적인 3가지 방법을 고려할 것이다.

(1) Sliced Inverse Regression (Li, 1991; SIR)

(2) Directional Regression (Li and Wang, 2007; DR)

(3) Seeded Method (Cook et al., 2007; seed)

위 세 가지 방법들을 다음 세 절에 걸쳐 소개하고자 한다.

2. 충분차원축소 방법

가. Sliced Inverse Regression

Sliced inverse regression(이하 SIR) 방법은 충분차원축소 방법 중 가장 오래된 역사를 지닌 방법 중 하나이다. 1991년 Li에 의하여 처음으로 소개되었으며, 그 이후로 다양한 응용 분야에서 사용되고 있다. 널리 알려진 차원축소 방법 중 하나인 주성분 분석(Principal Component

Analysis, PCA) 방법과의 차이점은 PCA 방법은 차원축소를 진행할 때 반응변수가 축소과정에서 사용되지 않지만, SIR 방법은 반응변수가 사용된다. 이는 SIR가 반응변수와 관련 있는 설명변수를 식별하는 데 더 유용할 수 있고, PCA보다 회귀분석에서 차원축소를 보다 효과적으로 수행할 수 있음을 암시하고 있다.

SIR가 중심부분공간을 추정하기 위해서는 다음의 선형성 조건이 만족되어야 한다: $E(X \mid \eta_Z^T Z)$는 $\eta_Z^T Z$에 대하여 선형이다.

위의 조건은 설명변수 X에 대해서만 적용되는 조건이다. 위의 선형성 조건은 중심부분공간을 생성하는 기저 η_Z에 대해서만 성립하면 된다. 이외의 다른 충분 차원축소 공간의 기저에 대해 성립할 필요는 없다. 따라서 위의 조건은 매우 강한 조건으로 보일 수 있다.

실제로 회귀분석에서의 분포적 혹은 이외 필요한 다양한 조건들은 조건부 분포인 $y \mid X$에 가정된다. 또한 일반적으로 회귀분석에서는 설명변수에 대해서는 별다른 조건을 부여하지 않는다. 하지만 보다 나은 적합을 위해 대부분의 경우 설명변수가 다변량 정규분포를 따르도록 Box-Cox 변환 방법을 사용한다. 왜냐하면 그러한 경우 다변량 정규분포 이론에 의해 $y \mid X$는 정규분포를 따르게 되고, 선형 회귀의 적합이 자료를 잘 설명할 수 있기 때문이다.

위의 선형성 조건은 설명변수 Z의 분포가 타원 대칭분포(elliptically contoured distribution)를 따른다면 모든 선형변환 $C^T Z$에 대해 성립한다. 이러한 타원 대칭분포 중 하나가 다변량 정규분포이고, 위에서 논의된 바와 같이 설명변수의 다변량 정규분포 가정은 실제 회귀분석에서 보편적으로 가정되고 있다. 또한 Hall and Li(1993)에 따르면 비록 Z이 타원 대칭분포를 따르지 않을지라도 설명변수의 차원이 높아지면 이 선형성 조건은 만족된다는 것을 증명하였다. 만약 선형성 조건을 만족시키고자 한다면 설명변수가 다변량 정규분포를 따르도록 변수 변환을 하는 것이 일반적이다. 이는 선형성 조건이 조건 자체가 분포를 가정하는 것처럼 강하지 않으며, 조건이 만족하지 않더라도 현실적인 대안이 있다는 점을 강조하고자 한다.

Li(1991)에 따르면 위의 선형성 조건이 만족되면, 다음의 식이 성립한다:

$$\Sigma^{-1/2} E(Z \mid y) \subseteq S_{y \mid Z}$$

중심부분공간 $S_{y \mid Z}$의 추정은 본질적으로 역회귀인 $E(Z \mid y)$을 추정함으로써 가능하다. 방법론에서 분포적 가정을 하지 않았기 때문에 $E(Z \mid y)$의 추정은 비모수적일 수 밖에 없다. $E(Z \mid y)$은 만약 반응변수가 이산형이라면 직관적이다. 반응변수가 이산형일 경우

$E(Z \mid y)$는 단순히 반응변수의 범주에 해당되는 Z의 표본평균으로 추정될 수 있다. 반응변수가 연속형인 경우 해당 반응변수를 이산형으로 범주화한다면 쉽게 $E(Z \mid y)$를 추정할 수 있을 것이다. Li(1991)에서는 이러한 연속형 반응변수의 그룹화를 슬라이싱(slicing)이라고 표현하였다. 슬라이싱을 할 때 자료가 겹치지 않도록 고르게 나누는 것은 중심부분공간의 정확한 추정을 위해 중요하다.

SIR는 일반적으로 선형적 추세를 보이는 회귀분석 문제에서 잘 작동하는 것으로 알려져 있다. 해당 방법론 적용 단계를 알고리즘으로 정리하면 다음과 같다.

알고리즘 Sliced Inverse Regression

1. 설명변수의 표본 평균과 표본 공분산을 계산한다.
$$\hat{\mu} = E_n(X), \quad \hat{\Sigma} = cov_n(X).$$
2. 설명변수를 표준화 한다.
$$Z_i = \hat{\Sigma}^{-\frac{1}{2}}\left(X_i - \hat{\mu}\right), \quad i = 1, \ldots, n.$$
3. 반응변수가 연속형인 경우 y를 h개의 분할 구간 $j_1, \ldots, j_l$ 로 나누고, 각 분할 구간에서의 표준화된 설명변수의 평균을 계산한다.
$$E_n(Z vert\, Y \in J_l) = \frac{E_n[ZI \mid y \in J_l]}{E_n[I(y \in J_l)]}, \quad l = 1, \cdots, h.$$
4. 공분산 행렬을 계산한다.
$$\hat{\Lambda} = \sum_{i=l}^{h} E_n[I(y \in J_l)]\, E_n(Z \mid y \in J_l) E_n\left(Z^T \mid y \in J_l\right).$$
5. $\hat{\Lambda}$의 고유벡터 중 가장 큰 고유치 d개에 대한 고유벡터를 $\hat{v_1}, \ldots, \hat{v_r}$라고 정의하면, $\hat{\eta_z} = (\hat{v_1}, \ldots \hat{v_d})$이고, $\hat{\eta} = \hat{\Sigma}^{-\frac{1}{2}}(\hat{v_1}, \ldots, \hat{v_d})$이 된다.

요약하자면, SIR은 설명변수와 반응변수 사이의 관계에 대한 가장 많은 정보를 포착하는 역회귀 함수의 공간에서 가장 중요한 방향을 식별하는 데 좋은 성능을 나타내는 방법이다. 설명변수를 이러한 방향으로 투영함으로써, 반응변수에 대한 대부분의 정보를 보존하는 설명변수의 저차원 표현식을 얻을 수 있다. SIR은 계산의 효율성으로 인해 변수가 아주 많은 고차원 데이터를 처리할 수 있다는 장점이 있으며, 데이터 분포에 대한 가정에 의존하지 않으므로 다양한 데이터에 적용이 가능한 방법이다. 하지만 SIR의 한계는 슬라이싱 개수에 민감하다는 점이며, 중심부분공간이 정확하게 식별되도록 충분한 수의 방향을 선택하는 것이 중요하다.

나. Directional Regression

SIR는 보편적으로 많이 사용되는 충분차원축소 방법이지만 위에서 언급된 방법론적 한계가 명백히 존재한다. 특히 자료가 반응변수와 설명변수의 관계가 다음과 같은 $y = x^2$과 같은 대칭 관계를 가질 때 SIR는 이러한 관계를 추정할 수 없다. 이 대칭 관계는 second-order polynomial 회귀로 실제 자료 분석에 자주 사용되는 선형 회귀 모형 중 하나이다.

SIR가 가지는 또 다른 방법론적 한계는 중심부분공간을 완전히 추정하지 못하는 것이다. 이를 위해 Li et al.(2005)은 contour regression을 제안하였다. contour regression은 완전하게 중심부분공간을 추정할 수 있는 장점이 있지만, 모든 contour의 방향을 계산해야 하는 복잡성으로 인해 상당한 시간이 소요된다는 결정적 단점이 있다. contour regression의 이러한 계산적 복잡성을 보완하기 위하여 개발된 directional regression(Li and Wang, 2007; DR)을 본 연구에서는 고려하고자 한다. DR은 역회귀의 두 조건부 적률을 사용하여 중심부분공간을 추정하는 방법으로써 반응 변수와 설명변수 간의 공분산을 최대화하는 방향 또는 축의 집합을 찾는 것을 기반으로 한다. DR은 역회귀의 2차 적률도 고려함으로써 contour regression이 가지는 계산적 복잡함을 해결하였으며, SIR가 가지는 비선형 회귀에서의 문제점도 완화할 수 있다. 또한 DR은 SIR가 가지는 slicing의 민감함에 강건하다는 장점이 있다. DR은 SIR과 같이 예측 모델의 정확도를 향상시키기 위한 전처리 단계로 사용될 수 있으며, DR은 다른 차원축소 기법보다 이상치 및 노이즈에 덜 민감하고, 계산 효율성이 높은 방법으로 알려져 있다.

DR을 사용하기 위해서는 SIR에서 필요한 선형성 조건 외에 다음의 등분산 조건을 만족해야 한다: $cov(Z \mid \eta_Z^T Z)$는 $\eta_Z^T Z$에 대해 일정하다.

등분산 조건은 선형성 조건과 마찬가지로 설명변수에만 해당되는 조건이다. 하지만 선형성 조건과는 다르게 타원 대칭분포에서 항상 성립하지 않지만, 다변량 정규분포에서는 만족한다.

Li and Wang(2007)에 따르면 선형성 조건과 등분산 조건이 만족되면, 다음의 식이 성립하고, 아래의 식을 규명하는 것이 방법론적으로 DR이 가지는 가장 핵심적인 부분이다.

$$(2I_p - E[(Z-\tilde{Z})(y,\tilde{y})])^2 \subseteq S_{y \mid Z}$$

자료를 이용한 $(2I_p - E[(Z-\tilde{Z})(y,\tilde{y})])^2$의 추정은 SIR와 같이 먼저 연속형 반응변수를 슬라이싱을 통하여 범주화한 후 각 범주별 설명변수의 평균과 공분산 행렬을 계산함으로써 가능해진다. DR을 이용한 중심부분공간의 추정은 다음과 같이 정리된다.

알고리즘 Directional regression

1. 설명변수의 표본 평균과 표본 공분산을 계산한다.
$$\hat{\mu} = E_n(X),\ \ \hat{\Sigma} = cov_n(X).$$
2. 설명변수를 표준화한다.
$$Z_i = \hat{\Sigma}^{-\frac{1}{2}}\left(X_i - \hat{\mu}\right),\quad i = 1, \ldots, n.$$
3. 반응변수가 연속형인 경우 y를 h개의 분할 구간 $j_1, \ldots, j_l$로 나누고, 각 분할 구간에서 다음을 계산한다.
$$M_{1l} = E_n(Z \mid y \in J_l);\ M_{2l} = E_n(ZZ^T \mid y \in J_l)$$
4. 이를 이용하여 다음의 3개의 행렬을 계산한다.
$$\Lambda_1 = \sum_{l=1}^{h} \hat{p_l} M_{1l}^2 \ \ ;$$
$$\Lambda_2 = (\sum_{l=1}^{h} \hat{p_l} M_{1l} M_{1l}^T)^2;$$
$$\Lambda_3 = (\sum_{l=1}^{h} \hat{p_l} M_{1l}^T M_{1l})(\sum_{l=1}^{h} \hat{p_l} M_{1l} M_{1l}^T).$$
여기서 $\hat{p_h}$은 h번째 수준에서의 관측 수의 비율을 의미한다.
5. 4번째에서 계산된 행렬을 이용하여 다음의 행렬을 계산한다.
$$\Lambda_{DR} = \ 2\Lambda_1 + 2\Lambda_2 + 2\Lambda_3 - 2I_p.$$
6. Λ_{DR}의 고유벡터 중 가장 큰 고유치 d개에 대한 고유벡터를 $\hat{v_1}, \ldots, \hat{v_r}$라고 정의하면, $\hat{\eta_z} = (\hat{v_1}, \ldots \hat{v_d})$이고, $\hat{\eta} = \hat{\Sigma}^{-\frac{1}{2}}(\hat{v_1}, \ldots, \hat{v_d})$이 된다.

다. Seeded method

앞에서 설명된 두 가지 방법뿐만 아니라 대부분의 충분차원축소 방법들은 방법론적으로 설명변수에 대한 표본공분산 행렬의 역행렬 계산이 필수적이다. 하지만 자료의 수보다 변수의 개수가 많은($n \leq p$) 경우에는 표본 공분산 행렬의 역행렬이 존재하지 않는다. 이것은 충분차원축소의 목적을 고려한다면 아이러니한 것이다. 이를 극복하기 위하여 Cook et al. (2007)은 역행렬의 계산을 요구하지 않는 방법론을 제안하였고, 이 방법론을 seeded method라고 한다.

Seeded method는 방법론 이름 그대로 아래의 관계를 만족하는 seed 행렬이라고 불리는 행렬 $v \in R^{p \times d}$을 먼저 찾아야 한다.

$$\Sigma^{-1} S(v) \subseteq S_{y \mid X} \Leftrightarrow S(v) \subseteq \Sigma S_{y \mid X}$$

여기서 $S(v)$는 v의 열에 의해서 생성되는 공간을 의미한다.

앞에서 언급한 선형성 조건이 만족하는 상황에서 seed 행렬 v로 사용될 수 있는 행렬의 후보는 다음과 같다.

$$E(X \mid y) : \Sigma^{-1}\{E(X \mid y) - E(X)\} \in S_{y \mid X} \Leftrightarrow S\{E(X \mid y) - E(X)\} \subseteq ; \Sigma S_{y \mid X}$$
$$cov(X, y) : \Sigma^{-1} cov(X, y) \in S_{y \mid X} \Leftrightarrow S\{cov(X, y)\} \subseteq \Sigma S_{y \mid X}$$

위의 seed 행렬 후보 중 $\Sigma^{-1}\{E(X \mid y) - E(X)\}$는 SIR가 중심부분공간을 추정하기 위해 사용되는 행렬이며, $\Sigma^{-1} cov(X, y)$는 최소제곱 추정량이다. 본 연구에서는 $cov(X, y)$를 seed 행렬로 사용한다.

그리고 중심부분공간을 포함하는 $M_{y \mid X}$이 존재한다고 가정하자. 그렇다면 다음의 관계가 성립한다.

$$\Sigma^{-1} S(v) \subseteq S_{y|X} \subseteq M_{y|X} \Leftrightarrow S(v) \subseteq \Sigma M_{y|X}.$$

Seeded method는 직접적으로 $S_{y|X}$을 추정하는 것이 아니라 seed 행렬을 이용하여 설명변수의 역행렬을 계산하지 않고 $M_{y|X}$의 기저를 추정하는 데 있다. 그러면 그 기저에 대한 투영 행렬을 일반적인 내적 공간이 아닌 $\langle a, b \rangle_{\Sigma}$로 정의하는 Σ-내적 공간에서 정의되는 투영행렬을 만들고 이를 이용하여 중심부분공간을 추정한다.

행렬 R를 $M_{y|X}$의 $p \times q$기저행렬이라고 하자. $\langle a, b \rangle_{\Sigma}$내적에 대해 $M_{y|X}$에 투영하는 직교 투영 연산자 $\mathrm{P}_{M_{y|X}(\Sigma)} = R(R^T \Sigma R)^{-1} R^T \Sigma$이다. 이 직교 투영 연산자의 계산에 역행렬이 필요하지만 $p \times p$행렬이 아닌 $q \times q$행렬인 $R^T \Sigma R$의 역행렬이다. 만약 q가 표본 수 n보다 작은 경우 역행렬 계산이 가능해진다. 만약 $\Sigma^{-1} v = S_{y|X}$라면 다음의 관계가 성립된다.

$$S_{y|X} = \mathrm{P}_{M_{y|X}(\Sigma)} S_{y|X} \Sigma^{-1} S(v) = S\big(R(R^T \Sigma R)^{-1} R^T \Sigma \Sigma^{-1} v\big) = S\big(R(R^T \Sigma R)^{-1} R^T v\big).$$

따라서 $R(R^T \Sigma R)^{-1} R^T v$는 중심부분공간을 생성하는 기저 행렬이 되고, 이를 통하여 중심부분공간을 추정할 수 있다. $R^T \Sigma R$의 역행렬은 일반적으로 무어-펜로즈 역행렬(Moore-Penrose inverse)로 계산이 가능하다.

위에서 언급된 행렬 R은 자료를 통하여 추정되어야 하며, 실제로 seed 행렬과 설명변수의 표본 공분산 행렬을 이용하여 추정한다. 행렬 R을 추정할 때 R의 열공간이 $S_{y|X}$를 포

함할 만큼 충분히 커야 하며 동시에 자료를 통해 계산이 용이할 수 있도록 합리적 작아야 한다. 이를 위해 Cook et al. (2007)은 행렬 R의 추정에 대해 v을 Σ에 반복적 투영하는 방법을 제시하고 있다.

$$R_u \equiv (v,\ \Sigma v,\ \cdots,\ \Sigma^{u-1}v),\quad u = 1,2,\cdots,u^*$$

여기서 u^*는 투영종료지수(termination index of projections)라고 한다. 모든 $u \geq 2$에 대해 $S(R_{u-1}) \subseteq S(R_u)$의 관계가 성립한다. 그러므로 적절한 $S(R_u) = S_{y|X}$를 보장하는 가장 작은 u를 찾는 것이 중요하다. 이에 대해서는 R_u의 정보가 더 이상 증가하지 않는 점을 찾아 정한다. 관련한 자세한 내용은 Um et al. (2018)을 참고하길 바란다. 해당 방법론 적용 단계를 알고리즘으로 정리하면 다음과 같다.

알고리즘 Seeded method

1. v로 사용될 수 있는 후보 행렬 중 하나의 커널 행렬을 선택한다. 본 연구에서는 $cov(X,y)$를 seed 행렬로 사용한다.
2. 설명변수의 표본 공분산 행렬 $\hat{\Sigma}$와 seed 행렬 $\hat{v}$를 계산한다.
3. 적당히 큰 값의 u에 대해 다음의 $\widehat{R_u}$를 계산한다.
$$\hat{R}_u = (\hat{v},\ \hat{\Sigma}\hat{v},\ \ldots,\ \hat{\Sigma}^{u-1}\hat{v}).$$
4. 계산된 $\widehat{R_u}$를 통해 u^*을 정하고, $\hat{R}_{u^*} = (\hat{v},\ \hat{\Sigma}\hat{v},\ \ldots,\ \hat{\Sigma}^{u^*-1}\hat{v})$.
5. $\hat{R}_{u^*}(\hat{R}_{u^*}^T\hat{\Sigma}\hat{R}_{u^*})^{-1}\hat{R}_{u^*}^T\hat{v}$를 계산하여 중심부분공간의 기저를 추정한다.

Ⅴ. 예측 모형 개발 및 평가

1. 예측 모형 개발 절차

뉴스 텍스트 지표는 15개 부문, 애널리스트 텍스트 지표는 41개 부문에 대한 지표로써 본 보고서에서 다루는 고차원 자료라고 할 수 있다. 이에 4장에서 소개된 충분차원축소 방법들을 적용하여 자료의 차원을 축소하였다. 충분차원축소 방법들은 결측치가 없는 완전 자료에서 적용이 가능하기 때문에 우선 결측치가 많은 MA의 자료는 차원축소 전에 선형 보간법을 이용하여 완전 자료로 만들었다.

또한 본 연구의 중요한 목적 중의 하나가 텍스트 지표의 변동성 안정화 및 노이즈 제거가 예측모형 개발에 효과가 있는지를 검토하는 것이다. 이에 다양한 조절 모수(lambda, λ) 값을 지정하여 HP filter를 TFI와 MA 자료에 적용하였다. 이때 조절 모수 값을 해석이 쉽도록 자료의 주기(frequency, ρ)지표로 치환하여 사용하였다. 주기와 조절 모수 사이에는 다음의 식이 성립한다.

$$\lambda = \ 6.25 \ \rho^4$$

월별 자료에서 사용된 주기 값은 0.5, 1, 1.5와 2이고, 분기별 자료에서는 0.25, 0.5, 0.75와 1이 사용되었다. 이렇게 결측치에 대해 선형보간 후 HP 필터링을 한 TFI와 MA에 대해 각각 앞 장에서 소개된 충분차원축소 방법을 적용하였다.

월별 예측 목표 변수인 선행지수 순환변동치, BSI 전산업 매출실적, BSI 전산업 매출전망에 대해서 TFI만 고려할 경우, 2005년부터의 자료가 있어 적지 않은 표본 수이기 때문에 슬라이스의 수를 10으로 하여 SIR와 DR를 적용하여, 최대 2차원으로 축소한다. 이후 모형화에서는 1차원으로 축소된 자료, 그리고 2차원으로 축소된 자료가 따로 사용된다. 그리고 MA만 그리고 TFI와 MA을 모두 고려한 경우에는 2019년부터의 월별 자료이기 때문에 표본의 수가 변수의 수보다 상대적으로 많이 크지 않기 때문에 seeded method를 적용하였고, 이때 사용된 seed행렬은 $cov(X,y)$이다. 따라서 이때 TFI와 MA는 각각 1차원으로 축소가 된다.

분기별 예측 목표 변수인 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기비에 대해서는 TFI만 고려한 경우 역시 월별 자료분석과 같이 슬라이스 수를 10으로 하여 SIR와 DR을 이용하여 차원축소를 하였다. 다만 MA도 고려하는 경우 자료의 수가 16개로

매우 적기 때문에 분기별에서는 MA자료가 사용되지는 않는다.

차원축소시 월별, 분기별 예측 목표 변수들과 TFI와 MA 간에 시차 효과(lag effect)가 존재할 수 있기 때문에, 월별 자료의 경우는 시차로 최대 3개월의 시차를 고려한 반면, 분기별 자료의 경우는 최대 1분기 시차만 고려하였다. 이에 대한 이유는 텍스트 자료의 경우 공식 통계보다 선행은 하지만, 뉴스와 리포트 기반이기에 긴 시차효과는 없으리라 판단했기 때문이다. 최적 시차는 시차별 차원축소된 변수와 다섯 개의 예측 목표 변수 간의 선형회귀 모형을 적합하여 가장 높은 adjust R^2 값을 보이는 시차를 검토하여 분석에 활용하였다. 월별 분석에서는 선행지수 순환변동치, BSI 전산업 매출실적과 BSI 전산업 매출전망에 대해 SIR을 이용하여 TFI만 차원축소를 한 경우 각각 2, 3와 0이 최적 lag이었고, DR을 적용하면 각각 2, 1와 3이었다. 반면 seeded method로 MA만 축소를 하는 경우에는 3, 1과 2이다. 마지막으로 seeded method로 TFI와 MA를 축소한 경우에는 선행지수 순환변동치, BSI 전산업 매출실적과 BSI 전산업 매출전망에 대해 각각 3, 1와 1이다. 분기별 분석에서는 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기에 대해 SIR을 이용하여 TFI의 차원축소 시 최적 lag는 각각 0과 1인 반면, DR을 이용하면 0과 0이었다.

예측 모형으로서는 시계열분석에서 가장 고전적으로 널리 사용되는 자기회귀(Autoregressive Regression, AR)모형을 적합하고 BIC(Bayesian Information Criterion)를 기준으로 가장 적정한 p를 분석하였다. 이를 바탕으로 AR모형을 기준으로 AR모형에 차원축소된 텍스트 자료 그리고 AR모형에 차원축소 전의 원텍스트 자료를 고려한 세가지 예측 모형을 적합하였다. AR모형에 텍스트 자료를 함께 고려한 예측 모형으로 차원축소된 텍스트 자료를 이용한 경우에는 선형회귀 모형을 적합하였고, 원래의 텍스트 자료를 이용한 경우에는 Lasso 회귀를 적합하였다. 모형의 예측력을 비교하기 위하여, rolling window를 활용한 one-step ahead cross-validation을 통하여 테스트 자료의 에러를 확인하였다. 이에 대해 다음과 같이 정리할 수 있다. 월별 자료의 경우는 2020년 12월까지의 자료를 학습 데이터셋으로, 이후 2021년 1월부터 2022년 12월까지 2년치 자료를 검증 데이터셋으로 사용하였다. 반면에 자료의 수로 인해 분기별 자료의 경우는 2021년 4분기까지의 자료를 학습 데이터셋으로, 이후 2022년의 4분기 자료를 검증 데이터셋으로 사용하였다.

네트워크 자료의 분석에 대해서는 이후의 절에서 상세히 설명하기로 한다.

2. 월별 텍스트 지표 분석

월별 자료의 예측 목표 변수는 선행지수 순환변동치, BSI 전산업 매출실적, BSI 전산업 매출전망이다. 앞에서 언급된 분석 방법을 적용하여 MAE를 계산하였다.

먼저 TFI만을 이용한 분석을 살펴보고, 각 예측 목표 변수별 최적의 결과가 <표 4>에 제시되어 있고, 상세한 결과는 <표 7>에 정리되어 있다.

〈표 4〉 **TFI만을 이용한 월별 지표 예측 결과**

모형 / 변수	변동성 안정화	자료 종류	차원축소 유무 및 방법	적합모형	MAE
선행지수 순환변동치	적용: 0.5	공식 통계 + TFI	Directional Regression	선형회귀/Lasso	0.112
BSI 전산업 매출실적	적용: 2	공식 통계 + TFI	해당사항 없음	선형회귀	2.285
BSI 전산업 매출전망	적용: 0.5	공식 통계 + TFI	Directional Regression	Lasso	2.241

<표 4>를 살펴보면, 우선 AR만 사용하기보다는 TFI의 자료를 사용하는 것이 예측력을 높여주고 있다. 또한 HP 필터링을 이용한 변동성 안정화가 예측을 보다 정확하게 해준다. 그리고 선행지수 순환변동치와 BSI 전산업 매출 전망에 대해서는 차원축소를 하는 것이 원자료를 사용하는 것보다 예측력을 높여주는 반면, BSI 전산업 매출실적에서는 차원축소를 하지 않은 원자료를 사용하는 것이 가장 높은 예측력을 보여주고 있다.

MA가 사용되는 월별 자료 분석에 대한 최적 결과는 <표 5>에 그리고 보다 상세한 결과는 <표 8>에 정리되어 있다.

<표 5>를 살펴보면, HP 필터링을 이용한 변동성 안정화는 선행지수 순환변동치에서만 최적의 예측력을 제시하고 있다. 그리고 텍스트 자료의 예측력 강화 측면에서는 세가지 예측 목표 변수에 대해 결과가 제시되고 있다. 선행지수 순환 변동치와 BSI 전산업 매출실적에서는 TFI와 MA을 모두 사용하는 것이, 마지막으로 BSI 전산업 매출 전망에서는 모두 사용하지 않는 것이 최적의 예측력을 나타낸다.

〈표 5〉 **MA 혹은 TFI와 MA을 모두 이용한 월별 지표 예측 결과**

모형 / 변수	변동성 안정화	자료 종류	차원축소 유무 및 방법	적합모형	MAE
선행지수 순환변동치	적용: 0.5	공식 통계+TFI+MA	해당사항 없음	Lasso	0.146
BSI 전산업 매출실적	미적용	공식 통계+TFI+MA	해당사항 없음	Lasso	2.111
BSI 전산업 매출전망	미적용	공식 통계	해당사항 없음	AR	1.772

월별 자료 분석에 있어 이렇게 차이가 나타나는 가장 큰 이유로 자료의 크기를 들 수 있다. TFI만 사용하는 경우에는 2005년부터의 월별 자료를 사용할 수 있는 반면, MA를 고려하면 2019년도의 자료만 사용하게 된다. 이러한 자료의 크기는 결과의 차이에 직접적인 영향을 줄 수 있다. 특히 변동성 안정화의 경우는 긴 시간에 걸친 자료가 있을 때 그 효과가 나타남을 확인할 수 있고, 자료의 차원축소 또한 많은 자료에서 보다 효과적인 차원축소가 가능함을 알 수 있다.

3. 분기별 텍스트 지표 분석

분기별 자료의 예측 목표 변수는 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기비이며, 두 경제지표를 예측하기 위해 월별 TFI 텍스트 지표만을 고려한다. MA 자료가 고려되지 않은 이유는 앞에서 언급하였듯이 2019년부터의 분기별 자료의 수가 매우 적어 분석의 결과에 대해 신뢰할 수 없기 때문이다. 예측에 대한 최적 결과는 <표 6>에 정리되어 있고, 상세 결과는 <표 9>를 참고하길 바란다.

<표 6>에서 확인할 수 있듯이, 두 예측 목표 변수에 대해 변동성 안정화는 더 나은 예측력을 확보하게 해준다. 또한 TFI 자료의 활용은 예측력을 높여줄 수 있다. 그리고 GDP 실질 계절조정계열 전기비에 대해서는 SIR을 이용한 차원축소가 예측력 향상에 효과적인 반면 GDP 실질 원계열 전년동기비의 경우에는 원자료의 TFI을 사용하는 것이 최적의 결과를 제공함을 확인할 수 있다.

〈표 6〉 **TFI만을 이용한 분기별 지표 예측 결과**

모형 / 변수	변동성 안정화	자료 종류	차원축소 유무 및 방법	적합모형	MAE
GDP 실질 계절조정계열 전기비	적용: 0.25 /0.5	공식 통계+TFI	SIR	Lasso	0.404
GDP 실질 원계열 전년동기비	적용: 0.25	공식 통계+TFI	해당사항 없음	Lasso	0.922

4. 네트워크 자료 분석

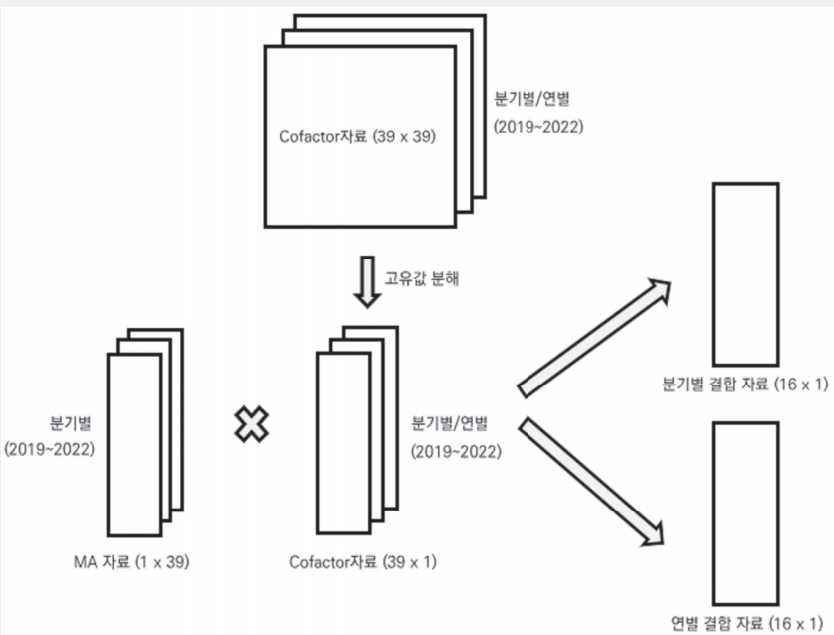
애널리스트 리포트 기반으로 추정한 산업별 유사도 네트워크 자료(cofactor)는 MA자료 중 변수명 1번에서 39번까지의 변수들에 대한 관계를 분기별 그리고 연별 계산한 39×39의

네트워크 행렬자료이다. 이 자료는 예측 목표 변수의 예측에 직접적으로 사용할 수 없다. 우선 편의상 MA의 1-39 변수의 자료를 T1이라고 한다.

우선 각각의 분기별 네트워크 행렬을 spectral decomposition을 이용하여 분해한 후 가장 큰 고유치와 상응하는 고유벡터를 구한다. 이 고유벡터를 l_i라고 정의한다. 이후 각 분기에 해당되는 T1의 관측값 $T1_i$ 와 위에서 구한 고유벡터를 l_i 를 곱하여, 즉 $l_i^T T1_i$, 새로운 관측값을 생성한다. 이렇게 계산한 관측값을 QN_i라고 하자.

이와 동일하게 연별 네트워크 행렬 역시 spectral decomposition을 이용하여 고유벡터를 구한 후 해당 연도에 대응되는 T1의 분기별 관측값을 곱하여 또 다른 관측값을 생성한다. 이 관측값을 AN_i라고 하자. 분기별 네트워크 MA 관측값인 QN_i 와 연별 네트워크 MA 관측값은 AN_i는 분석 가능한 형태의 자료이다. 이 두 자료에 대한 계산 과정은 <그림 2>을 참고하길 바란다.

〈그림 2〉 **애널리스트 리포트 기반 네트워크 자료와 MA 텍스트 지표 자료의 결합 과정**



QN_i와 AN_i가 실제로 MA의 자료와 어떠한 관계성을 가지고 있나 파악하기 위해 분기별 MA 자료를 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기비에 대해 seeded method를 적용하여 차원축소한 자료와 분기별 MA 자료에 대한 첫 번째 주성분에

대한 상관 분석을 실시하였다. 상관분석을 위하여 피어슨의 상관계수를 계산하였고, 이는 <표 10>에 정리되어 있다.

<표 10>을 살펴보면 QN과 AN은 예측 목표 변수를 고려하지 않은 MA의 첫 번째 주성분보다는 예측 목표 변수를 고려하여 축소된 변수들과 월등히 높은 상관관계를 갖고 있음을 확인할 수 있다. 또한 QN과 AN은 모두 seeded method로 축소된 MA 자료와 높은 상관관계를 갖고 있기 때문에, QN과 AN은 바로 해당 변수의 분석에 사용될 수 있음을 파악할 수 있어, 네트워크 자료의 실제적 유용성에 대한 가능성을 확인할 수 있다.

〈표 10〉 **네트워크 자료와 MA 차원축소 자료와의 상관 분석**

차원축소	네트워크자료	상관계수
GDP 실질 계절조정계열 전기비에 대해 seeded method를 이용한 차원축소	분기별 자료 (QN)	−0.625
	연별 자료 (AN)	−0.648
GDP 실질 원계열 전년동기비 대해 seeded method를 이용한 차원축소	분기별 자료 (QN)	−0.619
	연별 자료 (AN)	−0.619
MA의 첫번째 주성분	분기별 자료 (QN)	0.156
	연별 자료 (AN)	−0.443

Ⅵ. 결론 및 제언

본 연구는 노이즈 필터링 기법과 차원축소 등의 방법을 이용하여 텍스트 지표의 정상화에 대해 검토하고 실증분석을 통해 텍스트 지표의 활용 가능성을 높이고자 하였다. 노이즈 필터링의 방법으로 Hodrick and Prescott(1997)이 제안한 필터를 이용하여 뉴스 및 애널리스트 보고서 기반의 텍스트 지표에 다양한 조절 모수 값들을 적용함으로써 노이즈를 제거하였다. 이후 필터링 된 지표를 이용하여 월별 선행지수 순환변동치, BSI 전산업 매출실적, BSI 전산업 매출전망 그리고 분기별 GDP 실질 계절조정계열 전기비와 GDP 실질 원계열 전년동기비 예측 목표 변수에 대해 회귀분석에서 비모수적으로 차원을 축소하는 충분차원축소 방법론 중에서 SIR(Li, 1991), DR(Li and Wang, 2007), seeded method (Cook et al., 2007)을 적용하였다.

분석 결과 월별과 분기별 변수 모두 자료의 수가 많은 경우에 있어서는 텍스트 지표의 사용과 이에 대한 노이즈 필터링이 예측의 정확도를 높이는 것으로 나타나고 있다. 그리고 텍스트 지표 자료의 차원축소를 함에 따라 예측을 보다 더 정확하게 할 수 있음을 제시하고 있다. 하지만 자료의 수가 상대적으로 적은 경우 노이즈 필터링과 차원축소의 효과가 기대와는 달리 미비하였다. 이것이 시사하는 바는 자료가 충분할 경우 텍스트 지표의 사용은 노이즈 필터링과 차원축소를 통해 예측의 정도를 높일 수 있다는 것이다.

차원축소의 또 다른 장점으로는 두 텍스트 지표 자료를 저차원으로 변환함으로써 예측 목표 변수들과 차원축소된 지표들에 대한 시차 효과를 파악하는 것이 원지표를 고려하는 것보다 용이하다.

또한 애널리스트 지표의 네트워크 행렬자료는 그 자체로 분석에 직접적인 사용은 어려우나 행렬의 차원축소결과와 애널리스트 텍스트 지표와의 결합을 통해 실제 분석에서 사용되는 차원축소된 애널리스트 텍스트 지표와 높은 상관관계를 가지고 있어 그 자체만으로 의미 있는 자료가 될 수 있음을 분석을 통해 확인하였다.

본 연구에 관한 사후 연구 문제로 다음을 제안하고자 한다. 뉴스 기반 지표와 애널리스트 보고서 기반 지표에 대한 차원축소를 개별적으로 진행하였다. 하지만 애널리스트 보고서 기반 지표가 시간에 걸쳐 충분히 누적된다면 두 지표를 동시에 축소하는 방법이 고려될 수 있고, 그렇다면 두 지표 간의 상관성이 차원축소하는 과정에 고려될 수 있어 더 바람직한 차원축소가 이루어질 수 있으리라 기대할 수 있다.

그리고 두 텍스트 자료의 차원축소는 노이즈 필터링 이후의 자료에 대한 차원축소이기

때문에 SIR이나 DR 등과 같은 선형 공간의 차원축소 방법보다는 비선형 공간에서의 차원축소 방법이 더 적합할 수 있다.

끝으로 많은 결측치를 선형보간법으로 대체를 하였는데, 다른 보간법을 사용한 후 예측모형을 개발하여 정확도를 비교함으로써, 해당 텍스트 자료에 보다 적합한 보간법을 찾아보는 것도 의미 있는 연구가 될 것이다.

참고문헌

· 서범석(2022), "뉴스 텍스트를 이용한 경기 예측: 경제 부문별 텍스트 지표의 작성과 활용", BOK 이슈노트, No. 2022-18.

· _____(2023), "AI 알고리즘을 이용한 산업 모니터링: 증권사 리포트 텍스트 분석", BOK 이슈노트, No.2023-5.

· 성태제, 시기자(2013), 『연구방법론』, 서울: 학지사.

· Chen, C. C., Huang, H. H., Huang, Y. L. and Chen, H. H. (2021, October), "Constructing Noise Free Economic Policy Uncertainty Index", In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2915-2919.

· Cook, R.D., Li, B., and Chiaromonte, F. (2007), "Dimension reduction in regression without matrix inversion", *Biometrika*, 94, 569–584.

· Hall, P. and Li, KC. (1993), "On almost linearity of low dimensional projections from high dimensional data", *Annals of Statistics*, 21, 867–889.

· Hamilton, J. D. (2018). "Why you should never use the Hodrick-Prescott filter". *Review of Economics and Statistics*, 100(5), 831-843.

· Hodrick, R. J. and Prescott, E. C. (1997). "Postwar U.S. business cycles: An empirical investigation", *Journal of Money, Credit and Banking*. 29 (1): 1–16. doi:10.2307/2953682. JSTOR 2953682.

· Lim, Y., Ahn, Y. H., Yoo, J. K., Park, K. S. and Kwon, O. (2017), "Verifying identities of plant-based multivitamins using phytochemical fingerprinting in combination with multiple bioassays", *Plant Foods for Human Nutrition*, 72, 288-293.

· Lee, K., Choi, Y., Um, H. Y., and Yoo J. K. (2019), "On fused dimension reduction in multivariate regression", *Chemometrics and Intelligent Laboratory Systems*, 193, 103828

· Li, B. (2018). "Sufficient Dimension Reduction Methods and Applications with R", Chapman and Hall/CRC.

· Li, B. and Wang, S. (2007), "On directional regression for dimension reduction", *Journal of the American Statistical Association*, 102, 997–1008.

· Li, B., Zha, H. and Chiaromonte, F. (2005), "Contour regression: A general approach to dimension reduction", *Annals of Statistics*, 33(4), 1580-1616.

· Li, K. C.. (1991), "Sliced inverse regression for dimension reduction", *Journal of the American*

Statistical Association, 86, 316–327.

· Ravn, M. O., and Uhlig, H. (2002), "On adjusting the Hodrick-Prescott filter for the frequency of observations". *Review of economics and statistics*, 84(2), 371-376.

· Seo, B., Lee, Y., and Cho, H. (2022), "Machine-Learning-Based News Sentiment Index (NSI) of Korea", Bank of Korea WP, 15.

· Stratonovich, R. L. (1959), "Optimum nonlinear systems which bring about a separation of a signal with constant parameters from noise", *Radiofizika*, 2(6), 892-901.

〈표 1〉 TFI 텍스트 지표 세부 부문

연번	부문	연번	부문	연번	부문
1	건설투자	6	생산	11	자동차
2	구직	7	선박	12	정부지출
3	도소매	8	설비투자	13	주가전망
4	물가전망	9	세계교역	14	주택가격전망
5	반도체	10	실업	15	채용

〈표 2〉 MA 텍스트 지표 세부 부문

연번	부문	연번	부문	연번	부문
1	1차금속	15	부동산업	29	인쇄/기록매체복제
2	가구	16	비금속광물	30	자동차
3	가죽/가방/신발	17	사업시설/사업지원/임대업	31	전기/가스/증기
4	건설업	18	석유정재/코크스	32	전기장비
5	고무/플라스틱	19	섬유	33	전문/과학/기술
6	교육서비스	20	숙박업	34	전자/영상/통신장비 등
7	금속가공	21	식료품	35	정보통신업
8	금융	22	어업	36	조선/기타운수
9	기타개인서비스	23	예술/스포츠/여가	37	펄프/종이
10	기타기계/장비	24	운수/창고업	38	하수/폐기물처리업
11	기타제조업	25	음료	39	화학물질/제품
12	농업	26	의료/정밀기기	40	전산업
13	도매/소매	27	의료물질/의약품	41	자동차 및 조선/기타 운수
14	목재/나무	28	의복/모피		

〈표 3〉 네트워크 지표 세부 부문

연번	부문	연번	부문	연번	부문
1	1차금속	14	목재/나무	27	의료물질/의약품
2	가구	15	부동산업	28	의복/모피
3	가죽/가방/신발	16	비금속광물	29	인쇄/기록매체복제
4	건설업	17	사업시설/사업지원/임대업	30	자동차
5	고무/플라스틱	18	석유정재/코크스	31	전기/가스/증기
6	교육서비스	19	섬유	32	전기장비
7	금속가공	20	숙박업	33	전문/과학/기술
8	금융	21	식료품	34	전자/영상/통신장비 등
9	기타개인서비스	22	어업	35	정보통신업
10	기타기계/장비	23	예술/스포츠/여가	36	조선/기타운수
11	기타제조업	24	운수/창고업	37	펄프/종이
12	농업	25	음료	38	하수/폐기물처리업
13	도매/소매	26	의료/정밀기기	39	화학물질/제품

〈표 7〉

월별 경제 지표 예측 평균오차(MAE) 비교[1)]

(분석대상기간: 2005년 1월 ~ 2022년 12월)

모형 / 변수	변동성 안정화 lambda	AR 모형	원 TFI 지표		일차원으로 축소된 TFI 텍스트 지표[2)]				이차원으로 축소된 TFI 텍스트 지표[3)]			
			Linear Regression	Lasso Regression	SIR-Linear Regression	SIR-Lasso Regression	DR-Linear Regression	DR-Lasso Regression	SIR-Linear Regression	SIR-Lasso Regression	DR-Linear Regression	DR-Lasso Regression
선행지수 순환 변동치	미적용	0.124	0.124	0.124	0.126	0.127	***0.119***	***0.119***	***0.122***	***0.121***	0.124	0.124
	적용(0.5)		***0.122***	***0.121***	***0.123***	***0.123***	*0.112*	*0.112*	0.128	0.127	***0.116***	***0.115***
	적용(1)		0.139	0.134	0.130	0.129	***0.114***	***0.113***	0.124	***0.123***	***0.117***	***0.115***
	적용(1.5)		0.155	0.145	0.134	0.133	0.114	0.113	0.117	0.118	0.141	0.142
	적용(2)		0.151	0.142	0.129	0.130	0.127	0.127	0.116	0.116	0.141	0.142
BSI 전산업 매출실적	미적용	2.618	2.689	***2.462***	***2.517***	***2.523***	3.029	3.040	***2.586***	***2.591***	2.696	2.701
	적용(0.5)		2.678	***2.440***	2.708	2.717	3.008	3.017	2.898	2.908	2.741	2.752
	적용(1)		***2.567***	***2.446***	2.826	2.825	2.782	2.784	3.085	3.083	2.841	2.838
	적용(1.5)		***2.571***	***2.463***	2.646	2.650	2.746	2.749	2.906	2.888	3.151	3.133
	적용(2)		*2.285*	***2.328***	2.685	2.665	2.723	2.723	2.765	2.742	3.640	3.590
BSI 전산업 매출전망	미적용	2.554	2.845	2.564	2.561	2.571	***2.546***	2.574	***2.443***	***2.474***	***2.468***	***2.453***
	적용(0.5)		2.900	***2.436***	2.565	2.566	2.568	2.581	***2.540***	2.561	***2.271***	*2.241*
	적용(1)		2.765	***2.487***	2.724	2.729	2.623	2.639	3.144	3.125	2.949	2.897
	적용(1.5)		2.582	***2.446***	***2.542***	2.571	2.611	2.618	3.120	3.088	3.869	3.800
	적용(2)		***2.381***	***2.301***	2.566	2.589	2.598	2.595	2.840	2.732	4.187	4.097

주 : 1) 2021.1~2022.12 표본외 예측오차 평균
2) 첫 번째 차원축소 변수를 이용한 경우
3) 첫 번째, 두 번째 차원축소 변수를 이용한 경우

〈표 8〉

월별 경제 지표 예측 평균오차(MAE) 비교[1)]

(분석대상기간: 2019년 1월 ~ 2022년 12월)

모형 / 변수	변동성 안정화 lambda	AR	원 MA 텍스트 지표		차원축소된 MA 텍스트 지표		원 TFI + 원 MA 텍스트 지표		차원축소된 TFI & 차원축소된 MA 텍스트 지표	
			Linear Regression	Lasso Regression	Seed[2)]-Linear Regression	Seed-Lasso Regression	Linear Regression	Lasso Regression	Seed-Linear Regression	Seed-Lasso Regression
선행지수 순환 변동치	미적용	0.167	1.186	0.242	0.185	0.186	1.845	0.189	0.185	0.186
	적용(0.5)		4.270	0.176	0.198	0.213	4.551	*0.146*	0.198	0.214
	적용(1)		1.457	0.180	0.191	0.186	2.001	0.190	0.192	0.186
	적용(1.5)		6.303	0.189	0.180	0.173	2.990	0.199	0.181	0.180
	적용(2)		1.454	0.185	0.174	***0.164***	7.710	0.193	0.174	0.168
BSI 전산업 매출실적	미적용	2.642	73.228	2.959	5.172	5.166	135.319	*2.111*	5.173	5.167
	적용(0.5)		235.864	3.134	4.047	4.153	329.866	***2.270***	4.041	4.145
	적용(1)		40.759	***2.375***	3.730	3.766	191.953	***2.559***	3.807	3.847
	적용(1.5)		416.203	***2.396***	3.613	3.779	23.694	2.784	3.641	3.813
	적용(2)		77.485	***2.444***	3.613	3.619	32.270	3.328	3.631	3.639
BSI 전산업 매출전망	미적용	*1.772*	141.230	2.915	3.244	3.404	38.651	2.159	3.247	3.406
	적용(0.5)		42.587	3.586	3.552	3.674	63.975	2.336	3.547	3.669
	적용(1)		45.885	2.215	2.348	2.458	33.757	2.536	2.498	2.563
	적용(1.5)		64.845	2.141	2.447	2.491	237.391	2.705	2.459	2.501
	적용(2)		60.346	2.111	2.575	2.502	84.724	2.502	2.590	2.651

주 : 1) 2022.1~2022.12 표본외 예측오차 평균
2) Seeded method

〈표 9〉 **분기별 경제 지표 예측 평균오차(MAE) 비교[1)]**

(분석대상기간: 2005년 1분기 ~ 2022년 4분기)

모형 / 변수	변동성 안정화 lambda	AR	원 TFI 지표		일차원으로 축소된 TFI 텍스트 지표[2)]				이차원으로 축소된 TFI 텍스트 지표[3)]			
			Linear Regression	Lasso Regression	SIR-Linear Regression	SIR-Lasso Regression	DR-Linear Regression	DR-Lasso Regression	SIR-Linear Regression	SIR-Lasso Regression	DR-Linear Regression	DR-Lasso Regression
GDP실질 SA 전기비	미적용	0.523	1.348	1.344	0.682	***0.453***	0.650	0.614	0.832	***0.434***	0.631	0.579
	적용(0.25)		1.373	1.346	0.646	***0.464***	0.637	0.597	0.796	*0.404*	0.604	0.572
	적용(0.5)		1.442	1.396	0.536	***0.495***	0.627	0.578	0.734	*0.404*	0.588	0.543
	적용(0.75)		1.455	1.433	***0.502***	***0.508***	0.624	0.577	0.723	0.626	0.570	0.539
	적용(1)		1.464	1.440	***0.479***	***0.519***	0.607	0.565	0.719	0.630	0.554	0.559
GDP실질 원계열 전년 동기비	미적용	1.252	1.259	***0.938***	1.375	1.292	***1.106***	***1.066***	1.431	1.323	***1.043***	***1.047***
	적용(0.25)		***1.225***	*0.922*	1.349	1.273	***1.104***	***1.072***	1.412	1.297	***0.965***	***0.968***
	적용(0.5)		***1.062***	***0.925***	1.312	1.270	1.393	1.323	1.338	1.308	1.763	1.696
	적용(0.75)		***1.004***	***0.949***	1.319	1.277	1.390	1.321	1.319	1.277	1.390	1.321
	적용(1)		***0.973***	***0.980***	1.332	1.286	1.383	1.315	1.407	1.341	1.799	1.725

주 : 1) 2021.1~2022.4분기 표본외 예측오차 평균 2) 첫 번째 차원축소 변수를 이용한 경우
3) 첫 번째, 두 번째 차원축소 변수를 이용한 경우