서울 아파트매매가격지수 예측을 위한 베이지안 변수선택 기법*

이 창 훈** · 강 규 호*** · 안 지 희****

논문초록

정확한 장단기 주택가격 예측은 효율적인 부동산 정책수립과 운용, 부동산 투자 및 은행의 담보대출 위험관리 등에 필수적임에도 불구하고 잠재적인 예측변수의 수가 너무 많아 예측을 실시하는 데 기술적 어려움이 존재한다. 본 연구는 이러한 예측변수와 모형 불확실성 문제를 고려함으로써 보다 정확한 장단기 서울 아파트 매매가격지수 예측분포를 도출하고자 기존의 베이지안 변수선택 기법을 보완 및 확장한 새로운 변수선택 알고리즘을 제안한다. 본 연구에서 제시된 변수선택기법은 각 예측변수의 선택여부 뿐만 아니라 종속변수에 미치는 영향의 방향까지 식별한다는 점에서 기존 기법과 차별화된다. 표본 외 예측결과, 중장기 예측에서 본 연구의 베이지안 변수선택 기법을 적용한 모형이 다른 모든 경쟁모형들보다 예측력에서 우월한 것으로 나타났다. 반면 서울 아파트매매가격지수의 높은 지속성으로 인해 단기 예측에서는 p차 자기회귀모형의 예측력이 가장 우수했다.

핵심 주제어: 표본외 분포예측, 깁스-샘플링 알고리즘, 모형선택
경제학문헌목록 주제분류: R2, C11, C53

투고 일자: 2019. 9. 5. 심사 및 수정 일자: 2020. 2. 7. 게재 확정 일자: 2020. 3. 20.

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A2A01037796). 논문의 질적 개선에 큰 도움이 되는 중요한 논평을 해주신 익명의 심사위원께 대단히 감사드립니다.

** 제1저자, 고려대학교 경제학과 박사과정생, e-mail: rollin0807@korea.ac.kr
*** 교신저자, 고려대학교 경제학과 부교수, e-mail: kyuho@korea.ac.kr
**** 공동저자, 한국부동산연구원 부연구위원, e-mail: annji@kreri.re.kr

Ⅰ. 서 론

1. 연구배경 및 목적

정확한 장단기 주택가격 예측은 주택연금 산정, 부동산 규제의 완화 또는 강화, 규제 적용 범위 및 시점 결정, 가계부채 모니터링 및 대책 수립, 국민연금 자산 운용, 한국은행의 기준금리 결정 등 각종 부동산 관련 정책 및 제도 운용의 효율성 극대화에 필수적이다. 하지만 정책적인 필요성에도 불구하고 국내 주택가격의 통계적 예측연구는 찾아보기 힘든데, 실제 분석과정에서 직면하게 되는 불확실성 문제에 기인한 것으로 보인다.

예측에서 불확실성 문제를 일으키는 요인은 크게 두 가지로 구분된다. 첫 번째는 예측변수의 불확실성이다. 일반적인 선형회귀모형 하에서 K를 잠재적으로 가능한 예측변수의 수라고 할 때, K개의 예측변수들 중에서 종속변수인 주택가격 예측을 위해 어떤 변수를 선택할 것인지에 대한 불확실성이 존재한다. 일반적으로 주택가격은 물가, 경기변동, 금리, 주가 등 대부분의 거시 및 금융변수와 밀접한 관계를 갖기 때문에 K가 아주 큰 값을 가진다. 이 때문에 주택가격 예측연구는 전형적인 예측변수 불확실성 문제에 직면하게 된다.[1)]

두 번째는 모형 불확실성 문제이다. 모형 불확실성은 예측에서 최적 예측변수 집합 또는 최적시차 등이 시변할 가능성이 존재한다는 것을 의미한다. 우리가 특정시점에서 예측변수 불확실성 문제를 해결하여 가장 큰 예측력이 기대되는 최적모형을 찾더라도, 그 모형이 모든 시점에 대해서도 항상 가장 좋은 예측력을 보일 것이라고 기대하기는 어렵다. 모형의 예측력은 규제나 주택수요자의 선호변화 등 주택시장 내부의 구조적인 변화뿐만 아니라 경기 및 금리 변동 등 주택시장 외부환경에 인해 시변할 가능성이 크기 때문이다.

불확실성 문제 하에서 정확한 예측이 이루어지려면 예측변수 불확실성을 반영하여 최적 예측변수들을 찾아내는 변수선택 기법이 필요하고, 변수선택 알고리즘은 특정시점에 국한되는 것이 아니라 모든 예측시점에 대해서 동태적으로 시행되어야

---

1) 주어진 K로부터 매 시점 우리가 예측을 위해 고려해야 하는 총 모형의 수는 $J=2^K$와 같다. 예를 들어, K가 20개 이면, 총 모형의 수 J는 무려 1,048,576개에 달한다.

한다. 불확실성 문제를 고려하지 않으면 정확한 예측을 기대하기 어려우며, 의사결정의 효율성 극대화 측면에서 정책적 판단에 도움이 되는 정보를 제공하지 못할 가능성이 크다.

본 연구의 목적은 불확실성 문제를 고려한 정확한 주택가격 예측을 위해서, 베이지안 변수선택 기법(Bayesian Variable Selection)을 이용한 주택가격 분포예측(density forecast) 알고리즘을 개발하고 표본 외 예측을 통해 예측력을 비교평가하는 것이다. 주택가격의 지표로 전년동월대비 서울 아파트매매가격지수를 이용하며, 다양한 베이지안 변수선택 기법들 중에서 George and McCulloch(1993)의 베이지안 변수선택 기법(이하, BVS)을 수정 및 보완하여 표본 외 예측을 실시하고, 벤치마크 모형들과 예측력을 비교 평가한다.

베이지안 분포예측은 일반적인 점예측에 비해 정책적 판단에 있어 위험관리에 더욱 용이하다는 장점이 있다. 예를 들어, 임계치가 주어져 있을 때, 분포로부터 향후 주택가격이 임계치를 상회할 확률을 계산할 수 있으며 이와 같은 정보는 주택가격 급등조기경보지수로 의사결정자가 정책판단을 하는데 유용하게 사용될 수 있다.

다양한 베이지안 변수선택 기법들 중에서 BVS 기법을 차용한 이유는 상대적으로 다른 기법들에 비해 변수선택의 정확성을 잃지 않으면서, 컴퓨팅 속도 측면에서 월등히 빠르기 때문이다. 베이지안 접근법에는 베이지안 모형 평균(Bayesian Model Averaging, BMA)과 같이 오래전부터 전통적으로 사용되는 기법뿐만 아니라 Raftery et al. (2010)의 동태적 모형 평균(Dynamic Model Averaging, DMA), Onorante and Raftery(2016)의 Dynamic Occam's Window(DOW) 등과 같이 비교적 최근에 개발되어 예측력 향상이 기대되는 기법들이 있다. BMA와 같은 전통적인 기법은 이론적 근거가 존재하며 가장 표준적이지만, 방법론상 모형들의 가중치로 사용되는 주변우도(Marginal Likelihood)를 계산하는 과정에서 K의 수가 50보다 크면 과도한 계산비용이 발생하여 현실적으로 적용하기 어려운 단점이 존재한다. DMA 기법은 BVS 기법에 비해 상대적으로 변수선택이 정확할 것으로 기대되지만 K가 아주 작은 경우에만 적용 가능하다는 단점이 있다. BMA와 마찬가지로 알고리즘 내에서 과도한 연산과정을 요구하여 K가 10보다 크면 컴퓨터 메모리 부족 문제로 프로그램 실행이 쉽지 않다. DOW 기법은 DMA의 컴퓨팅 문제를 개선한 방법으로 DMA보다 축소된 모형 공간에서 최적화하는 기법이지만, 연구가

시뮬레이션 실험에서만 이루어지고 논문에서 구체적인 알고리즘을 제시하지 않아 적용하기 어려운 단점이 존재한다. 반면, BVS 기법은 베이지안 방법론에서 표준적으로 사용되는 변수선택 기법이면서 모형의 형태가 일반적인 선형회귀모형에 기반 하기 때문에 K의 수가 50보다 큰 경우에도 비교적 다른 변수선택 기법에 비해 적용하기 쉬우며 컴퓨팅 속도가 월등히 빠르다는 장점이 있다.

본 논문에서는 BVS 기법을 그대로 적용하지 않고 예측력 향상을 도모하기 위해 알고리즘의 일부를 수정하고 보완하였다. 첫째, BVS 기법은 알고리즘 내에서 연구자에 의해 설정되는 조정 파라메터(tuning parameter)에 민감하게 반응하여 파라메터 값이 적절하게 설정되지 않을 시, 변수선택 정확성이 크게 떨어진다는 단점이 있다. 이와 같은 문제를 완화하기 위해서 본 연구에서는 조정 파라메터 또한 추정해야 하는 모형 파라메터로 가정하여 계층 모형(hierarchical model)을 설정하고 사후 샘플링 하였다. 둘째, 특정 예측시점에 대해서 한 번만 변수선택을 실시하는 BVS 기법과 다르게 동 시점에 변수선택을 여러 번 반복하여 선택된 예측변수들 중에서도 임의의 기준을 만족하지 못하는 예측변수들은 제외되도록 하였다. 마지막으로 예측변수선택의 정확도를 높이기 위해 예측변수가 종속변수에 미치는 영향이 양인지 음인지를 구분되어 선택되도록 설정하였다.

본 연구의 베이지안 변수선택 알고리즘의 예측력을 비교평가 하기 위해 두 가지의 벤치마크 모형군을 설정하였다. 첫 번째 모형군은 종속변수의 시차변수만 예측변수로 고려한 1차 자기회귀모형(First-order Auto Regressive model)과 p차 자기회귀모형이다. 두 번째는 BVS 모형군으로 본 연구가 제시한 모형 가정과 분포예측 알고리즘 적용 여부에 따라 BVS을 세분화 한 것이다. 대표적으로 동태적으로 변수선택 하지 않고, 첫 예측시점에만 변수선택 기법을 적용한 모형, George and McCulloch(1993)의 변수선택 기법을 적용한 모형 등이 있다. 벤치마크 모형에 대한 자세한 내용은 Ⅲ장에 설명한다. 예측력 평가 기준은 베이지안 방법론에서 표준적으로 사용되는 사후 예측 정보기준(Posterior Predictive Criterion, PPC)과 평균제곱근오차(Root Mean Square Error, $RMSE$)를 사용한다.

예측결과, 2015년 12월부터 2018년 11월까지 총 3년의 표본 외 예측기간을 종합적으로 고려하였을 때, 본 연구가 제시한 변수선택 기법의 예측력이 나머지 벤치마크 모형들보다 우월한 것으로 나타났다. 특히, 단기를 제외한 중기와 장기 예측에서 서울 아파트매매 가격지수에 대한 예측력이 크게 향상되는 것으로 나타났다. 단

기 예측에서는 p차 자기회귀모형($AR(p)$)의 예측력이 가장 뛰어났다. 이는 서울 아파트매매가격지수의 높은 지속성이 반영된 것으로 단기에서는 지난 기의 가격지수 정보 이외에 추가적인 예측변수의 정보는 불필요하다는 것을 의미한다. 결론적으로 효율적인 주택 정책 수립과 운용, 주택자산관리 및 은행의 주택담보대출 위험관리 등 정확한 중장기 예측이 요구되는 의사결정과정에 본 연구가 기여할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 서론에서는 선행연구와 기존 연구와의 차별성을 추가로 논의한 다음, 제Ⅱ장에서 예측모형을 소개하고, 제Ⅲ장은 예측 알고리즘과 표본외 예측력 평가 기준에 대해 설명한다. 예측결과는 제Ⅳ장에서 다뤄지며, 제Ⅴ장에서 연구결과를 요약하고 추가 연구방향을 제시하고자 한다.

2. 선행연구

국내의 주택가격지수와 관련된 연구는 주택가격의 결정요인분석과 주택가격의 예측모형 및 방법론의 비교 연구로 분류될 수 있다. 우선 주택가격 결정요인에 관한 연구들은 주택시장 변수뿐만 아니라 금융 및 거시경제 변수, 정부정책 등 각 연구에서 선택된 변수들의 중요성을 강조한다.

예를 들어, 윤성민 외(2016)은 다양한 거시경제 변수들이 주택가격에 장, 단기적으로 미치는 영향을 분석하였다. 장기적으로는 이자율, 주택담보대출금, KOSPI지수 등의 변수들이 주요 결정요인인 것으로 나타났고, 단기에서는 이자율, 주택금융신용보증, 경기동행지수 등이 주택가격에 영향을 미치는 것으로 나타났다. 이외에 손종칠(2010), 김윤영(2012), 전해정・박헌수(2012)의 연구에서는 CD금리, 국민소득, 소비지출, 주거용 건설투자, 물가, 환율, 주가, 산업생산지수, 회사채수익률 등이 주택가격의 주요 결정요인으로 보고되었다.

부동산 정책과 관련하여 곽승준・이주석(2006)은 주택가격의 변동성이 변화하는 시점과 부동산 정책 시행시기를 비교하여 정책이 주택가격 변동성에 기여하는 바가 있는지 분석하였다. 연구결과, 몇몇 일치하는 시점이 있지만 2001년 이후에는 2003년의 부동산 정책을 제외하고는 일치하지 않아 정책이 주택시장 안정화에 크게 기여하지 못한다는 결론을 도출하였다. 노영학・김종호(2012)는 부동산 정책을 완화정책과 억제정책으로 구분하여 정부의 정책발표 횟수를 변수로 사용해 아파트매

매가격지수에 미치는 영향을 회귀분석으로 연구하였다. 연구결과, 완화정책은 유효하게 영향을 미치나 미비하고, 억제정책은 오히려 가격을 상승시켜서 정책의 효과가 없다는 결론을 도출하였다. 이와 관련하여 전해정(2014)은 패널 회귀분석을 통해 주택가격에 대해 규제강화정책과 완화정책의 통계적 유의미성이 없음을 보였다.

다음으로 국내의 주택가격 예측모형 및 방법론 비교 연구는 AR, $ARIMA$, VAR, $VECM$ 모형 등의 시계열 모형에 기반한 빈도주의 접근법으로 표본 외 예측을 실시하고 예측오차를 기준으로 예측력을 비교하는 연구가 주를 이룬다. 비교적 최근에는 주택가격 예측력의 향상을 위해 베이지안 방법론을 도입하거나 머신러닝 기법을 이용해 기존의 시계열 모형들과 비교하는 연구가 활발히 진행 중이다.

손정식 외(2002)는 실질 GDP 성장률, 회사채수익률 등의 거시변수를 이용한 VAR 모형과 $ARIMA$ 모형으로 주택매매가격 변동률에 대한 예측력을 비교하였다. 예측결과, VAR 모형의 예측력이 $ARIMA$ 모형에 비해 상대적으로 우수한 것으로 나타났다.

이영수(2014)는 단일변수 시계열 모형들의 예측력을 비교하기 위해서 $ARIMA$, $GARCH$, 국면전환, 비관측요인 모형들을 사용하여 전국 아파트가격지수를 예측하였다. 예측결과, 표본 외 예측에서는 비관측요인 모형의 예측력이 가장 우수한 것으로 보고하였다.

김경외 · 김영효(2015)는 모형 불확실성을 반영하기 위해서 베이지안 방법론을 적용하여 서울과 부산의 아파트매매 가격지수를 예측하였다. BMA 기법을 이용하여 총 11개의 거시변수를 대상으로 최적 모형을 찾았고, 표본 외 예측력 측면에서 BMA 기법이 AR 모형보다 뛰어나다는 결과를 도출하였다. 함종영 · 손재영(2016)은 주택가격에 대해서 일반적인 VAR과 베이지안 VAR($BVAR$) 모형의 예측력을 비교 · 평가하여 $BVAR$ 모형의 예측력이 VAR 모형보다 우수하다는 것을 보였다.

송상윤(2016)은 실질가계부채증가율, 실질가계소득증가율, 대출금리, 주택가격등락률의 총 4개 변수를 이용한 $BVAR$ 모형 하에서 각 변수들의 동태적 상관관계를 분석하고, 표본 외 예측을 실시하였다. 추정 및 예측에서 $BVAR$ 모형에 2가지의 베이지안 변수선택 기법을 적용하였는데, 첫 번째는 Korobilis(2013)의 변수선택(Variable Selection, $BVAR$-VS) 기법이고, 두 번째는 George, Sun and Ni

(2008)의 Stochastic Search Variable Selection($BVAR$-$SSVS$) 기법이다.[2] 연구결과, 주택가격등락률에 대해서 $BVAR$-$SSVS$의 예측력이 일반적인 VAR, $BVAR$, $BVAR$-VS 보다 우수한 것으로 나타났다.

배성완・유정석(2018)은 심층신경망, LSTM(Long Short Term Memory networks) 등의 머신 러닝 방법을 적용한 모형들과 시계열 모형인 AR, VAR, $BVAR$ 모형들의 서울 아파트매매가격지수에 대한 예측력을 비교하였다. 예측결과, 머신 러닝 방법이 기존의 시계열 모형들보다 예측력이 뛰어나다는 결과를 도출하였다.

3. 선행연구와의 차별성

주택가격의 결정요인과 관련해서 선행연구들로부터 소수로 구성된 예측변수들의 집합을 생성하기는 쉽지 않다. 기존 연구마다 사용하는 변수들의 범위가 넓고, 기존 연구들의 결과를 직접적으로 비교하는 것이 불가능하며, 시차까지 고려되면 변수의 수가 기하급수적으로 증가하기 때문이다. 따라서 예측변수 불확실성은 주택가격 분석에서 반드시 고려되어야 하며, 이를 반영하는 방법론 중에서 쉽게 사용할 수 있는 방법이 베이지안 변수선택 기법이다.

본 연구는 기존 국내연구들의 결과를 참고하여 주택시장의 변수뿐만 아니라 다양한 금융 및 거시경제 변수들을 잠재적인 예측변수로 선정하였다. 참고로, 부동산 정책의 경우에는 수치화된 대용변수가 존재하지 않고 적절한 대용변수를 고안하는 것이 어려워 불가피하게 예측변수 집합에 포함되지 않았다.

기존의 주택가격 예측은 빈도주의 방법으로 AR, $ARIMA$, VAR 등의 시계열 모형을 이용하여 점예측을 하는 것이 일반적이다. 하지만 점예측은 방법론상으로 간단하다는 장점이 있는 반면, 미래 주택가격의 변동 가능폭이 예측되지 않기 때문에 위험관리 측면에서 베이지안 분포예측에 비해 상대적으로 유용성이 제한적이다. 또한 본 연구에서 사용한 베이지안 변수선택 기법은 예측변수와 모형의 불확실성을

2) Korobilis(2013)와 George, Sun and Ni(2008)의 기법은 변수선택에서 예측변수의 계수에 대한 가정의 차이가 있다. 예를 들어, k-번째 예측변수가 변수선택 단계에서 선택되지 않으면 Korobilis(2013)는 k-번째 예측변수의 계수 α_k를 0이라고 가정하고, George, Sun and Ni(2008)는 0에 가깝도록 제약을 부여할 뿐 0으로 고정 시키지는 않는다. 여기서 George, Sun and Ni(2008)의 기법은 George and McCulloch(1993)의 기법을 VAR 모형에 적용되도록 확장한 것이다.

동시에 고려할 수 있다는 이점이 있다.

베이지안 방법론에 기반한 선행연구들 중에서 본 연구와 가장 유사한 것은 김경외 · 김영효(2015) 와 송상윤(2016) 의 연구이다. 김경외 · 김영효(2015) 는 모형 불확실성을 반영하기 위해 11개의 예측변수들을 이용하여 총 2,048개의 모형들을 대상으로 BMA 기법을 적용하였다. BMA는 베이지안 방법론에서 가장 표준적인 예측기법이지만 과도한 계산비용 때문에 BVS 기법에 비해 상대적으로 적은 예측변수를 사용한 것으로 보인다. 송상윤(2016) 은 4개의 거시변수를 사용하여 $BVAR$ 모형에 George, Sun and Ni(2008) 의 변수선택 기법을 적용하였다. VAR은 모형의 특성으로 인하여 변수의 수가 아주 클 경우, 시차가 늘어날수록 계수의 수가 기하급수적으로 증가하기 때문에 주택가격 예측에 적합하지 않다.

본 연구는 20개의 예측변수와 종속변수를 제외한 예측변수들의 최대시차를 24까지 고려하여 총 480개의 예측변수들을 사용하였다. 또한 동태적 변수선택을 실시하여 예측변수와 모형의 불확실성이 충분히 반영될 것으로 기대된다.

국내연구들 중에서 예측변수의 수가 아주 큰 경우에 예측변수와 모형의 불확실성을 동시에 고려한 연구는 강규호(2018) 가 있다. 강규호(2018) 는 AR, VAR, ADL (Autoregressive Distributed Lag) 모형에 베이지안 머신 러닝 기법을 적용하여 우리나라 가계대출을 예측하였다.[3] 베이지안 머신 러닝 알고리즘은 변수선택, 모형선택, 예측조합 단계로 이루어져 있는데, 변수와 모형선택 이후, 하나의 모형이 아닌 선택된 여러 모형으로 예측을 실시하고 예측분포들을 조합하는 방법이다.

강규호(2018) 는 변수선택 기준과 예측분포 조합의 가중치로 사후예측우도(Posterior Predictive Likelihood, PPL) 을 사용한다. PPL은 베이지안 방법론에서 표준적으로 사용되는 예측력 평가 기준이며, 예측구간에 대해서 예측대상의 실제 실현치가 예측분포에서 갖는 사후밀도들의 평균치에 해당한다. PPL은 베이지안 방법론에서 이론적으로 가장 좋은 예측력 평가 기준이지만 실증분석에서는 예측구간이 클수록 모형의 예측력을 과소 또는 과대평가할 가능성이 크고, 변수선택이 훈련표본(training sample) 에 민감하며, 계산상 부담도 BVS 기법에 비해 상대적으로 훨씬 크다.[4] 이와 같은 이유로 본 연구는 PPL 대신 BVS 기법을 선택하였다.

---

3) 강규호(2018) 의 머신 러닝 기법은 지도학습(supervised learning) 이다.

4) 예를 들어, 어떤 모형 M이 특정 예측구간에 대해서 한 시점을 제외한 나머지 모든 예측시점에서는 최적모형이라고 하자. 만약, 예측력이 떨어지는 시점에서 모형 M의 예측분포가 실제

Ⅱ. 모 형

예측모형과 분포예측 알고리즘을 설명하기에 앞서 본 연구에서 사용된 BVS 기법과 파라메터 사후 샘플링 방법에 대해 먼저 설명하고자 한다. 다음의 전형적인 베이지안 선형회귀모형을 고려하자.

$$Y|\beta,\sigma^2 \sim N(X\beta,\sigma^2 I_T)$$

단, T는 표본크기, K는 잠재적으로 가능한 예측변수의 수, $Y=(y_1,y_2,...,y_T)'$는 종속변수, $\beta=(\beta_1,\beta_2,...,\beta_K)'$는 계수벡터, $x_t=(x_{1,t},x_{2,t},...,x_{K,t})'$는 t시점의 예측변수벡터, $X=(x_1,x_2,...,x_T)'$는 예측변수행렬을 나타낸다.

K개의 예측변수가 주어져 있을 때, BVS 기법에서 변수선택을 한다는 것은 선택되지 않은 예측변수들의 계수 값이 0이 되도록 하여 종속변수에 영향을 주지 않도록 제약하는 것과 동일하다. 즉, 예측변수들의 표본분산을 1로 표준화하였을 때, 깁스 샘플링 알고리즘 과정에서 선택되지 않은 예측변수의 계수는 0에 아주 가깝게 샘플링 되어야 한다. 이를 위해, 표준적인 깁스 샘플링 알고리즘과 달리 BVS 기법은 β가 서로 다른 크기의 사전 분산을 갖는 혼합 정규 사전분포를 따른다고 가정하며, $k=1,2,...,K$에 대해 $\Pr[\gamma_k=1]$의 확률로 베르누이 분포를 따르는 파라메터 γ_k를 도입한다. 단,

$$\Pr[\gamma_k=1]=\Pr[\gamma_k=0]=0.5$$

이며, γ_k는 k-번째 예측변수의 계수 β_k의 사전분산 b_k에 영향을 주는 파라메터이다. 사전분산 b_k는 주어진 γ_k로부터 다음과 같이 결정된다.

$$b_k = I(\gamma_k=1)\times B_1 + [1-I(\gamma_k=1)]\times B_0$$

실현치를 극단치로(outlier)로 취급하면 PPL값이 비정상적으로 계산되어 모형 M의 예측력을 과소평가하게 된다.

단, $B_0 < B_1$이며, 일반적으로 B_0에 대해서는 0에 가까운 값을, B_1에 대해서는 상대적으로 충분히 큰 값을 가정한다. 여기서 γ_k가 주어져 있을 때, $k = 1,2,...,K$에 대해 β_k의 사전분포는 다음과 같다.

$$\beta_k|\gamma_k \sim N(0, b_k).$$

만약 $\gamma_k = 0$이면, 사전평균이 0이고 분산이 $b_k = B_0$이므로 β_k는 0에 가까운 값이 샘플링될 것이다. 반면, $\gamma_k = 1$인 경우에는 분산이 충분히 크므로 0이 아닌 충분히 크거나 작은 값이 샘플링될 것이다. 따라서, γ_k의 값은 $k-$번째 예측변수 $x_{k,t}$가 모형에 포함되는지의 여부를 나타낸다. 즉, $\gamma_k = 1$이면 예측변수 $x_{k,t}$가 선택되었다는 것을 의미한다.

분산 σ^2에 대해서는 표준적인 깁스 샘플링에서와 같이 켤레분포인 역감마분포를 가정한다.

$$\sigma^2 \sim IG(\nu_0/2, \delta_0/2).$$

여기서 σ^2는 분석의 편의를 위해 $\gamma(=(\gamma_1,\gamma_2,...,\gamma_K)')$와 독립이라고 가정한다.[5] 모형 파라메터의 결합사후분포는 $\beta,\sigma^2,\gamma|Y$이고, B_0와 B_1은 조정 파라메터로서 β 사후분포의 서포트를 잘 반영하도록 연구자에 의해 설정되어야 한다. 지금까지 설명한 모형이 George and McCulloch (1993)의 BVS 기법에서 가장 표준적인 모형 설정이다.

위와 같은 표준적인 BVS 기법은 추정이 쉬우면서 예측변수의 수가 큰 경우에도 다른 베이지안 변수선택 기법에 비해 계산비용이 적다는 장점이 있지만, 모형의 단순성으로부터 두 가지 한계점을 가지고 있다. 첫째, 지시변수인 γ_k의 값으로부터 k번째 예측변수가 모형에 포함되는지 여부는 판단이 가능하지만 해당 예측변수가 종속변수에 양의 영향을 주는지 음의 영향을 주는지의 방향성은 직접적으로 판단할 수 없어 변수선택 결과를 해석하는데 불편함이 발생한다. 둘째, 변수선택 결과가

5) σ^2은 β를 통해서 γ와 상관관계가 존재할 수 있다. 예를 들어, 선택되는 변수가 많아져 β의 차원이 커질수록 σ^2은 작아질 수 있다.

조정 파라메터인 B_0와 B_1에 민감하게 반응하여, 적절한 값으로 설정하지 않으면 정확한 변수선택을 기대하기 어려우며 결과적으로 예측력 또한 떨어진다.[6] B_0와 B_1을 적절한 값으로 설정한다는 것은 β_k의 사전분포가 실제 β_k의 서포트를 잘 반영하도록 설정하는 것을 의미하지만 β_k는 알려지지 않은 확률변수이므로 자료의 정보에 근거하여도 설정이 쉽지 않다.

본 연구에서는 위와 같이 전통적인 BVS 기법에서 발생하는 한계점을 완화하여 정확한 변수선택과 예측력을 극대화하기 위해서 γ_k와 β_k의 사전분포에 대한 가정을 변형하는 한편, 조정 파라메터들을 모형 파라메터로 간주하여 연구자에 의해 임의로 설정되는 것이 아니라 알고리즘 내에서 사후샘플링 되도록 하는 계층모형(hierarchical model)의 형태를 가정하였다. 구체적으로 γ_k는

$$\Pr[\gamma_k = 1] = \Pr[\gamma_k = -1] = \Pr[\gamma_k = 0] = 1/3$$

의 확률로 1, -1, 또는 0의 값을 갖는다. 또한 γ_k가 주어져 있을 때, β_k의 사전평균 μ_k와 사전분산 b_k은 다음과 같이 결정된다.

$$b_k = I(\gamma_k = 1) \times B_1 + I(\gamma_k = -1) \times B_{-1} + I(\gamma_k = 0) \times B_0,$$
$$\mu_k = I(\gamma_k = 1) \times \overline{\beta_1} + I(\gamma_k = -1) \times \overline{\beta_{-1}} + I(\gamma_k = 0) \times \overline{\beta_0},$$

단, $\overline{\beta_1} > 0$, $\overline{\beta_{-1}} < 0$, $\overline{\beta_0} = 0$이며, $\beta_k | \gamma_k \sim N(\mu_k, b_k)$이다. 여기서 B_{-1}은 B_1과 동일한 값을 갖는다고 가정한다. 기존의 모형 설정에서는 β_k의 사전평균이 γ_k에 관계없이 항상 0이었던 것과 다르게 γ_k의 값에 따라 사전평균이 양의 값, 음의 값, 또는 0을 갖는다. 따라서 γ_k 값은 변수선택의 여부 뿐 만 아니라 예측변수가 종속변수에 어떤 영향을 주는지까지 나타낸다. 예를 들어, $\gamma_k = -1$은 $k-$번째 예측변수 $x_{k,t}$가 모형에 포함되며 종속변수에 음의 영향을 준다는 것을 의미한다.

모형 파라메터들의 결합사후분포(joint posterior distribution)는 기존의 $\beta, \sigma^2, \gamma | Y$에서

6) O′Hara and Sillanpää(2009)를 참고하기 바란다.

$$\beta, \sigma^2, \gamma, \overline{\beta}, B \mid Y$$

로 확장되어 계층모형의 형태를 취한다. 단, $\overline{\beta} = (\overline{\beta_1}, \overline{\beta_{-1}})$이며 $B = (B_1, B_{-1}, B_0)$이다. $i = 1, -1$에 대해서 $\overline{\beta_i}$의 사전분포는 $N(\overline{\mu_i}, V_{\overline{\beta_i}})$를 가정하고, B_1, B_{-1}, B_0의 사전분포는 σ^2와 동일하게 역감마분포를 가정한다.

1. 파라메터 사후 샘플링

본 연구의 모형은 계층모형이지만, 선형회귀모형에 기반하므로 복잡한 알고리즘을 요구하지 않고, 깁스샘플링 기법으로 간단하게 추정 할 수 있는 장점을 갖고 있다. 깁스샘플링은 시뮬레이션의 매 반복시행에서 각 파라메터들의 완전 조건부 분포(full conditional distribution)로부터 파라메터들을 샘플링한다. 모형 파라메터들의 완전 조건부 분포와 샘플링 순서는 다음과 같이 요약된다.

〔모형 파라메터 깁스샘플링〕

1. β를 $\beta \mid Y, \gamma, \overline{\beta}, B, \sigma^2$에서 샘플링한다.
2. σ^2를 $\sigma^2 \mid Y, \beta$에서 샘플링한다.
3. γ를 $\gamma \mid \beta$에서 샘플링한다.
4. $\overline{\beta_1}$와 $\overline{\beta_{-1}}$를 각각 $\overline{\beta_1} \mid \beta, B_1, \gamma$와 $\overline{\beta_{-1}} \mid \beta, B_{-1}, \gamma$에서 샘플링한다.
5. B_1, B_{-1}, B_0를 각각 $B_1 \mid \beta, \gamma, \overline{\beta_1}$, $B_{-1} \mid \beta, \gamma, \overline{\beta_{-1}}$, $B_0 \mid \beta, \gamma$에서 샘플링한다.

가정에 따라 σ^2는 γ에 의존하지 않는다. γ는 β를 통해서만 Y에 영향을 주기 때문에 γ를 샘플링하기 위해서는 β의 정보만 있으면 된다. 따라서 γ의 완전 조건부 분포는 자료 Y에 의존하지 않는다. B_1, B_{-1}, B_0은 모든 k에 대해서 γ_k가 주어져 있을 때 β_k의 사전분산이므로 β, γ, $\overline{\beta}$에만 의존한다. $\overline{\beta_1}$와 $\overline{\beta_{-1}}$ 또한 β_k의 사전평균이므로 β, γ, B의 정보만으로 사후샘플링이 가능하다.

Ⅲ. 베이지안 변수선택 기법을 이용한 분포예측

본 장에서는 예측모형과 연구에서 진행되는 표본 외 예측과정, 베이지안 변수선택 기법의 예측력 평가 방법 및 서울 아파트매매가격지수 분포예측 알고리즘에 대해 설명한다.

1. 예측모형

본 연구의 예측대상인 서울 아파트매매가격지수 예측을 위한 예측모형에 대해 설명한다. HP_t를 t 시점 서울 아파트매매 가격지수라고 할 때, 종속변수 y_t는 t 시점의 전년동월대비 서울 아파트매매 가격지수 증가율로 단위는 퍼센트(%)이며, $y_t = \ln HP_t - \ln HP_{t-12}$로 계산된다. 예측시계($h$)에 대해서 h-기 이후 전년동월대비 증가율 예측치는 y_{t+h}이며, y_{t+h}가 주어져 있을 때, h-기 이후 서울 아파트매매가격지수 수준 예측치(HP_{t+h})는

$$HP_{t+h} = \exp[y_{t+h} + \ln HP_{t+h-12}]$$

로 계산된다.

종속변수의 시차가 p이고, 예측변수의 최대시차가 q인 h-기 앞 전년동월대비 서울 아파트매매가격지수 증가율 예측모형은 $h = 1,2,\ldots,H$에 대해서

$$y_{t-1+h} | F_{t-1}, \theta \sim N\left(\phi + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{i=1}^{K}\sum_{j=1}^{q} \beta_{i,j} x_{i,t-j}, \sigma^2\right) \quad (3.1)$$

로 표현된다. F_t는 t 시점까지의 모든 관측 자료의 집합이며, θ는 모형 파라메터들의 집합이고, H는 최대 예측시계이다.

예측모형은 수식의 표현만 다를 뿐, 선형회귀모형이므로 앞 장에서 설명한 깁스 샘플링 방법으로 동일하게 추정 및 분석이 가능하다. 변수선택과 관련해서 동일한 설명변수라도 시차가 다르면 서로 다른 예측변수로 간주한다. 예를 들어, $x_{1,t-1}$과 $x_{1,t-2}$는 서로 다른 예측변수로 $x_{1,t-1}$의 선택이 $x_{1,t-2}$의 선택을 보장하는 것은

아니다. 추가로 일반적인 예측모형에서 상수항은 항상 포함되므로, 상수항은 변수선택 대상에서 제외한다.

2. 표본 외 예측

베이지안 변수선택 알고리즘에 대해 구체적으로 설명하기에 앞서 본 연구에서 사용되는 예측기법과 표본 외 예측과정을 간략하게 설명한다.

본 연구에서 예측은 직접예측기법(direct forecasting)에 기반하며, 예측과정에서 추정은 이동회귀기법(rolling regression)을 사용한다. 예측변수 중요도의 시변성을 고려하여 과거에 중요한 예측변수의 가중치를 줄여 더 정확한 변수선택을 하기 위한 것이다.

전년동월대비 서울 아파트매매가격지수의 전체 표본 크기를 T라고 하고, 표본 외 예측 구간의 크기(out-of-sample size)를 OSS라고 하자. 우리의 관심은 예측시계(h)별로 OSS 구간의 매 시점을 주어진 자료(in-sample)를 이용하여 예측하는 것이다. $T_0 = T - OSS$로 정의하면, 우리가 예측해야 할 표본 외 예측구간은 T_0+1 시점부터 $T_0 + OSS(= T)$ 시점까지이다. 여기서 T_0를 training 표본의 크기라고 한다. 주의할 것은 표본 외 예측에서 예측시계 h의 역할이다. 우리는 h 값에 관계없이 T_0+1 시점부터 T 시점까지 항상 동일한 구간을 예측한다. 단, h는 동일한 구간을 예측할 때, 사용되는 종속변수 자료의 정보량에 영향을 주며 단기일수록 더 많은 정보를 사용하고, 장기일수록 정보의 양이 줄어든다.[7]

3. 예측력 평가 방법

본 연구의 목적은 전년동월대비 서울 아파트매매 가격지수의 분포예측을 위한 베이지안 변수선택 알고리즘을 개발하고 예측력을 비교평가 하는 것이다. 변수선택 기법 하에서의 모형과 벤치마크 모형들의 예측력 평가를 위해 본 논문에서는 베이

---

7) 예를 들어, $h=1$이고 $OSS=2$인 경우에 T_0+1 시점과 T_0+2 시점의 예측분포는 각각 $y_{T_0+1}|F_{T_0}$와 $y_{T_0+2}|F_{T_0+1}$와 같다. 단, $F_{T_0}=(y_1,y_2,...,y_{T_0})$. 반면, $h=2$인 경우에는 $y_{T_0+1}|F_{T_0-1}$과 $y_{T_0+2}|F_{T_0}$로 사용되는 정보의 양이 하나씩 줄어든다.

지안 분포예측에서 표준적으로 사용되는 예측력 평가 기준인 사후 예측 정보기준(Posterior Predictive Criterion, 이하 PPC)과 평균 제곱근 오차(Root Mean Square Error, 이하 $RMSE$)를 이용한다.

MCMC 시뮬레이션의 크기가 N일 때, 분포예측이 완료되면 예측시계(h) 별로 모든 표본 외 예측구간 시점에 대해 N개의 예측치로 구성된 OSS개의 예측분포가 도출된다. 특정 h와 $g=1,2,...,OSS$에 대해 예측분포 $\left\{y_{T_0+g}^{(j)}\right\}_{j=1}^{N}$가 주어져 있을 때, T_0+g 시점의 PPC 값을 $PPC(h,g)$라고 하자. $PPC(h,g)$는 다음과 같이 예측분포의 분산과 예측오차의 제곱의 합으로 정의된다.

$$PPC(h,g) = Var(y_{T_0+g}) + [y_{T_0+g}^{r} - E(y_{T_0+g})]^2$$

단, $Var(y_{T_0+g})$는 T_0+g 시점에 대한 예측분포의 분산이며, $E(y_{T_0+g})$는 동 예측분포의 평균이고, $y_{T_0+g}^{r}$는 T_0+g 시점의 종속변수 실현치이다. 정의에 의해 PPC값이 작으면 작을수록 더 좋은 예측력을 의미한다.

4. BVS 분포예측 알고리즘

이 장에서는 BVS 기법을 이용한 전년동월대비 서울 아파트매매 가격지수의 분포예측 알고리즘에 대해서 설명한다. 알고리즘은 크게 변수선택 단계와 분포예측 단계로 구성되며, 세부적으로는 총 5 단계로 실시된다. 구체적인 설명에 앞서 명칭에 의한 혼란을 방지하기 위해 본 연구에서 제시하는 BVS 기법을 $BVS(3)$ 기법 또는 모형이라 하고, George and McCulloch(1993)의 BVS 기법을 $BVS(2)$ 기법 또는 모형이라고 칭한다. 괄호안의 숫자는 지시변수 γ가 취할 수 있는 값의 수를 의미한다.

변수선택 단계에서는 모든 표본 외 예측구간 시점과 예측시계에 대해서 $BVS(3)$ 기법에 기반하여 예측변수를 선택한다. 단, 특정 예측시점에 대해 변수선택을 한 번 실시하는 $BVS(2)$ 알고리즘과는 다르게 매 시점마다 변수선택을 여러 번 반복 시행하고, 선택된 예측변수들 중에서도 상대적으로 중요도가 낮은 예측변수들을 제외하여 예측력 향상을 도모한다. 예측변수들의 상대적 중요도를 측정하기 위해 사

후포함확률을 계산하고, 각 예측변수들은 사후포함확률이 임의의 기준을 만족시키지 못하면 예측에서 제외된다. 예를 들어, $BVS(3)$ 기법의 첫 번째 반복에서 0.1를 기준으로 하는 경우 변수선택이 완료된 후 사후포함확률이 0.1보다 작은 예측변수들은 모형에서 제외된다. 기준 값은 매 반복 시행마다 0.1 씩 증가하여 마지막 $BVS(3)$ 기법 시행에서는 0.7까지 증가한다. 요약하면 매 예측시점 마다 $BVS(3)$ 기법을 총 7번 반복하여 상대적으로 중요하지 않은 예측변수들을 걸러내어 예측력을 향상시키는 것이다. 기준값의 설정은 일종의 최적화 과정으로 다양한 값을 설정해보고 분포예측 알고리즘을 적용하여 예측력이 가장 좋아지는 수준에서 결정한다. 이와 같은 방식으로 변수선택을 반복하여 최적화 과정을 거친 $BVS(3)$ 기법을 $IBVS(3)$ (Iterated $BVS(3)$) 기법 또는 모형이라고 하겠다. 본 연구에서는 예측시계 별로 기준값이 달라지도록 하였으며, 기준설정에 대한 구체적인 내용은 Ⅳ장에서 설명한다.

예측 단계에서는 변수선택 단계에서 최종적으로 선택된 예측변수들을 대상으로 $BVS(3)$ 기법을 다시 적용하여 전년동월대비 서울 아파트매매 가격지수 분포예측을 실시한다. $BVS(3)$ 기법을 다시 적용하는 이유는 이전 단계에서 선택된 예측변수들의 중요성이 높다고 하더라도 예측변수 불확실성이 존재하기 때문에 이를 반영하기 위해서이다.

베이지안 변수선택 기법을 이용한 분포예측 알고리즘은 다음과 같다.

〔변수선택〕

1 단계: $MCMC$ 시뮬레이션 크기 N, training 표본크기 T_0, 표본 외 예측구간 크기 OSS, 예측시계(h), 예측변수의 최대시차(P)의 크기를 설정한다.

2 단계: $g=1,2,...,OSS$와 $h=1,2,...,H$에 대해서 T_{0+g} 시점의 y_{t-1+h} 분포예측을 위한 $BVS(3)$ 기법을 적용한다. 알고리즘이 예측변수들의 사후포함확률을 계산하여 저장한다.

3 단계: 사후포함확률이 기준값보다 작은 예측변수들은 모형에서 제외시킨다. 기준값을 증가시키고 2 단계로 돌아간다. 만약 기준값이 최대 기준값 보다 커지면 반복을 멈추고 4 단계로 넘어간다.

〔분포예측〕

4 단계: $g=1,2,...,OSS$와 $h=1,2,...,H$에 대해서 변수선택 단계에서 선택된 예측변수들을 이용하여 $BVS(3)$ 기법을 적용해 T_{0+g} 시점의 y_{t-1+h} 분포예측을 실시한다. $BVS(3)$ 기법의 n번째 반복시행에서 모형 파라메터 집합 $\theta^{(n)}$와 $\gamma^{(n)}$이 주어져 있을 때, y_{t-1+h}의 조건부 예측분포는 다음과 같다.

$$y_{t-1+h}^{(n)} | F_{t-1}, \theta^{(n)}, \gamma^{(n)} \sim N\left(\phi^{(n)} + \sum_{j=1}^{p} \alpha_j^{(n)} y_{t-j} + x' \beta_\gamma^{(n)}, \sigma^{2(n)}\right)$$

단, 변수선택 단계에서 최종적으로 선택된 예측변수의 수가 K_{vs}라고 할 때, x는 차원이 $K_{vs} \times 1$인 예측변수들의 벡터이고, $\beta_\gamma^{(n)}$은 주어진 $\gamma^{(n)}$으로부터 선택되지 않은 예측변수들의 계수가 0이 되도록 한 $K_{vs} \times 1$ 벡터이다. 예를 들어, $K_{vs}=3$, $\gamma^{(n)}=(0,1,1)'$, $\beta^{(n)}=(1,2,3)'$이라면 $\beta_\gamma^{(n)}=(0,2,3)'$이 된다.

5 단계: 분포예측이 완료되면, 모든 g와 h에 대해서 $PPC(h,g)$를 계산해서 저장한다.

5. 벤치마크 모형

본 논문에서 제시한 $IBVS(3)$ 기법의 예측력을 비교 평가하기 위해서 다음의 벤치마크 모형들을 설정하였다. 첫 번째 모형군은 전통적으로 예측에 많이 사용되는 1차 자기회귀 모형(이하, $AR(1)$)과 p차 자기회귀 모형(이하, $AR(p)$)이다. $AR(p)$ 모형은 다음과 같이 표현된다.

$$y_{t-1+h} | F_{t-1}, \theta \sim N\left(\mu + \sum_{i=1}^{p} \alpha_i y_{t-i}, \sigma^2\right)$$

$AR(1)$ 모형은 다른 예측변수들을 전혀 고려하지 않기 때문에 변수선택에 관계없이 일반적인 베이지안 분포예측 기법으로 예측을 실시한다. $AR(p)$ 모형의 경우

시차 p는 예측력을 기준으로 동태적으로 결정된다. 구체적으로 설명하면 허용 가능한 최대시차 P를 설정하고, 각각의 예측시계와 표본 외 예측구간 시점에 대해서 최대시차 내에서 고려 가능한 모든 $AR(p)$ 모형들을 추정하고 예측을 실시한 다음 $RMSE$를 기준으로 최적시차를 결정한다. 예를 들어, 최대시차 $P=2$이고, 예측시계 $h=1$이라고 할 때, 첫 번째 표본 외 예측구간 시점 T_0+1에 대해서 $AR(1)$ 모형과 $AR(2)$ 모형으로 예측을 실시한 후, 더 나은 $RMSE$ 값을 나타내는 모형의 시차를 T_0+1 시점의 최적시차로 결정하는 방법이다.

두 번째는 BVS 모형군으로 반복적으로 변수선택을 하지 않는 $BVS(3)$ 모형과 동태적으로 변수선택을 하지 않고, 최초의 표본 외 예측시점에 대해서만 $IBVS(3)$ 기법을 적용한 후, 이후의 예측시점에 대해서는 앞에서 선택된 예측변수들을 이용하여 분포예측을 실시하는 모형(Fixed $IBVS(3)$, 이하 $FIBVS(3)$)이다. 예를 들어, 특정 예측시계에서 T_0+1 시점에 선택된 예측변수의 수가 10개라면 나머지 모든 표본 외 예측구간 시점에 대해서는 앞서 선택된 10개의 변수들만 이용해서 예측을 실시한다.

추가로 George and McCulloch(1993)의 $BVS(2)$ 모형과 $BVS(2)$에 본 연구의 분포예측 알고리즘을 적용한 $IBVS(2)$ 모형을 벤치마크 모형으로 설정하였다. $BVS(2)$는 $BVS(3)$와 비교했을 때 동태적으로 변수를 선택하는 방식은 동일하지만 지시변수 γ_k가 취할 수 있는 값에 대한 가정이 다르다.[8]

예측력 평가를 위한 모형집합은 다음과 같다.

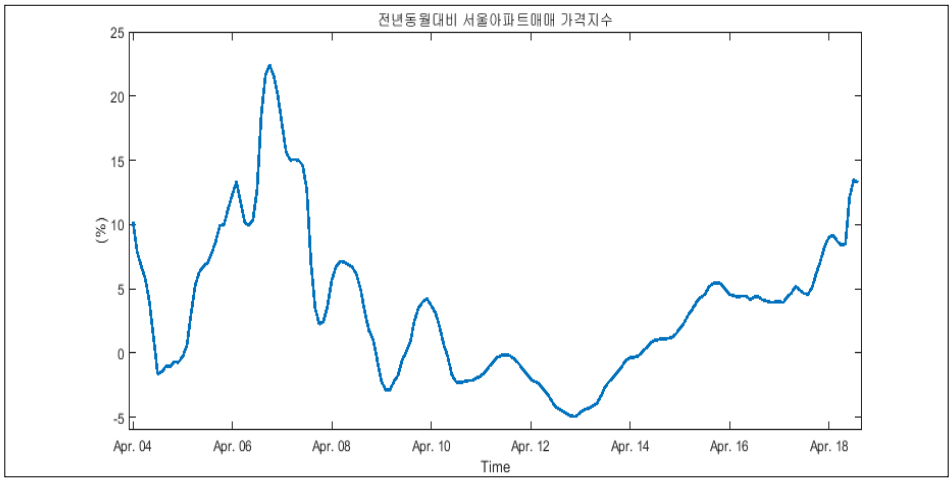
$$\{AR(1), AR(p), BVS(3), IBVS(3), FIBVS(3), BVS(2), IBVS(2)\}$$

위와 같은 모형들로 벤치마크 모형을 설정한 이유는 예측시계별로 $IBVS(3)$의 예측력을 (1) 추가적인 예측변수의 효과, (2) 모형의 시변성 혹은 동태적 변수선택의 효과, (3) 지시변수 방향성의 효과, (4) 반복적 변수선택의 효과와 같이 네 가지 측면에서 비교평가하고 측정하기 위한 것이다. 추가적인 예측변수의 효과는

8) $BVS(2)$의 경우 B_0와 B_1을 조정 파라메터로 연구자가 직접 설정해야 하지만, 무정보 사전분포(Non-informative prior) 하에서는 좋은 예측력을 기대하기 어렵다고 판단하여 $BVS(3)$와 동일하게 계층모형으로 가정하였다.

AR 모형과 나머지 모형들의 비교, 동태적 변수선택의 효과는 $IBVS(3)$와 $FIBVS(3)$ 또는 $AR(1)$과 $AR(p)$의 비교, 개선된 베이지안 샘플링의 효과는 $BVS(3)$와 $BVS(2)$ 또는 $IBVS(3)$와 $IBVS(2)$의 비교, 마지막으로 반복적인 변수선택의 효과는 $BVS(3)$와 $IBVS(3)$ 또는 $BVS(2)$와 $IBVS(2)$의 비교로부터 측정이 가능하다.

〈Figure 1〉 Seoul apartment housing purchase price index growth rate(%, YoY)



6. 자료 및 예측변수

〈Figure 1〉은 종속변수인 전년동월대비 서울 아파트매매가격지수 변동률의 시계열을 그림으로 나타낸 것이다. 그림에서 나타나듯이 종속변수의 지속성이 상당히 큰 것으로 보여 종속변수의 1기 시차변수를 예측변수로 추가하였다. 자료는 월별자료이며 총 기간은 2003년 4월부터 2018년 11월까지이다.

〈Table 1〉은 전년동월대비 서울 아파트매매 가격지수의 분포예측을 위해 사용되는 예측변수들의 목록이다. 서론에서 언급하였듯이 주택가격은 주택시장 변수 뿐만 아니라 다양한 거시경제변수와 금융변수의 영향을 받아 변동할 가능성이 크다. 본 연구에서는 종속변수를 포함한 총 21개의 예측변수를 사용하며, 각각의 예측변수들은 상관관계를 기준으로 주택가격 및 미분양주택, 건설 및 인구이동률, 물가, 금리 및 환율, 경기변동의 5개의 그룹으로 나누어진다.

〈Table 1〉 Predictor list

Category	SN	Data	Source
House Price, Unsold New Housing	1	Apartment Housing Purchase Price Index (Seoul, Growth Rate(YoY))	KB Kookmin Bank
	2	Apartment Housing Purchase Price Index (Whole Country, Growth Rate(YoY))	
	3	Apartment Jeonse Price Index (Seoul, Growth Rate(YoY))	
	4	Housing Jeonse Price Composite Index (Seoul, Growth Rate(YoY))	
	5	Land Price Index (Seoul, Growth Rate(YoY))	Korea Appraisal Board
	6	Unsold New Housing (Seoul)	Ministry of Land, Infrastructure, and Transport
Construction, Migration	7	Statistics on Building Commencement Works (Total, Seoul)	
	8	Statistics on Building Commencement Works (Residential, Seoul)	
	9	Population Migration Rate (Seoul)	Statistics Korea
	10	Population Migration Rate (Whole Country)	
Inflation	11	Consumer Price Index (Seoul, Growth Rate(YoY))	
	12	M2 (Growth Rate(YoY))	The Bank of Korea
Interest Rate, Exchange Rate	13	Yields of Treasury Bonds (3-year)	
	14	Loans & Discounts By Fund (For Housing, Growth Rate(YoY))	
	15	Mortgage Rate (Loans to Households(Houses))	
	16	Exchange Rate (Won/US dollar, Growth Rate(YoY))	
Business Fluctuations	17	Leading Composite Index (Whole Country, Growth Rate(YoY))	Statistics Korea
	18	Coincident Composite Index (Whole Country, Growth Rate(YoY))	
	19	Index of All Industry Production (Construction, Growth Rate(YoY))	
	20	Business Survey Index (Business Condition, Construction)	The Bank of Korea
	21	KOSPI Index (Growth Rate(YoY))	Korea Exchange

7. 사전분포

예측모형 파라메터들의 사전분포는 앞 장에서 설명한 선형회귀모형의 사전분포와 동일한 분포를 가정한다. $\beta_{k,j}$는 γ가 주어져 있을 때, 사전 평균이 $\mu_{k,j}$이고, 사전 분산이 $b_{k,j}$인 혼합 정규분포를 따른다고 가정한다. $\mu_{k,j}$와 $b_{k,j}$는 식 2.2과 식 2.1에 따라 계산된다. 시뮬레이션 최초 β를 샘플링 하는 시점에 B_0, B_1, B_{-1}은 초기값으로 주어지며, 각각 사전분포의 평균값인 0.3, 0.01, 0.01로 설정하였다. 변수선택과 관계없이 항상 모형에 포함되는 μ에 대해서는 $N(0,5)$을 가정한다. σ^2, B_0, B_1, B_{-1}의 사전분포는 역감마분포이며, σ^2에 대해선 무정보 사전분포를 가정하고, B_0, B_1, B_{-1}에 대해선 사전평균이 초기값과 같아지도록 설정한다. $\overline{\beta_1}$와 $\overline{\beta_{-1}}$의 사전평균은 각각 0.5와 -0.5, 사전분산은 동일하게 0.01로 가정한다. George and McCulloch(1993)의 표준적인 BVS 기법에서는 $B_0 < B_1$을 가정하지만 본 연구에서는 β의 부호가 결정된 이후에 다른 부호의 값이 샘플링되지 않도록 B_1과 B_{-1}의 초기값을 B_0보다 상대적으로 더 작게 설정하였다. γ에 대해서는 무정보 사전분포로 $\Pr[\gamma_{k,j}=1]=\Pr[\gamma_{k,j}=0]=\Pr[\gamma_{k,j}=-1]=1/3$로 설정하였다.

Ⅳ. 예측 결과

표본 외 예측구간은 3년($OSS=36$)으로 2015년 12월부터 2018년 11월까지를 예측시계 별로 예측한다. 예측시계는 단기와 중기 및 장기의 예측력 차이를 비교하기 위해서 $h \in \{1,6,12\}$로 설정하였다.

분포예측 알고리즘의 변수선택 단계에서 $IBVS(3)$, $IBVS(2)$, $FIBVS(3)$ 모형에 대한 기준값은 $h=1$인 경우에 0.1에서 시작하여 0.1씩 증가하며 최대 기준값은 0.9이다. $h=6$인 경우에는 0.1에서 시작하고 최대 기준값은 0.5로 설정하였다. $h=12$일 때는 $h=6$일 때와 동일한 기준값을 적용하였다. 각 예측시계별 기준값은 0과 1사이의 다양한 값들 중에서 가장 좋은 예측력을 나타내는 값으로 설정한 것이다. $BVS(3)$와 $BVS(2)$에 대해서는 예측시계 별 마지막 기준값인 {0.9, 0.5, 0.5}를 이용해서 변수선택을 한 번만 수행한다.

예측변수 최대시차 $Q=24$로 설정하였다. 이는 주택시장에서 시장가격과 거래량의 조정이 서서히 이루어지는 특성 때문이다. 특히, 아파트나 고층 오피스텔 등은 공급까지 많은 시간이 소요되고 정부 정책에 의한 제약을 많이 받기 때문에 주택의 공급은 일반적으로 비탄력적이다. 반면 주택의 수요는 탄력적인 편이기 때문에 수요·공급의 시차 불일치 혹은 불균형이 발생할 수 있다. 그리고 가격하방성이 존재하여 가격이 하락하면 주택소유자는 거래를 기피하고 시장에서는 오히려 거래량이 줄어드는 현상이 나타날 수 있다. 또한 한국 주택시장의 경우 임대차 거래에서 전세계약이 차지하는 비중이 높고, 임대차 계약은 일반적으로 2년 이상 지속된다. 이와 같은 주택시장의 특성으로 인하여 주택가격은 주택가격의 결정요인이나 외부충격으로 인한 영향을 장기간 받게 된다. 이에 본 연구에서는 예측변수가 주택가격에 긴 시차를 두고 영향을 끼칠 수 있다고 판단하고, 예측변수들 중에서 '건축물 착공실적'을 바탕으로 예측변수 최대시차를 설정하였다. 아파트 착공 후 공사기간은 건설규모 등에 따라 다르겠지만 일반적으로 2 또는 3년으로 알려져 있고, 이 점을 반영하여 예측변수 최대시차를 24로 설정하였다. 종속변수를 제외한 총 20개의 예측변수와 상수항을 포함한 총 예측변수의 수는 481개이다.

본 연구에서와 같이 그룹별 예측변수들간에 상관관계가 높고 시차가 큰 예측변수들을 사용하는 경우에 다중공선성 문제가 발생할 가능성이 크다. 따라서, 본 연구에서는 변수선택에 앞서 다중공선성 문제를 완화하고 예측력을 향상시키기 위해 상수항과 종속변수 시차변수를 제외한 20개의 예측변수들에 대해 주성분분석(Principal Component Analysis)을 적용하였다. 특히, 효율적인 문제완화를 위해 20개의 예측변수들에 대해서 일괄적으로 주성분분석을 적용하지 않고, 〈Table 1〉의 상관관계를 기준으로 구분한 5개의 예측변수 그룹에 대해서 그룹별 주성분분석을 적용하였다.

주성분분석은 $AR(1)$과 $AR(p)$ 모형을 제외한 $BVS(3)$, $IBVS(3)$, $FIBVS(3)$, $IBVS(2)$, $BVS(2)$ 모형에 동일하게 적용된다. 논문에 보고하지 않았지만 주성분분석을 적용하는 경우, 적용하지 않는 경우에 비해 예측력이 크게 향상되는 것으로 나타났다.

추가로 $BVS(3)$, $IBVS(3)$, $FIBVS(3)$, $BVS(2)$, $IBVS(2)$ 모형에는 종속변수의 시차변수도 예측변수로 포함되는데, 이때 시차는 $AR(p)$ 모형의 예측에서 결정된 최적시차 p로 설정하였다. $AR(p)$와 BVS 모형들의 종속변수 최적시차가 모두

동일할 것이라고 기대하긴 어려우나 최대시차 P가 큰 경우, 모형 추정에서 과도한 계산비용이 발생하기 때문에 분석의 편의상 이와 같은 방법을 사용하였다.

다음 장부터는 모형들의 예측결과에 대해 (1) 예측변수의 효과, (2) 동태적 변수선택의 효과, (3) 지시변수 방향성의 효과, (4) 반복적 변수선택의 효과의 네 가지 측면에서 비교 모형들의 예측결과를 분석한다.

1. 예측변수 효과

종속변수의 시차변수를 제외한 예측변수의 추가적 도입이 국내 주태가격 예측력 향상에 유효한 효과가 있는지 알아본다. 〈Table 2〉는 비교모형들의 PPC와 $RMSE$ 값들을 나타낸 것으로 모든 표본 외 예측구간에 대한 예측시계별 평균이다. AR 모형군과 BVS 모형군을 비교했을 때, 단기 예측($h=1$)을 제외한 중, 장기 예측에서 BVS 모형군의 예측력이 AR 모형군 보다 우월한 것으로 나타났다. 이러한 결과는 전년동월대비 서울 아파트매매 가격지수변동률과 서울 아파트매매 가격지수 수준의 예측분포를 그린 〈Figure 3, 4, 5, 6, 7, 8〉에서도 확인할 수 있다. 특히 〈Figure 4, 5, 7, 8〉에 나타나듯 중기와 장기에서 $AR(p)$ 모형은 예측력이 급격하게 악화되는 반면, BVS 모형군은 상대적으로 좋은 예측력을 유지하고 있다. 따라서, 위험관리 측면에서 추가적인 예측변수를 고려하는 것이 의사결정에 더 유용한 정보를 제공할 것이다.

반면 단기 예측에서는 〈Table 2〉에 나타나듯이 PPC와 $RMSE$ 모든 기준에서 $AR(p)$ 모형이 BVS 모형군 보다 우월한 예측력을 보여 단기 예측에서 추가적인 예측변수는 불필요한 것으로 판단된다. 이러한 특징은 〈Table 2〉의 BVS 모형군들의 단기 예측 결과에서도 나타난다. $RMSE$ 결과를 살펴보면 $BVS(3)$와 $BVS(2)$ 모형이 각각 $IBVS(3)$와 $IBVS(2)$ 모형보다 단기 예측력에서 우월한데, $IBVS(3)$와 $IBVS(2)$ 모형의 경우 평균적으로 선택된 예측변수의 수는 각각 14.49개와 15.08개로, $BVS(3)$와 $BVS(2)$가 각각 5개와 5.8개 인 것에 비해 많아 선택되는 변수의 수가 작을수록 예측력이 좋아지는 경향이 있으며, 추가적인 변수가 오히려 예측력을 악화시키는 것으로 판단된다. 이러한 단기 예측결과는 〈Figure 1〉에 나타나듯이 종속변수의 지속성이 강하기 때문인 것으로 보인다.

2. 동태적 변수선택의 효과

본 연구의 주 목적 중 하나는 국내 주택가격 예측에서 동태적 변수선택 또는 모형 시변성의 중요성을 파악하는 것이다. 이는 $IBVS(3)$와 $FIBVS(3)$의 비교 평가로부터 알 수 있다. 〈Table 2〉에서 확인되듯이 PPC 기준 중기 예측을 제외하고는 모든 예측력 기준과 예측시계에 대해 $IBVS(3)$ 모형이 $FIBVS(3)$ 모형보다 예측력이 우월하다. 두 모형의 예측력 차이는 분포예측 그림을 통해 쉽게 이해할 수 있는데, 〈Figure 3, 4, 5, 6, 7, 8〉의 (b)와 (c)를 보면 단기에서 장기로 갈수록 그 차이가 커짐을 확인할 수 있다. $FIBVS(3)$의 경우 장기로 갈수록 예측치가 실현치에서 크게 벗어나며 예측분포의 분산도 커지는 경향을 보인다. 따라서 국내 주택가격의 정확한 예측을 위해서는 추가적인 예측변수 뿐만아니라 동태적 변수선택을 통해서 모형의 시변성을 고려할 필요가 있다.

〈Table 2〉 Out-of-sample forecasting performance of the models

	PPC						
h	$AR(1)$	$AR(p)$	$IBVS(3)$	$IBVS(2)$	$BVS(3)$	$BVS(2)$	$FIBVS(3)$
1	1.79	**0.86**	0.90	0.97	1.29	1.27	1.63
6	22.05	19.92	1.80	**1.43**	1.98	1.84	1.71
12	37.70	35.77	**1.11**	1.45	1.31	1.48	4.64
	20.51	18.85	**1.27**	1.28	1.53	1.53	2.66
	RMSE						
h	$AR(1)$	$AR(p)$	$IBVS(3)$	$IBVS(2)$	$BVS(3)$	$BVS(2)$	$FIBVS(3)$
1	0.87	**0.66**	0.86	0.90	0.84	0.86	1.16
6	3.10	2.91	**0.91**	1.07	0.98	1.18	1.10
12	4.43	4.42	**0.92**	1.12	0.97	1.08	1.99
	2.80	2.67	**0.90**	1.03	0.93	1.04	1.42

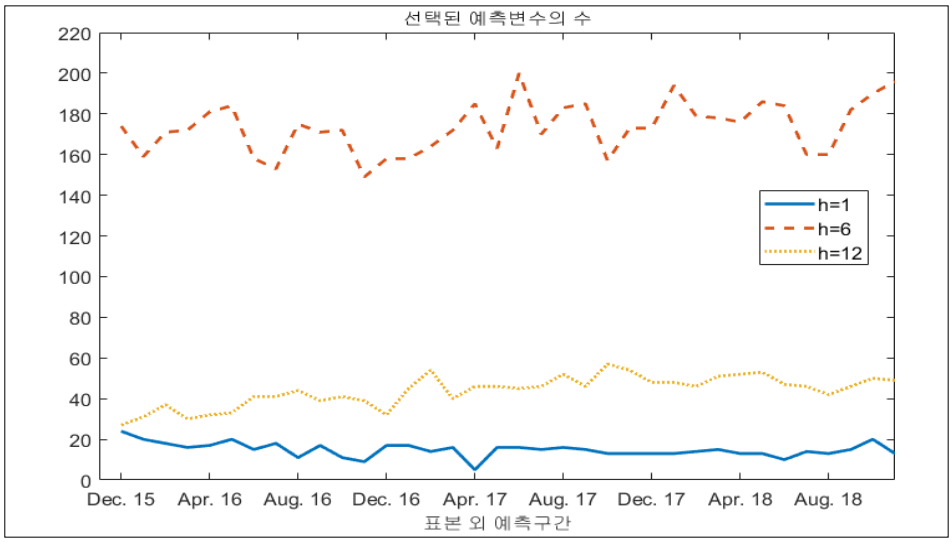
Note: This table report *PPC* and *RMSE* values averaged over the out-of-sample period and all forecasting horizons. The out-of-sample period is from 2015:M12 to 2018:M11. Bold numbers indicate the smallest *PPC* and *RMSE* values in each row of the table.

〈Figure 4〉를 보면 중기 예측에서 $FIBVS(3)$ 모형은 예측분포 분산이 $IBVS(3)$ 모형보다 더 작긴 하지만, 예측분포 신뢰구간이 실현치를 포함하지 못하는 구간이 많이 나타나 위험관리 측면에서는 $IBVS(3)$ 모형이 더 유용하다고 판단된다. 또한

$AR(p)$와 $AR(1)$의 비교를 통해서도 동태적 변수선택의 효과를 확인할 수 있다. 이는 $AR(p)$ 모형의 경우 예측력을 기준으로 자기회귀의 최적시차가 동태적으로 결정될 수 있도록 하였고, $AR(1)$의 경우 그렇지 않기 때문이다. 〈Table 2〉에 나타나듯이 큰 폭은 아니지만 $AR(p)$ 모형이 $AR(1)$ 모형에 비해 예측력이 향상 된 것을 확인할 수 있다.

〈Figure 2〉는 표본 외 예측구간에 대해서 본 연구의 $IBVS(3)$ 기법을 적용했을 때 선택된 예측변수 수의 시계열을 그린 것이다. 그림에서 나타나듯이 모든 예측시계에 대해서 선택된 예측변수의 수는 일정하지 않으며 예측 시점에 따라 변동한다. 상대적으로 장기 예측에서 적은 수의 예측변수가 선택되며 중기 예측에서 가장 많은 예측변수가 선택되는 것으로 나타났다.

〈Figure 2〉 The number of selected predictors in $IBVS(3)$ model



3. 지시변수 방향성에 의한 효과

본 연구에서는 George and McCulloch (1993)의 방법과 다르게 지시변수 γ에 대한 가정을 수정하여 방향성을 부여함으로써 예측변수의 선택 여부 뿐만아니라 선택된 예측변수가 종속변수에 어떤 방향으로 영향을 주는지를 분석하였다. 지시변수의 역할을 개선한 베이지안 샘플링이 국내 주택가격 예측력에 미치는 효과는

IBVS(3)와 *IBVS*(2) 또는 *BVS*(3)와 *BVS*(2)의 예측력 비교로부터 확인할 수 있다. *IBVS*(3)와 *IBVS*(2) 비교 시, 〈Table 2〉에 따르면 중기의 *PPC* 값을 제외한 모든 예측력 평가 기준과 예측시계에 대해 *IBVS*(3)의 예측력이 *IBVS*(2) 모형보다 뛰어나며, *BVS*(3)와 *BVS*(2) 모형 비교에서도 같은 결과를 보여준다. 이러한 효과는 방향성이 부여된 지시변수가 예측에 중요한 변수를 더 잘 식별하여 추정과 예측에서 효율성을 증대시켰기 때문인 것으로 판단된다.

〈Table 2〉와 〈Figure 4〉의 (c), (d)를 보면 중기 예측에서 *IBVS*(3)의 *PPC* 값이 *IBVS*(2) 보다 큰 원인은 예측오차가 아닌 상대적으로 큰 예측분포 분산에 기인한 것인데, 이는 *IBVS*(3) 모형에서 선택되는 예측변수의 수가 *IBVS*(2) 모형보다 크기 때문인 것으로 판단된다. 중기 예측에서 동일한 기준값 0.5를 부여했음에도 불구하고 *IBVS*(3) 모형에서는 〈Figure 2〉에 나타나듯 평균적으로 173.47개가 선택된 반면, *IBVS*(2) 모형에서는 41.86개로 큰 차이를 보였다.

4. 반복적 변수선택의 효과

본 연구가 제시한 분포예측 알고리즘은 예측시점마다 특정 사후포함확률을 기준으로 한번 만 변수선택하는 George and McCulloch (1993)의 방법과 다르게 예측력 향상에 도움이 되는 중요 변수가 제외되지 않도록 낮은 기준값에서 시작해 조금씩 증가시켜 일종의 최적화 또는 스무딩 과정을 거친다는 것이다. 이러한 분포예측 알고리즘의 효과를 측정하기 위해서 *IBVS*(3)와 *IBVS*(2) 모형에 대해 예측시점마다 단기에서 0.1씩 증가시켜 0.9까지 9번 반복하였고 중기와 장기에 대해서는 0.5까지 5번 반복 시행하였다. *BVS*(3)와 *BVS*(2) 모형에 대해서는 분포예측 알고리즘의 마지막 기준값만을 이용해서 한 번만 변수선택하였다.

〈Table 2〉와 예측분포 그림을 보면 *IBVS*(3)와 *BVS*(3), *IBVS*(2)와 *BVS*(2) 사이에 예측력 차이가 극명하진 않으나 한 번만 변수선택할 때 보다 개선된 효과가 있다. 특히, *IBVS*(3) 모형은 단기의 *RMSE* 값을 제외하면 *BVS*(3) 모형보다 모든 면에서 예측력이 좋다. 참고로 논문에 보고하지 않았지만, 중기와 장기에 대해 사후포함확률 기준이 강할수록 *BVS*(3)와 *BVS*(2)의 예측력은 급격하게 나빠지는 반면, *IBVS*(3)와 *IBVS*(2)의 예측력은 완만하게 하락하는 것으로 나타났다. 이러한 결과는 예측변수들의 조합에 따라 특정 예측변수의 중요성 또는 사후포함확률

이 상대적으로 커질 수도 작아질 수도 있다는 것을 의미한다. 따라서, 국내 주택가격 예측에 있어 특정 기준값으로 한 번만 변수선택하는 경우, 실제로 예측력 향상에 도움이 되는 예측변수들이 제외될 수 있으므로, 점진적으로 변수선택을 반복하는 과정이 정확한 주택가격 예측에 중요한 역할을 할 것이다.

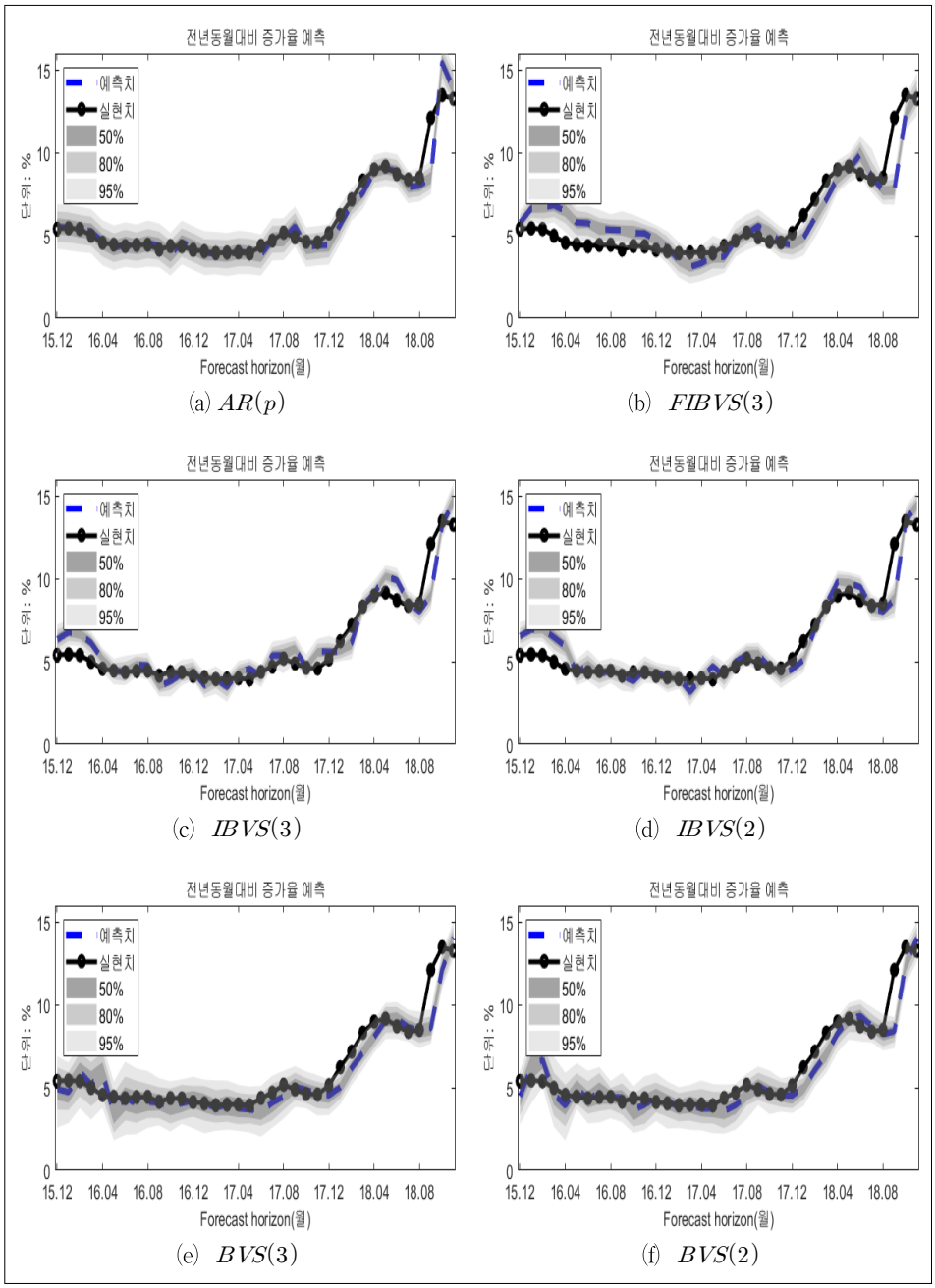
5. 모형별 예측력 종합평가

전년동월대비 서울 아파트매매 가격지수 변동률에 대한 2015년 12월부터 2018년 11월까지의 예측시계별 표본 외 예측결과, 단기 예측에서는 $AR(p)$모형의 예측력이 가장 우수한 것으로 나타났다. 이는 〈Figure 1〉에 나타나듯이 종속변수의 강한 지속성에 의한 것으로 보이며, 본 연구가 제시한 분포예측 알고리즘을 적용하는 경우 예측력이 다소 악화되었다. 추가적인 예측변수의 도입이 오히려 불필요한 정보를 포함하는 것이 되어, 동태적인 변수선택, 지시변수의 방향성, 반복적 변수선택의 효과가 미비한 것으로 판단된다.

중기 예측에서는 PPC 기준 $IBVS(2)$ 모형, $RMSE$ 기준 $IBVS(3)$ 모형의 예측력이 가장 우월한 것으로 나타났다. PPC와 $RMSE$를 종합적으로 고려했을 때는 〈Figure 4, 7〉의 (c)와 (d)에 나타나듯 $IBVS(2)$의 예측분포 신뢰구간이 실현치를 포함하지 못하는 경우가 $IBVS(3)$에 비해 많아 위험을 과소평가할 우려가 있어 위험관리 측면에서는 $IBVS(3)$ 모형이 낫다고 판단된다.

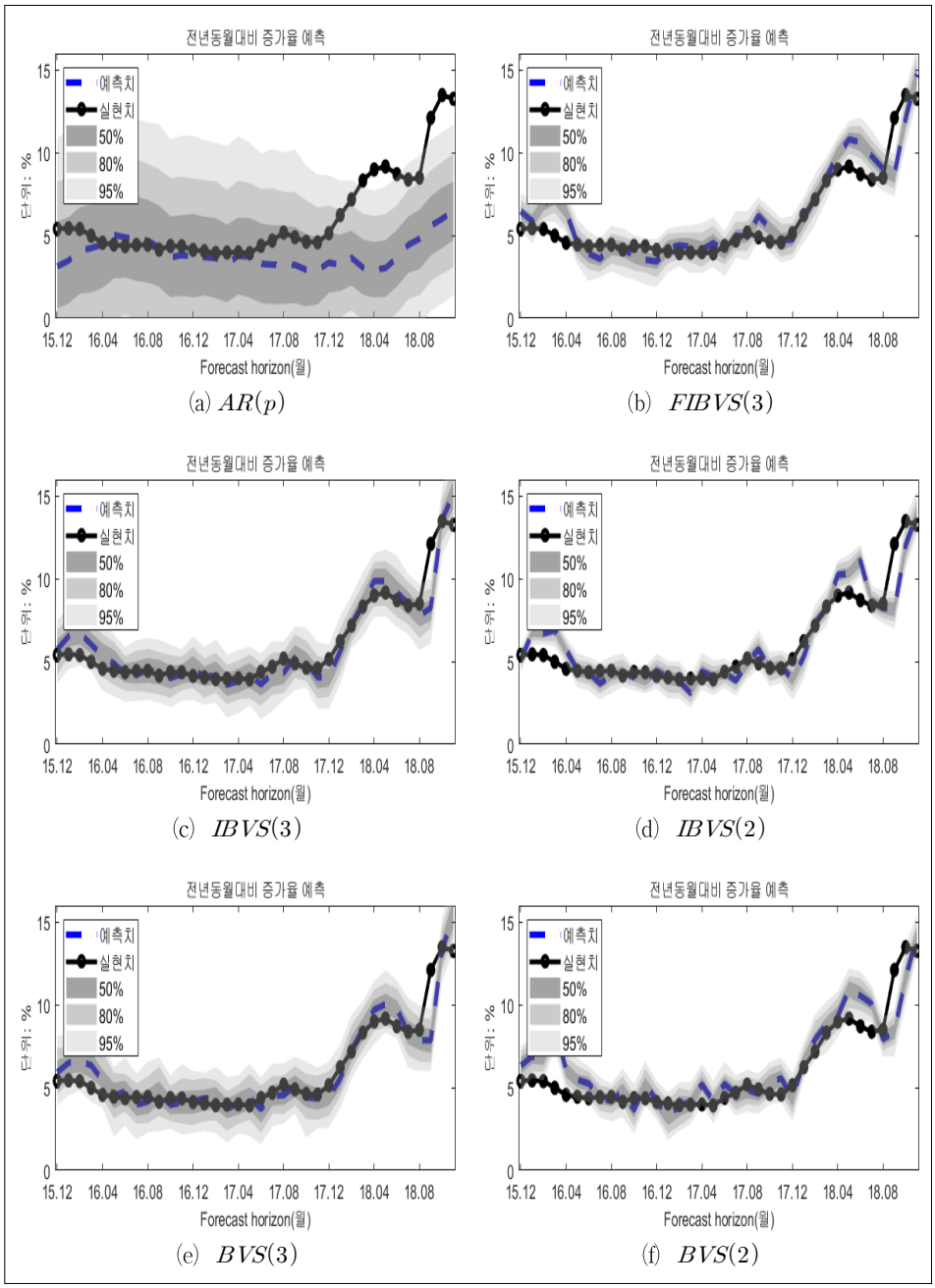
장기 예측의 경우 〈Figure 3, 4, 5〉에 나타나듯 $AR(p)$ 모형의 예측력은 장기로 갈수록 급격하게 악화되는 반면, $FIBVS(3)$ 모형을 제외한 BVS 모형군은 완만하게 하락하나 예측력을 유지하고 있다. 그 중에서도 〈Table 2〉와 같이 $IBVS(3)$ 모형의 예측력이 모든 예측력 평가 기준에서 가장 우월한 것으로 나타났다. 종합적으로 모든 예측시계에 대해서는 $IBVS(3)$ 모형은 PPC와 $RMSE$ 값이 각각 1.27과 0.90으로 가장 작게 나타나 가장 좋은 예측력을 보였다. 추가로 $IBVS(3)$, $BVS(3)$, $BVS(2)$의 PPC 값이 중기에 예측력이 나빠졌다가 장기에서 다시 향상되는 경향이 있는데, 이는 〈Figure 2〉에서 알 수 있듯이 선택된 변수의 수가 단기, 장기에 비해 중기 예측에서 상대적으로 많고, 변동성 또한 커서 예측분포의 분산을 크게 만들기 때문인 것으로 판단된다.

〈Figure 3〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index growth rate(%, YoY) by the models ($h=1$): 2015:M12-2018: M11



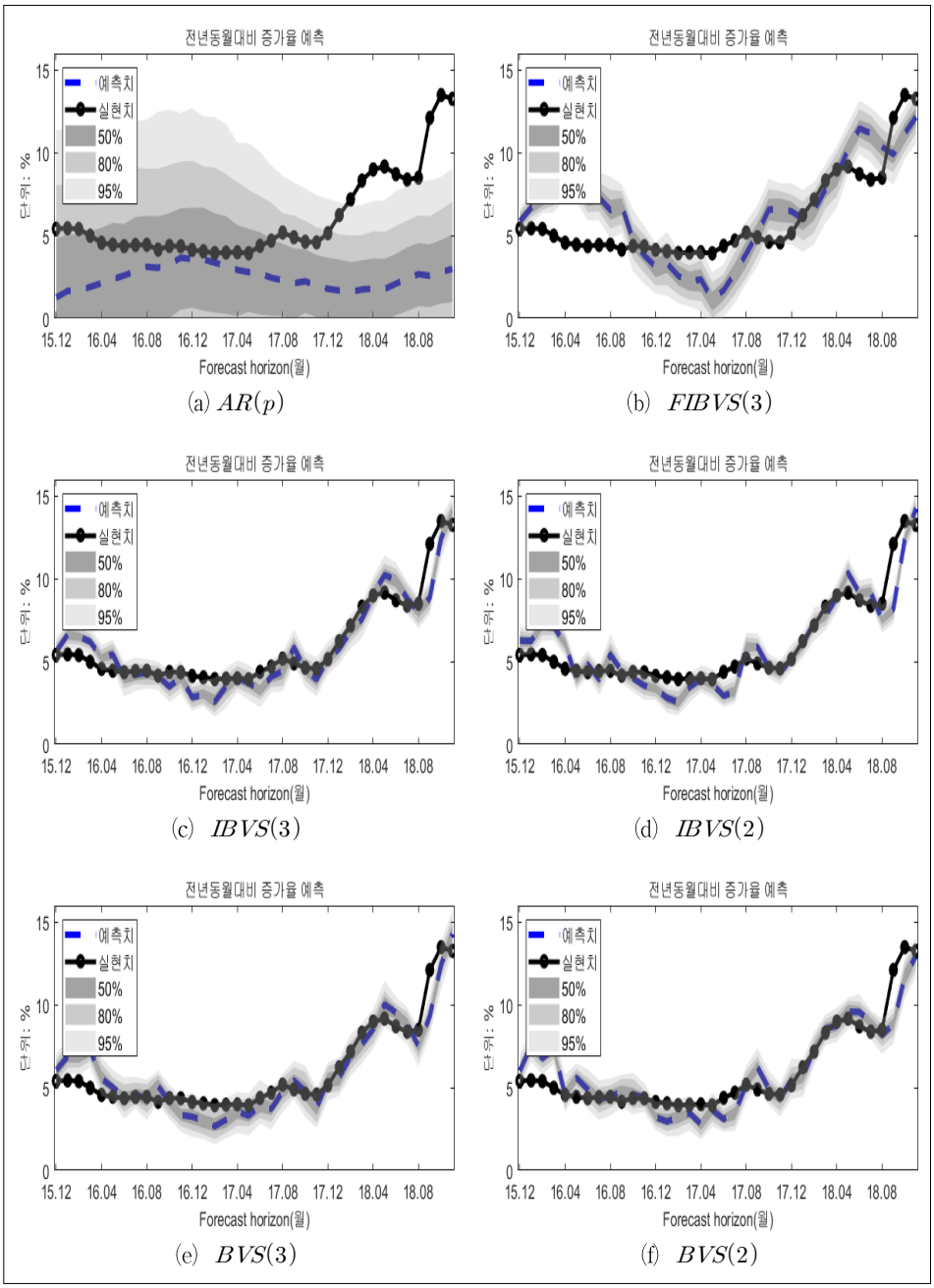
Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

〈Figure 4〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index growth rate(%, YoY) by the models ($h=6$): 2015:M12-2018: M11



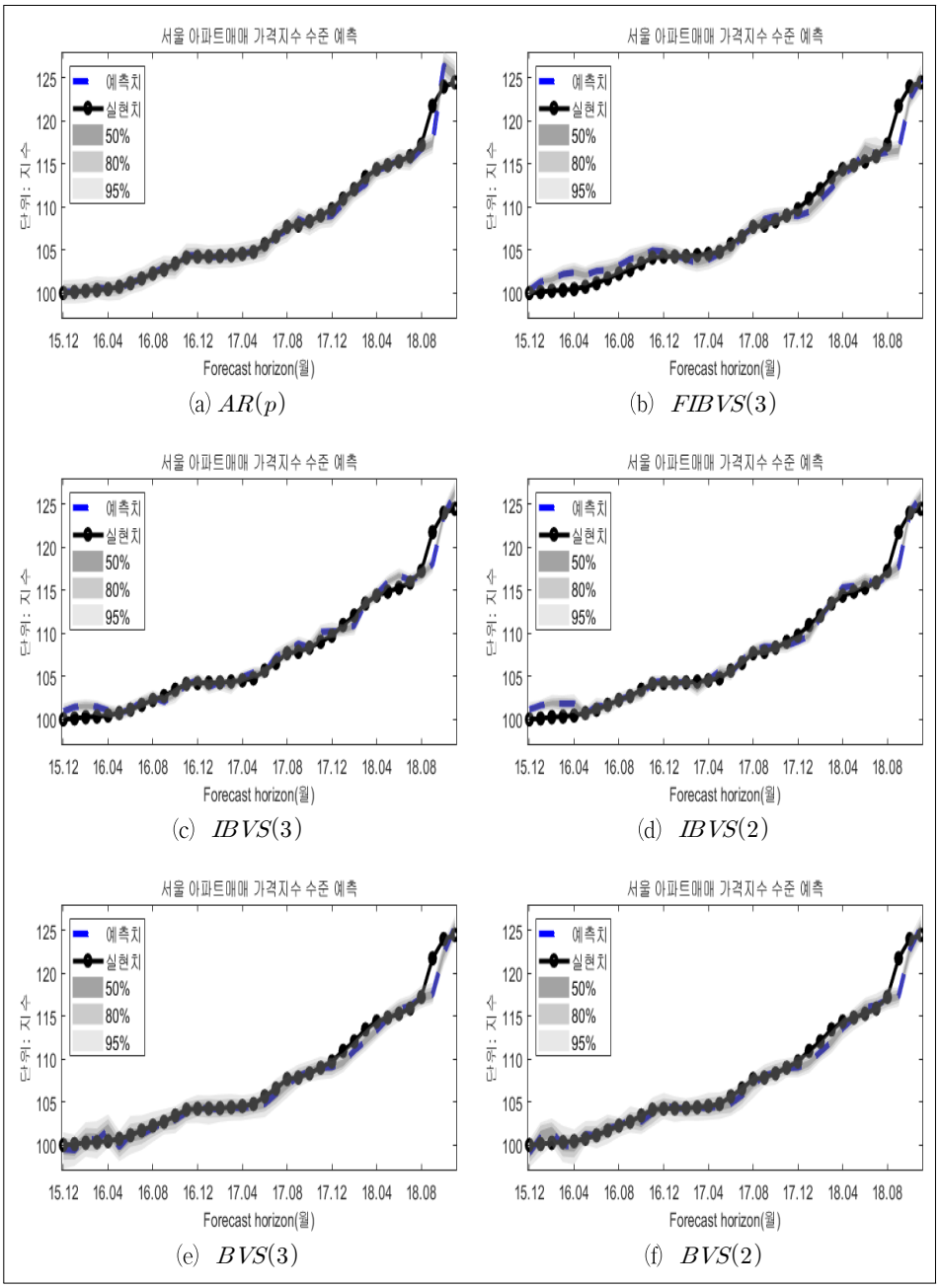
Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

〈Figure 5〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index growth rate(%, YoY) by the models ($h=12$): 2015:M12-2018: M11



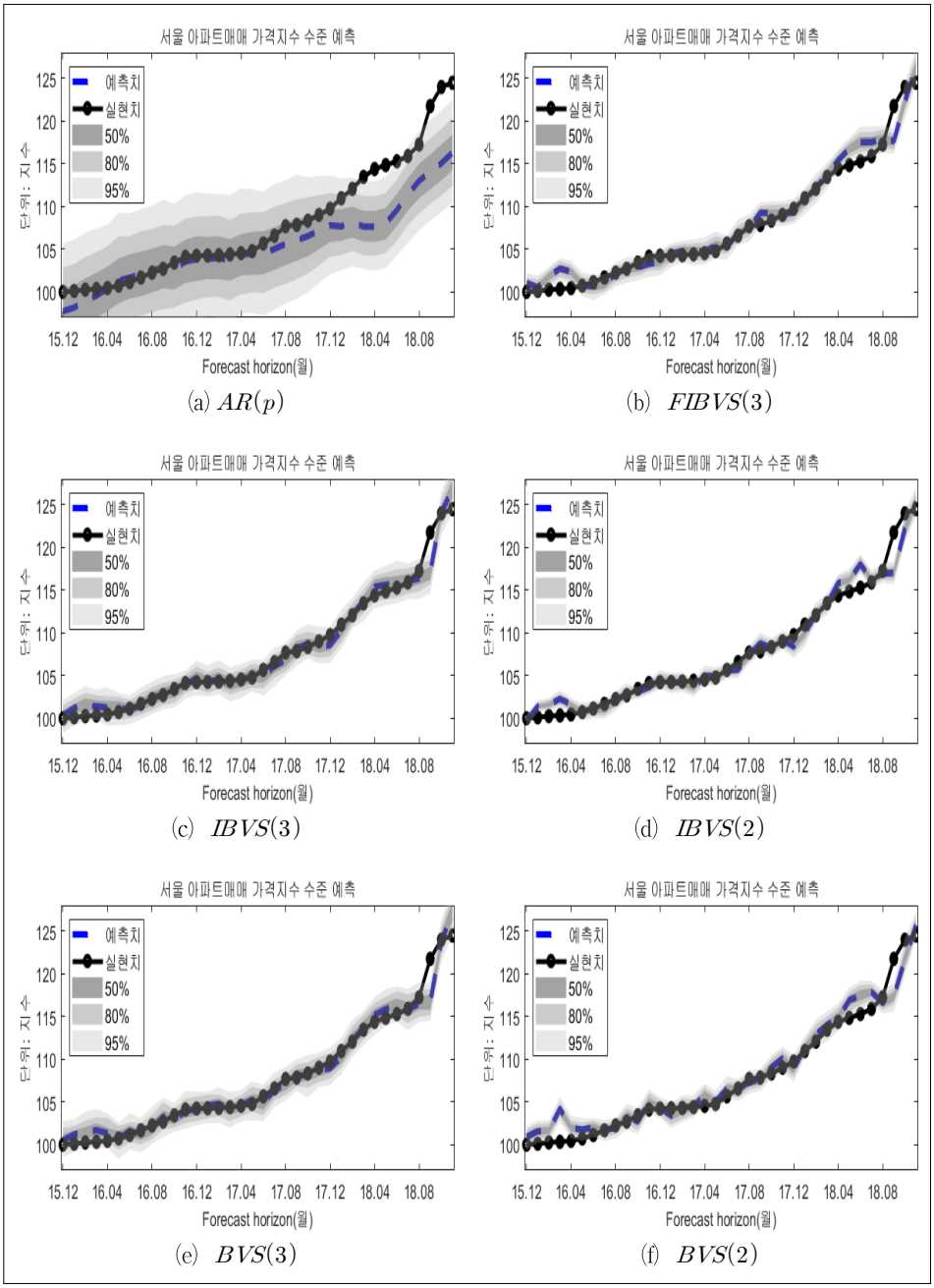
Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

〈Figure 6〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index by the models ($h=1$): 2015:M12-2018: M11



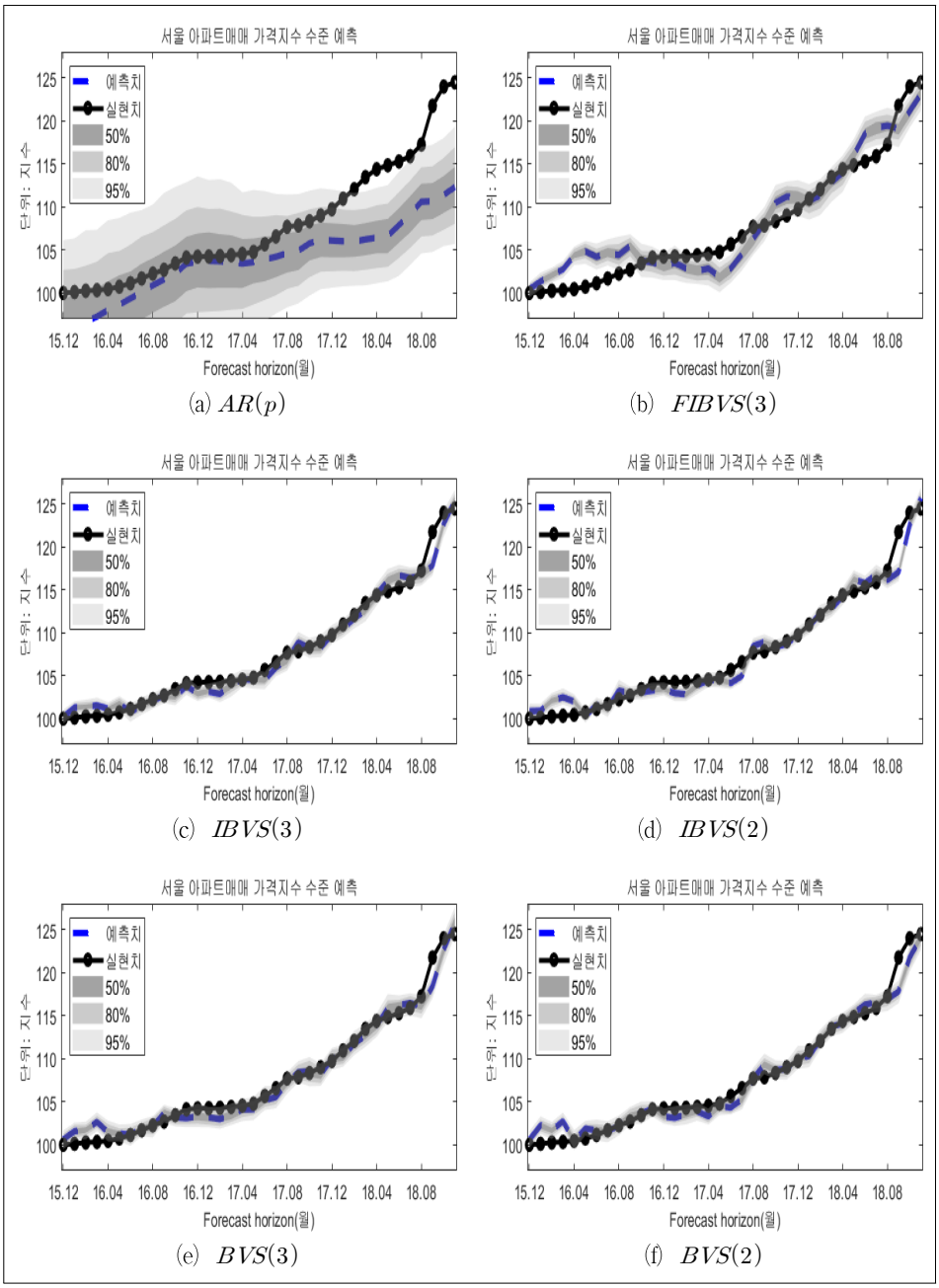
Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

〈Figure 7〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index by the models ($h=6$): 2015:M12-2018: M11



Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

〈Figure 8〉 Out-of-sample predictive distributions of Seoul apartment housing purchase price index by the models ($h=12$): 2015:M12-2018: M11



Note: Dotted line is the mean of the posterior predictive distribution. 50%, 80%, and 95% credit intervals are shaded.

V. 결 론

본 연구는 부동산 정책의 합리적인 의사결정에 필수적인 정확한 주택가격 예측을 위해 George and McCulloch (1993)의 변수선택 모형에 대한 가정과 분포예측 알고리즘을 수정하여 확장하고 표본 외 예측을 통해 벤치마크모형들의 예측력을 비교 평가하였다. 전년동월대비 서울 아파트매매 가격지수 변동률을 대상으로 하였으며, 모형의 조정 파라메터를 모형 파라메터로 가정하고, 변수선택에서 핵심이 되는 지시변수에 대한 가정을 개선해 방향성을 부여하여 변수선택 여부뿐만 아니라 어떤 방향으로 영향을 주는지 해석할 수 있게 하였다. 표본 외 예측결과, 본 연구의 $IBVS(3)$ 모형이 모든 예측시계에 대해 종합적으로 고려하였을 때, 예측력이 가장 우수한 것으로 나타났으며 특히, 모든 예측력 평가 기준에서 장기의 예측력이 가장 큰 폭으로 향상되는 것으로 나타났다.

국내 주택가격 예측에 대한 본 연구의 예측결과는 다음의 시사점을 갖는다. 첫째, 단기 예측에서 예측변수의 도입과 예측 알고리즘 적용은 오히려 분석의 효율성을 하락시켜 예측력을 악화 시킬 수 있다. 둘째, 예측시계가 길수록 예측변수들의 상대적 중요도가 커진다. 셋째, 예측변수들의 중요성은 시변에 따라 또는 예측변수 조합에 따라 달라질 수 있으므로 동태적으로 변수선택을 해야하며, 예측시점 마다 한 번만 실시하는 것보다 반복시행 하는 것이 식별에 효율적이다. 넷째, 베이지안 변수선택 기법에서 지시변수에 방향성을 부여하는 경우, 정확한 예측변수 선택에 도움이 되어 예측력을 향상 시킬 수 있다. 이와 같은 결과로부터 부동산 정책 결정과 주택담보대출 위험관리 등의 정확한 장기 분포예측이 요구되는 분야에 본 연구가 기여할 수 있을 것으로 기대된다.

본 연구는 연구의 특성상 여러 한계점을 가진다. 첫째, large number of variables 또는 large model space의 조건에 기반한 예측을 실시하기 때문에 과도한 계산비용으로 VAR, BMA 등과 같이 전통적으로 예측에 널리 사용되는 시계열 모형군들과의 직접적인 예측력 비교를 시행하지는 못하였다. 따라서 본 연구가 서울아파트매매가격지수 예측력을 개선할 수 있는 새로운 변수선택기법을 제시하였으나, 이 기계학습 알고리즘이 주택가격 예측을 극대화시킨다고는 주장할 수 없다. 둘째, 변수선택 기법을 적용하는데 있어 BIC, AIC 등과 같은 전통적인 기법은 고려하지 않는다. 이 또한 large model space 조건에 의한 것으로 상수항과 종속변수

시차변수를 제외했을 때, 예측시점마다 2^{480}개에 달하는 BIC를 계산하는 것은 불가능에 가깝다. BIC에 근거한 변수선택 기법과 관련하여 Chen and Chen (2008)은 본 연구와 유사한 환경의 large model space에서 변수선택이 가능하게 하는 extended BIC를 제시하였다. 이후의 연구에서 extended BIC 기법과 본 연구의 베이지안 기법의 비교가 흥미로운 연구주제가 될 것이다. 셋째, 예측변수들에 대해 주성분분석을 적용하여 어떤 변수가 선택되고 어떤 영향을 주는 지는 해석할 수 없다. 주택가격의 주요 결정요인과 파급과정을 분석하기 위해서는 구조적인 모형이 설정되어야 할 것이다. 특히 정책당국자의 입장에서는 주택가격예측과 더불어 주택시장상황에 따라 규제의 효과를 정확하게 예측하는 것이 무엇보다 중요함에도 이런 부분이 분석되지 않았다는 점도 본 연구의 한계이다. 향후 이런 주제들이 후속 연구를 통해 보완될 것으로 기대한다.

■ 참 고 문 헌

1. 강규호, "베이지안 머신 러닝을 이용한 은행권 주택담보대출 예측,"『금융안정연구』, 제19권 제1호, 2018, pp. 99-129.

   (Translated in English) Kang, Kyu-Ho, "Mortgage Loan Prediction: Bayesian Machine Learning Approach," *Financial Stability Studies*, Vol. 19, No. 1, 2018, pp. 99-129.

2. 곽승준 · 이주석, "부동산정책이 주택가격의 변동성 변화에 미치는 영향 - 주택가격의 변동성 변화 시점을 중심으로,"『주택연구』, 제14권 제2호, 2006, pp. 175-194.

   (Translated in English) Kwak, Seung-Jun, and Joo-Suk Lee, "The Impacts of Public Policy on Housing Volatility Changes," *Housing Studies Review*, Vol. 14, No. 2, 2006, pp. 175-194.

3. 김경외 · 김영효, "모델의 불확실성을 반영한 아파트가격지수 예측 모형 연구 -BMS, BMA를 중심으로-,"『부동산분석』, 제1권 제1호, 2015, pp. 27-49.

   (Translated in English) Kim, Keung-Oui, and Young-Hyo Kim, "A Study on Forecasting Model for Apartment Housing Price Index Reflecting Model Uncertainty: Focused on BMS, BMA," *Journal of Real Estate Analysis*, Vol. 1, No. 1, 2015, pp. 27-49.

4. 김윤영, "우리나라 주택시장의 매매 · 전세 가격변동 거시결정요인의 동태분석,"『경제학연구』, 제60집 제3호, 2012, pp. 127-153.

(Translated in English) Kim, Yun-Yeong, "Macroeconomic Determinants of Housing and Housing Lease Prices' Dynamics in Korea," *The Korean Economic Review*, Vol. 60, No. 3, 2012, pp. 127-153.

5. 노영학 · 김종호, "부동산정책이 주택가격에 미치는 영향연구,"『부동산학보』, 제50집, 2012, pp. 108-122.

(Translated in English) Noh, Young-Hak, and Gong-Ho Kim, "A Study on the Impact on Real Estate Policy of Housing Prices," *Korea Real Estate Academy Review*, Vol. 50, 2012, pp. 108-122.

6. 배성완 · 유정석, "머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측,"『주택연구』, 제26권 제1호, 2018, pp. 107-133.

(Translated in English) Bae, Seong-Wan, and Jung-Suk Yu, "Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model," *Housing Studies Review*, Vol. 26, No. 1, 2018, pp. 107-133.

7. 손종칠, "통화정책 및 실물 · 금융변수와 주택가격간 동태적 상관관계 분석,"『경제학연구』, 제58집 제2호, 2010, pp. 179-219.

(Translated in English) Son, Jong-Chil, "Dynamic Analysis of Correlations among Monetary Policy, Real and Financial Variables and Housing Prices," *The Korean Economic Review*, Vol. 58, No. 2, 2010, pp. 179-219.

8. 손정식 · 김관영 · 김용순, "부동산가격 예측모형에 관한 연구,"『주택연구』, 제11권 제1호, 2002, pp. 49-76.

(Translated in English) Son, Jung-Shik, Kwan-Young Kim, and Yong-Sun Kim, "A Study on the Forecasting Model of Real Estate Market : The Case of Korea," *Housing Studies Review*, Vol. 11, No. 1, 2002, pp. 49-76.

9. 송상윤, "BVAR-SSVS와 TVP-BVAR 모형을 이용한 가계부채, 주택가격 및 대출금리의 동학에 관한 연구,"『통계연구』, 제21권 제3호, 2016, pp. 67-98.

(Translated in English) Song, Sang-Yoon, "The Dynamic Relationships between Household Debts, Housing Prices and Interest Rates in Korea: BVAR-SSVS and TVP-BVAR Approach," *Journal of The Korean Official Statistics*, Vol. 21, No. 3, 2016, pp. 67-98.

10. 윤성민 · 손승화 · 이정인, "지역주택가격 변동의 장단기 결정요인에 관한 실증분석,"『부동산학보』, 제67집, 2016, pp. 198-211.

(Translated in English) Yoon, Seong-Min, Seung-hwa Son, and Jung-In Lee, "Empirical Analysis on the Long-Run and Short-Run Determinants of Regional House Price Dynamics," *Korea Real Estate Academy Review*, Vol. 67, 2016, pp. 198-211.

11. 이영수, "단일변수 시계열 모형들의 주택가격지수 예측력 비교,"『부동산학연구』, 제20권 제4호, 2014, pp. 75-94.

(Translated in English) Lee, Young-Soo, "Comparing the Forecasting Performance of Univariate Time Series Models with Korean House Price Index," *Journal of the Korea Real Estate Analysts Association*, Vol. 20, No. 4, 2014, pp. 75-94.

12. 전해정, "패널공적분을 이용한 거시경제변수 및 주택정책이 주택매매가격에 미치는 영향에 관한 연구,"『부동산학보』, 제57집, 2014, pp. 251-263.

(Translated in English) Chun, Hae-Jung, "Empirical Study on Impact of Macroeconomic Variables and Policy on Housing Prices Using Panel Cointegration," *Korea Real Estate Academy Review*, Vol. 57, 2014, pp. 251-263.

13. 전해정 · 박헌수, "주택시장과 거시경제변수 요인들간의 동태적 상관관계 분석,"『주택연구』, 제20권 제2호, 2012, pp. 125-147.

(Translated in English) Chun, Hae-Jung, and Heon-Soo Park, "A Study on Dynamic Correlations between the Housing Market and Factors of Macroeconomic Variables," *Housing Studies Review*, Vol. 20, No. 2, 2012, pp. 125-147.

14. 함종영 · 손재영, "사전확률분포를 이용한 주택시장 예측모형 비교 연구-Bayesian VAR모형을 중심으로,"『부동산 · 도시연구』, 제8권 제2호, 2016, pp. 25-38.

(Translated in English) Ham, Jong-Young, and Jae-Young Son, "Applying the Bayesian Vector Autoregressive Model in House Price Prediction," *Review of Real Estate and Urban Studies*, Vol. 8, No. 2, 2016, pp. 25-38.

15. Chen, J., and Z. Chen, "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, Vol. 95, No. 3, 2008, pp. 759-771.

16. George, E. I., and R. E. McCulloch, "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, Vol. 88, No. 423, 1993, pp. 881-889.

17. George, E. I., D. Sun and S. Ni, "Bayesian Stochastic Search for VAR Model Restrictions," *Journal of Econometrics*, Vol. 142, No. 1, 2008, pp. 553-580.

18. Korobilis, D., "VAR forecasting using Bayesian Variable Selection," *Journal of Applied Econometrics*, Vol. 28, No. 2, 2013, pp. 204-230.

19. O'Hara, R. B., and M. J. Sillanpää, "A Review of Bayesian Variable Selection Methods: What, how and Which," *Bayesian Analysis*, Vol. 4, No. 1, 2009, pp. 85-117.

20. Onorante, L., and A. E. Raftery, "Dynamic Model Averaging in Large Model Spaces using Dynamic Occam's Window," *European Economic Review*, Vol. 81, 2016, pp. 2-14.

21. Raftery, A. E., M. Kárný and P. Ettler, "Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill," *Technometrics*, Vol. 52, No. 1, 2010, pp. 52-66.

A Bayesian Variable Selection Method for Seoul Apartment Price Index Prediction*

Changhoon Lee** · Kyu Ho Kang*** · Jihee Ann****

Abstract

Accurate house price forecasts are essential for efficient policy-making, investment, and risk management of mortgage loan. Nevertheless, there are few empirical studies on the Korean house price prediction. This seems to be because of the large number of variables. In this study, we provide a new Bayesian variable selection method for the Seoul apartment price index forecasting, considering the uncertainty of the variables. To do this, we extend the standard Bayesian variable selection by using a more flexible and interpretable spike-and-slab prior. This method consists of two stages: variable selection and predictive density simulation. According to our out-of-sample forecasting experiment, the proposed model outperforms the standard variable selection method and first-order auto-regressive (AR(p) model in the medium- and long-term horizons. Meanwhile, the AR(p) is found to be the best for the short-term forecasting due to the high persistence of the price index growth.

Key Words: out-of-sample density forecasting, Gibbs-sampling, model selection
JEL Classification: R2, C11, C53

Received: Sept. 5, 2019. Revised: Feb. 7, 2020. Accepted: March 20, 2020.

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01037796).

** First Author, Ph. D. Candidate, Department of Economics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea, Phone: +82-2-3290-5132, e-mail: rollin0807@korea. ac. kr

*** Corresponding Author, Associate Professor, Department of Economics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea, Phone: +82-2-3290-5132, e-mail: kyuho@korea. ac. kr

**** Co-Author, Associate Research Fellow, Korea Real Estate Research Institute, 3F, 52, Bangbae-ro, Seocho-gu, Seoul 06705, Korea, Phone: +82-2-520-5026, e-mail: annji@kreri. re. kr