碩士學位論文

커널능형회귀분석에서 앙상블기법을 이용한 효율성 연구

韓國外國語大學校 大學院
統 計 學 科
韓 善 雨

碩士學位論文

커널 능형 회귀분석에서 앙상블기법을 이용한 효율성 연구

指導 李 碩 浩 敎授

이 論文**을** 碩士學位請求論文**으로** 提出**합니다.**

2015年 **12**月

韓國外國語大學校 大學院
統 計 學 科
韓 善 雨

이 論文**을** 韓善雨**의** 碩士學位論文**으로** 認定**함**

2015年 **12**月 日

審 査 委 員 _____ (**인**)

審 査 委 員 _____ (**인**)

審 査 委 員 _____ (**인**)

韓國外國語大學校 大學院

요약

본 연구는 커널능형회귀(kernel ridge regression)에서 배깅(bagging) 및 랜덤포레스트(random forests)와 같은 앙상블 기법(ensemble method)을 이용하여 추정량의 분산을 축소시킴으로써, 추정량의 정확성을 높이는 방법을 제안한다. 다양한 상황의 커널회귀분석 문제에서 기존의 방법론과의 비교를 컴퓨터 모의실험을 통해 수행하였으며, 다수의 실제데이터에 적용하였다. 이를 통해 본 연구에서 제안한 방법이 일반 커널회귀분석보다 우수한 성능을 보임을 확인하였다.

목 차

표 목 차

그 림 목 차

1 서 론

1.1 연구배경 및 목적

회귀분석(regression analysis)은 여러 통계 분석기법 가운데 가장 널리 쓰이는 것 중 하나로, 경제, 경영, 심리, 교육 등의 사회과학과 물리, 화학, 생물, 환경 등의 자연과학 및 공학, 농학 및 의약학 등 거의 모든 학문 분야에서 광범위하게 사용되고 있다 (강기훈, 2009). 변수(variable) 간의 함수적 관계를 밝히려는 통계적 분석 방법을 회귀분석이라 하고, 많은 분야에서 반응변수(dependent variable)와 설명변수(independent variable) 간의 상호 관련성을 찾는 연구가 활발히 진행되고 있다. 최근에는 컴퓨터 기술의 발전으로 방대하고, 다양한 형태의 데이터를 실시간으로 처리할 수 있게 됨에 따라 사회 전반적인 분야에서 빅데이터(big data)를 활용하여 사업 전략을 수립하는데 회귀분석 문제가 빈번하게 적용된다. 이미 해외국가에서 정부과제나 기업의 사업전략을 수립함에 빅데이터와 회귀분석을 결합한 분석이 수차례 성공사례로 보고되었다. 하지만 다양한 분야에서 다뤄지는 빅데이터는 단순히 관측치의 수만 증가함을 의미하지 않고 설명변수의 수도 상당히 증가하여 의미있는 결과를 도출하는 데 많은 어려움을 가져온다. 예를 들어 미래를 예측하는 회귀분석 문제에서 설명변수의 증가는 모델에서 그만큼 많은 모수를 추정해야 함을 의미한다. 데이터를 잘 설명하는 모델을 찾기 위해서 각 설명변수의 차수(polynomial order)와 변환(transformation) 여부를 결정하는 문제가 발생하고 설명변수 간의 교호작용(interaction effect) 여부 역시 고려해야 한다. 따라서 모델이 복잡해짐에 따라

고려해야 할 사항이 비례하게 많아지므로 알맞은 모델을 찾는 것은 현실적으로 큰 어려움이 따른다. 또 다른 문제로 설명변수가 많아짐에 따라 유사 설명변수 간의 높은 상관관계를 가지는 다중공선성(multicollinearity) 문제도 발생할 개연성이 높아진다. 일반적으로 다중공선성 문제가 존재하면 추정치의 분산이 과도하게 커지는 문제가 발생하여 추정량 및 예측값의 신뢰성이 떨어지고 잘못된 예측결과나 분류결과를 가져오는 심각한 문제를 발생시킬 수 있다. 따라서 본 논문에서는 위와 같은 문제점을 부분적으로 해결하기 위한 추정기법에 관한 연구를 수행하였다.

1.2 연구방법 및 구성

일반적으로 회귀분석 문제에서 반응변수에 영향을 주는 설명변수의 변환식이 복잡하거나 설명변수간의 교호작용을 고려하는 경우, 복잡한 모형을 사전에 설정하기 않으면서 이를 자동으로 모형화하는 방법으로 커널기법(kernel method)을 이용한다. 커널 기법은 분류(classification) 문제인 지지도벡터기계(support vector machine)에서 시작되었던 분석 기법이지만, 최근에는 분류문제뿐만 아니라 예측(regression)문제에서도 활발한 연구가 진행되고 있다. 커널기법과 회귀분석을 결합한 커널능형회귀법(kernel ridge regression)은 복잡한 형태의 회귀모델을 분석전에 설정할 필요없이 데이터 기반으로 자동으로 설정되어 반응변수를 비교적 정확하게 예측하는 장점이 있다. 또한, 다수의 서브모형(submodels)을 종합하여 예측력을 높히고, 유사한 설명변수들의 증가로

발생하는 다중공선성 문제를 극복하기 위해 앙상블 기법(ensemble method)을 제안한다. 앙상블기법으로, 동일한 분포적 성질을 갖는 다수의 서브모형을 평균함으로써 추정치의 분산을 줄이는 데 초점을 맞춘 배깅기법(bagging method)과 전체 모집단의 샘플(sample)을 추출하는 과정에서 일부의 설명변수를 선택하여 서브모형 간의 상관성을 줄여 최종모형의 분산을 줄이는 랜덤 포레스트기법(random forests)을 사용한다. 따라서 본 논문은 커널능형회귀에 두 가지의 앙상블 기법을 적용하여 회귀모형의 예측력을 높이는 방안을 제시한다.

본 논문은 총 6장으로 구성되어 있다. 2장에서 커널능형회귀법에 대해 개관하고, 3장에서 앙상블 기법을 소개한다. 4장에서는 커널능형회귀에 앙상블 기법을 적용한 분석기법을 제안한다. 5장에서는 제안된 방법의 성능을 입증하기 위한 모의실험 설계방법과 분석결과를 서술하였고, 실제 자료에 적용하여 얻은 결과를 수록하였다. 6장은 결론으로서 본 연구의 결과를 요약하였다.

2 커널능형회귀분석법

2.1 능형 회귀분석(ridge regression)

다중회귀분석(multivariate regression analysis)은 하나의 반응변수와 여러 개의 설명변수 사이의 함수적 관계를 모델링(modeling)하여 예측모형을 구축하는 통계적 분석방법이다. n개의 데이터 $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$이 있다

고 가정하자. 자료의 개수가 n이고 k개의 설명변수를 가지는 다중선형회귀모형은 다음과 같이 나타낼 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

위 식에서 ϵ_i은 평균이 0이고 분산은 σ^2이며 서로독립인 확률오차(random error)이다. 이를 행렬형태로 표현하면 다음과 같다.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

여기에서 $y = (y_1, y_2, \ldots, y_n)^T$는 길이가 n인 반응변수벡터이고,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

는 크기 $n \times p$ $(p = k + 1)$인 자료행렬, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$는 길이가 p인 회귀계수벡터, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$은 길이가 n인 오차벡터라고 하면, 위의 다중선형회귀모형은 다음과 같은 행렬식으로 다시 쓸 수 있다.

$$y = X^T \beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n).$$

위에서 0은 길이가 n인 영벡터이고 I_n는 크기가 n인 단위행렬이다. 일반적으로 회귀계수 β의 추정은 잔차의 제곱을 최소화하는 최소제곱법(least squares estimation; LSE)을 사용한다. 최소제곱법으로 얻은 정규방정식(normal equation)의 행렬형태는 다음과 같고,

$$X^T X \hat{\beta} = X^T y$$

자료행렬 X가 full-rank일 경우, 회귀계수 β는 다음과 같은 유일한 해로 주어진다.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

능형회귀(ridge regression)는 추정량에 약간의 편의(bias)를 주는 대신, 분산(variance)을 큰 폭으로 줄여 추정량의 정확도를 높이고 모형의 예측력을 향상시키는 분석 방법이다. 이를 위해 회귀계수를 0에 가깝게 축소추정(shrinkage estimation)하여 예측력을 높힌다. 능형회귀의 회귀계수를 추정하는 방법은 기존의 최소제곱법에서 회귀계수에 대한 L_2 벌점함수(L_2 penalty)를 추가한 목적함수로부터 얻게 된다.

$$\min_{\beta} \sum_{i=1}^{n} \|y_i - \beta^T X_i\|^2 + \lambda \|\beta\|^2 \quad (\lambda > 0).$$

$\lambda\|\beta\|^2$는 회귀계수의 L_2 노름(norm)의 제곱에 비례하도록 벌점을 부여한다. 양의 실수를 갖는 λ는 조절모수(tuning parameter)로써, 일반적으로 λ가 커지면 능형회귀의 추정량의 편의는 커지는 반면 유연성(flexibility)이 감소해 분산은 작아지는 효과가 있다. 반대로 λ가 작아지면 능형회귀의 추정량은 최소제곱추정량에 가까워져 편의는 작아지는 반면, 유연성이 증가해 분산은 커지는 효과가 있다. 모형의 예측력은 최적의 λ를 설정함으로써 이룰수 있는데, 일반적으로 CV(cross validation)를 이용해 test MSE(mean squared error)의 추정량을 가장 작게 하는 λ값을 선택하여 분석에 사용한다.

2.2 커널능형회귀분석법

반응변수의 영향을 주는 설명변수의 함수식이 복잡한 비선형 형태로 주어진다면, 데이터를 잘 설명하는 회귀모형을 사전에 식별하는 데에 많은 어려움이 따른다. 비선형모형을 다항함수 및 푸리에함수(Fourier functions) 등의 알려진 기저함수(basis function)를 이용해 적합(fitting)하게 되면 필요 이상의 많은 수의 기저함수를 필요로 하게 되며, 또한 설명변수 간의 교호작용을 고려하는 경우 복잡한 기저함수를 고려해야만 한다. 이런 이유로 최근 비선형구조를 선형구조로 변환시킬 수 있는 커널트릭 기법(kernel-trick method)이 많은 관심을 받고 있다. 커널트릭 기법이란 데이터에 적절한 형태의 사상함수(mapping function) Φ를 취하여 설명변수 공간을 고차원의 특성공간(feature space)으로 변형하여 특성공간 하에서 선형모형을 적합하는 기법이다. 여기서 "트릭"이라는 말이 붙은 이유는, 특성공간으로 변형할 때, 명시적으로 변환함수 Φ를 설정하는 것이 아니라 커널(kernel)함수를 이용하여 내부적으로 변환함수가 설정되고 그 형태는 데이터에 기반하여 자동으로 정해짐으로, 사용자가 그 형태를 구체적으로 볼 필요가 없기 때문이다.

자료의 개수가 n개인 데이터에서 설명변수 $x_1, x_2, \ldots, x_n$이 p차원 유클리디안 공간(Euclidean space)에 분포한다고 가정하자. 이에 대한 일반적인 선형모형은 다음과 같다.

$$y = X^T\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n).$$

p차원 공간상의 점으로 표현되는 설명변수 $x_1, x_2, \ldots, x_n$을 사상함수 Φ를

이용해 다음과 같이 변환한다.

$$x_1, x_2, \ldots, x_n \mapsto \Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n).$$

변환된 설명변수 $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$이 놓이는 공간을 특성공간이라 한다. 변환된 설명변수를 이용하여 특성공간 상에서의 선형모형은 다음과 같다.

$$E(y|x_1, \ldots, x_n) = d_1\Phi(x_1) + d_2\Phi(x_2) + \cdots + d_n\Phi(x_n)$$

위 방정식에서 새로운 회귀계수를 $d = (d_1, d_2, \ldots, d_n)^T$로 설정하였다. 변환된 설명변수 $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$와 새로운 회귀 계수 d를 이용하여 원래 데이터의 반응변수 y를 설명하기 위해서는 다음과 같은 $\Phi(X)$의 내적을 계산해야 한다.

$$\sum_{i'=1}^{n} \langle \Phi(x_i), \Phi(x_{i'}) \rangle d_{i'} = \sum_{i'=1}^{n} K_{i,i'} d_{i'}.$$

여기에서 $K_{i,i'}$는 특성공간 상의 점들 간의 내적인 $\langle \Phi(x_i), \Phi(x_{i'}) \rangle$이고 이를 $n \times n$ 크기의 행렬 $K = (K_{i,i'})$로 표현할 수 있다. 커널트릭에서 중요한 점은, 내적을 계산할 때 변환함수 Φ를 설정하고 이를 이용하여 내적을 계산하는 것이 아니라 커널함수(kernel function) $k(\cdot, \cdot)$을 이용하여 내적을 계산한다는 점이다. 즉, $K_{i,i'} = \langle \Phi(x_i), \Phi(x_i') \rangle = k(x_i, x_i')$로 계산한다. 커널함수는 양정치(positive definite)이며 reproducing 성질을 가지는 함수로 정의된다. 커널 이론에 의하면 사용되는 커널함수에 따라 이에 대응하는 변환함수가 존재한다. 자세한 이론은 Schölkopf and Smola (2002)를 참조하고 간략한 설명은 Hastie et al. (2009)를 참고하길 바란다. 내적을 계산시 커널함수 사용의 장점은 설명변수의 차원을 높히거나 교호작용 항을 증가시킴에 따라 계산의 양이

기하급수적으로 증가하게 되는 반면, 커널함수를 사용하면 많은 양의 계산을 비교적 쉽게 해결할 수 있다는 점이다.

위의 과정에서 y와 이에 대한 기대값 Kd를 결합시키면 다음과 같은 최소제곱추정량을 얻을 수 있다.

$$\hat{d} = K^{-1}y.$$

일반적으로 위의 식은 K가 역행렬이 존재하는 경우 성립한다. 하지만 일반적으로 K의 역행렬은 존재하지 않는다. 따라서 능형회귀와 같이 능형회귀 형태의 벌점(ridge-type penalty)함수를 이용한다. 즉, 최소제곱법의 최적화문제를 바꾸어 다음을 최소화 하는 문제로 변환된다.

$$\min_{d} \|y - Kd\|^2 + \lambda d^T K d, \quad \lambda \geq 0.$$

위의 최적화 문제에서 회귀계수 d는 다음과 같다.

$$\hat{d} = (K + \lambda I_n)^{-1}y.$$

훈련자료(trainning data)의 X로 모델을 적합하고, 검증자료(test data)의 X^*로 모델을 평가하는 경우에는 X와 X^* 사이의 내적으로 만들어진 새로운 K^*와 훈련자료의 적합과정에서 얻은 $\hat{d}$을 이용하여 검증자료의 예측값 $\hat{y}^*$을 얻을 수 있다.

$$\hat{y}^* = K^{*T}\hat{d}.$$

3 앙상블 기법(ensemble method)

앙상블 기법(ensemble method)은 기계학습(machine learning) 분야에서 주로 활용되는 기법으로 그 목적은 향상된 추정치를 얻기 위해서 여러 추정치의 예측값을 결합하는 것이다. 그 대표적인 방법은 배깅(bagging)기법과 랜덤포레스트(random forests method) 기법이 있다. 이 두 방법은 서로 독립적인 여러 추정치의 예측값을 평균 내는 방법으로 추정치의 분산을 상당히 줄여 기존의 한번 분석해 얻은 추정치보다 예측 면에서 향상된 결과를 얻는다. 앙상블 기법을 커널능형회귀에 이용하여 추정치의 분산을 줄여 보다 우수한 예측력을 가지는 새로운 회귀분석 방법을 연구에서 제안한다.

3.1 배깅(bagging)

배깅 기법의 주요 목적은 서로 다른 표본(sample)에서 얻은 추정치를 하나로 결합해 분산을 줄여 예측력을 향상시키는데 있다. 하지만 실제로 대부분의 경우에서는 하나의 데이터만 존재하는 경우가 많다. 이 경우 하나의 훈련자료에서 반복적으로 표본을 추출하는 붓스트랩(bootstrap) 기법을 사용하여 여러 개의 붓스트랩 훈련자료(bootstrap training data)를 얻는다. 이러한 이유로 배깅 기법은 bootstrap aggregation으로 명명되었고 이에 대한 약자로 bagging이라 불린다.

평균화를 통해 분산이 줄어드는 현상은 다음의 간단한 예를 통해 쉽게 이해될 수 있다. 분산이 σ^2인 서로 독립적인 n개의 관측치 $Z_1, Z_2, \ldots, Z_n$이

있다고 가정하자. 그렇다면 관측치 Z의 평균 $\bar{Z}$의 분산은 σ^2/n이다. 즉 어떤 관측치에 평균을 취하면 관측치의 수와 비례하여 분산이 작아지게 된다. 동일한 맥락에서, 공통모집단에서 추출한 서로 다른 표본에 각각 동일한 모형을 적합 시키면 추정량 혹은 예측치는 서로 독립이다. 이러한 예측 결과에 평균을 취하게 되면 낮은 분산을 갖는 추정치를 얻을 수 있다.

서로 다른 B개의 붓스트랩 훈련자료가 있다고 가정하자. 각각의 데이터에서 얻은 추정치를 다음과 같이 나타내자.

$$\hat{f}^1_{\text{bag}}(x),\ \hat{f}^2_{\text{bag}}(x),\ \ldots,\ \hat{f}^B_{\text{bag}}(x).$$

각각의 추정치를 결합해 하나의 향상된 추정치를 얻기 위해서 다음과 같은 방식으로 평균을 취하면 상대적으로 낮은 분산을 지닌, 예측력이 우수한 배깅추정치(bagging estimate)를 얻을 수 있다.

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b_{\text{bag}}(x).$$

3.2 랜덤포레스트(random forests)

랜덤포레스트는 서로 다른 표본으로부터 얻은 추정치를 결합해 분산을 줄이는 방식은 배깅과 동일하지만, 각 표본에서 추정시, 모든 변수를 활용하지 않고 일부분만 선택하여 모형을 적합한다는 점이 다르다. 각 붓스트랩 훈련자료는 모집단인 훈련자료와 동일한 분포를 지니고 있으므로 각 붓스트랩 훈련자료로부터 적합된 모형은 서로 유사할 것이고, 이는 추정치들의 양의 연관성을

지닐 개연성이 높다. 이는 배깅의 기본 개념인 독립자료에 대한 평균이 분산을 줄인다는 기본전제가 어긋날 가능성이 높다는 의미이다. 하지만, 각 붓스트랩 훈련자료에 대해 랜덤하게 일부의 설명변수만 사용하여 모형 적합을 하게 되면, 사용되는 변수가 상이하여 서로 연관성이 낮아지게 된다. 물론, 변수의 개수가 줄어든 만큼 개개의 붓스트랩 추정량은 편의가 커져 예측력이 낮아지게 된다. 하지만, 상관성을 줄임(decorrelation)으로 인한 분산의 감소가 이를 상쇄하여, 전체적으로 예측력이 향상하는 장점이 있다.

전체 변수의 개수가 p인 훈련자료가 있다고 가정하자. 서로 다른 붓스트랩 훈련자료를 추출할 때 전체 p개의 설명변수 중 무작위로 m개를 선택하여 붓스트랩 훈련자료에 포함한다. 일반적으로 선택되는 변수의 개수는 $m \approx \sqrt{p}$을 사용한다. 각각의 붓스트랩 훈련자료에서 얻은 추정치를 다음과 같이 나타내자.

$$\hat{f}_{\mathrm{rf}}^{1}(x),\ \hat{f}_{\mathrm{rf}}^{2}(x),\ \ldots,\ \hat{f}_{\mathrm{rf}}^{B}(x).$$

위의 추정치들은 배깅 기법을 이용해 얻은 추정치보다 서로 비연관(uncorrelated)되어 더 큰 폭으로 분산을 줄일 수 있다. 위 과정을 비연관화(decorrelating)라고 한다. 위의 각각의 추정치를 결합해 하나의 향상된 추정치를 얻기 위해서 다음과 같은 식으로 평균을 취하면 상대적으로 낮은 분산을 가지는 예측력이 우수한 랜덤포레스트 추정치(random forests estimate)를 얻을 수 있다.

$$\hat{f}_{\mathrm{rf}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{\mathrm{rf}}^{b}(x).$$

4 앙상블 기법을 이용한 커널능형회귀분석법

3장에서 기술한 바와 같이, 앙상블 기법 중 배깅, 랜덤포레스트 기법 모두 추정치의 분산을 줄이는 방법으로 예측력을 향상시키는 이점을 가지고 있다. 이 두 가지 방법을 기존에 활용되고 있는 커널능형회귀분석에 적용한다면 더 우수한 예측력을 가지는 추정치를 얻을 수 있을 것이라는 직관적인 이해를 할 수 있다. 따라서 본 연구에서는 기존의 커널능형회귀에서 배깅 기법과 랜덤포레스트 기법을 이용하여 추정치의 분산을 안정시켜 향상된 예측력을 가지는 새로운 회귀 분석방법을 제안한다.

4.1 배깅 기법을 이용한 커널능형회귀분석법

배깅 기법을 이용한 커널 능형 회귀법은 다음과 같다. 전체 관측치의 수가 n개 이고, p차원 유클리디안 공간상에 설명변수가 존재하는 훈련자료가 있다고 가정하자. 훈련자료에서 중복을 허용하여 무작위로 n개의 자료를 추출하여 이를 다음과 같이 표기한다.

$$(y_1^b, x_1^b),\ (y_2^b, x_2^b),\ \ \ldots,\ (y_n^b, x_n^b).$$

위의 설명변수 $x_1^b, x_2^b, \ldots, x_n^b$를 다음과 같은 변환을 통해 특성공간에 위치시킨다.

$$x_1^b, x_2^b, \ldots, x_n^b \mapsto \Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b).$$

변환된 설명변수 $\Phi(x_1^b), \Phi(x_2^b), \ldots, \Phi(x_n^b)$를 이용해 다음과 같은 새로운

선형모형을 적합한다.

$$E(y^b|x_1^b,\ldots,x_n^b) = d_1^b\Phi(x_1^b) + d_2^b\Phi(x_2^b) + \cdots + d_n^b\Phi(x_n^b).$$

위의 내용은 개념적 설명이며 실제 분석을 위한 내적의 계산은 커널함수를 이용한다. 이는 다음과 같다. 변환된 설명변수와 회귀계수 $d^b = (d_1^b, d_2^b, \ldots, d_n^b)^T$를 이용하여 반응변수인 y^b의 예측모형 적합을 위해서 다음과 같은 $\Phi(X^b)$의 내적을 계산한다.

$$\sum_{i'=1}^{n}\langle\Phi(x_i^b),\Phi(x_{i'}^b)\rangle d_{i'}^b = \sum_{i'=1}^{n} K_{i,i'}^b d_{i'}^b.$$

이때, 내적의 계산은 주어진 커널함수 $k(\cdot,\cdot)$을 이용하여 다음과 같이 계산하고 $K^b = (K_{i,i'}^b)$인 $n \times n$ 행렬을 구성한다.

$$K_{i,i'}^b = \langle\Phi(x_i^b),\Phi(x_{i'}^b)\rangle = k(x_i^b, x_{i'}^b).$$

따라서 b번째 붓스트랩 훈련자료를 통한 모형적합은 아래의 최적화문제로 풀게 된다.

$$\|y^b - K^b d^b\|^2 + \lambda {d^b}^T K^b d^b, \quad \lambda \geq 0.$$

회귀계수 d^b는 다음과 같은 식을 풀어주면 얻을 수 있다.

$$\hat{d^b} = (K^b + \lambda I_N)^{-1} y^b.$$

따라서 예측값 $\hat{y}^b$는 다음과 같이 얻을 수 있다.

$$\hat{y}^b = {K^b}^T \hat{d^b}.$$

붓스트랩 훈련자료의 X^b로 모델을 적합하고, 검증자료의 X^*로 모델을 평가하는 경우에는 X^b와 X^* 사이의 내적으로 만들어진 새로운 K^{b*}와 붓스트랩 훈련자료를 이용한 $\hat{d}^b$를 이용해 다음과 같이 검증자료의 예측값을 계산한다.

$$\hat{y}_{\text{bag}}^{b*} = K^{b*T}\hat{d}^b.$$

위와 같은 과정을 $b = 1, 2, \ldots, B$를 거쳐 반복하여 $\hat{y}_{\text{bag}}^{b*}$를 얻는다. 이를 평균하면 배깅 기법의 최종 추정치인 $\hat{y}_{\text{bag}}^{*} = \sum_{b=1}^{B} \hat{y}_{\text{bag}}^{b*}/B$을 얻을 수 있다.

4.2 랜덤포레스트 기법을 이용한 커널능형회귀분석법

랜덤포레스트 기법을 이용한 커널 능형 회귀법은 다음과 같다. 전체 관측치의 수가 n개 이고, p차원 유클리디안 공간상에 존재하는 훈련자료가 있다고 가정하자. 훈련자료의 설명변수 p개 중 m개를 랜덤하게 선택하고 중복을 허용하여 무작위로 n개의 자료를 추출하여 이를 다음과 같이 표기하자.

$$(y_1^b, x_1^b), (y_2^b, x_2^b), \ldots, (y_n^b, x_n^b).$$

여기에서 $m \approx \sqrt{p}$를 사용하였고, m이 정수가 되지 않을 경우는 소수점 버림하였다. 여기서의 붓스트랩 훈련자료의 설명변수 x_i^b가 배깅에서의 설명변수와 다른 점은, 배깅의 경우 x_i^b는 모든 설명변수값을 취하고 있으므로 길이가 p인 벡터인 반면, 랜덤포레스트의 경우 x_i^b의 길이는 m이며, p개의 설명변수 중 랜덤하게 선택된 m개만 포함된다는 점이다. 주의할 점은 각 $b = 1, 2, \ldots, B$의 붓스트랩 훈련자료마다 m개의 설명변수를 포함하되, 각 붓스트랩 훈련자료별로 선택된 m개의 설명변수들은 다르다는 점이다.

이후 최종예측값을 계산하는 방법은 배깅와 동일하므로 계산법은 생략한다. 대신 배깅 예측값과 구분하기 위해 랜덤포레스트 예측값을 $\hat{y}^*_{\mathrm{rf}} = \sum_{b=1}^{B} \hat{y}^{b*}_{\mathrm{rf}}/B$ 으로 표기한다. 여기서 $\hat{y}^{b*}_{\mathrm{rf}}$는 각 붓스트랩 훈련자료를 통해 얻은 검증자료의 예측값이다.

5 모의실험 및 실증 분석

본 장에서는 기존에 활용되고 있는 2가지 커널회귀방법과 본 논문에서 제안하는 2가지 방법 총 4가지 방법에 관해 모의실험과 실증 분석으로 각각의 성능을 비교하였다. 모든 모의실험 및 실증 분석은 통계프로그램 R을 이용하였다.

5.1 모의실험

모의실험은 설명변수의 개수 p가 1, 2, 5, 10 인 경우 각각에 훈련자료의 개수 n을 50, 100, 200, 400으로 달리하여 총 16가지의 상황을 가정하여 4가지 방법론에 대해 비교 분석하였다. 비교하고자 하는 방법은 다음과 같다.

- KR : kernel ridge regression
- KRS : kernel ridge regression using sub-sampling
- KRB : kernel ridge regression using bagging
- KRR : kernel ridge regression using random forests

위의 모형 중, KRS 방법은 Huh (2015)에 의해 제안된 방법이다.

각 상황에서 위의 4가지 분석방법을 다음의 단계(step)에 따라 자료를 생성하여 모의실험을 진행하였다.

(Step1) 훈련자료(training data)와 평가를 위한 검증자료(test data)를 각각 생성한다. 반응변수는 $y_i|x \sim N(\mu_i, 0.5^2)$를 통해 생성한다. 이때 평균 μ_i는 다음과 같이 주어진다.

$$\mu_i = 1 + \sum_{j=1}^{p} \left(\frac{x_{ij}}{2}\right)^j.$$

훈련자료의 개수는 $i = 1, 2, \ldots, n$이다. 각 상황에 따라 n은 50, 100, 200, 400으로 설정하였다. 검증자료의 관측치 수는 1000개로 고정한다. 설명변수 x_{ij}은 $(-2, 2)$사이의 균일분포에서 독립적으로 추출하였다.

(Step2) $\|y - Kd\|^2 + \lambda d^T K d$에 사용되는 능형 모수(ridge parameter) λ를 생성한다.

능형 모수 λ는 5-fold cross validation으로 최적인 λ를 선택하여 분석에 사용하였다. KR에서는 한번 분석하여 얻은 λ를 사용하였고, KRS, KRB, KRR에서는 100번 반복하여 얻은 λ 중 test RMSE를 최소로 하는 λ를 선택하였다.

(Step3) $\hat{d} = (K + \lambda I_n)^{-1} y$ 식을 이용하여 추정치를 구한다.

KRS에서 모델 적합에 사용한 sub-sampling의 자료의 수는 훈련자료의 수가 50, 100, 200일 경우 전체 자료의 0.75만큼을 사용하였고, 자료의 수가 400일 경우 전체 자료의 0.65만큼 사용하여 분석하였다. KRB, KRR

에서는 서로 다른 1000개의 붓스트랩 훈련자료에 대하여 $\hat{y}_{\text{bag}}^{b*}$와 $\hat{y}_{\text{rf}}^{b*}$를 얻은 후, 이를 평균하여 $\hat{y}_{\text{bag}}^{*}$ 및 $\hat{y}_{\text{rf}}^{*}$를 계산한다.

(Step4) 검증자료에 대한 RMSE(root mean squared error)를 구한다.
4가지 방법에 대하여 $RMSE = \sqrt{\frac{(\hat{y}-\hat{y}_{\text{test}})^2}{N}}$를 계산한다. 여기서 $\hat{y}_{\text{test}}$는 각 방법으로 통해 얻은 검증자료의 예측치를 의미한다.

(Step5) 각 방법을 통해 얻은 RMSE의 결과가 항상 동일하지 않음으로 100회 반복하여 얻은 결과를 평균하여 모형의 성능을 비교한다.

KR, KRS, KRB, KRR 4가지 방법들을 적용하여 모의실험 결과를 얻었다. 모의실험 설계에서 제안한 16가지 상황에 대하여 각각 100회 반복하여 얻은 예측치의 RMSE 값을 비교하여 박스 그림(box-plot)과 표로 나타내었다.

다음은 설명변수 p를 1, 2, 5, 10개로 고정하고 관측치 n을 50, 100, 200, 400의 경우로 바꿔가며 얻은 결과를 나타낸 박스 그림이다.

그림 1: 설명변수 $p = 1$이고 n에 따른 RMSE 비교그림.

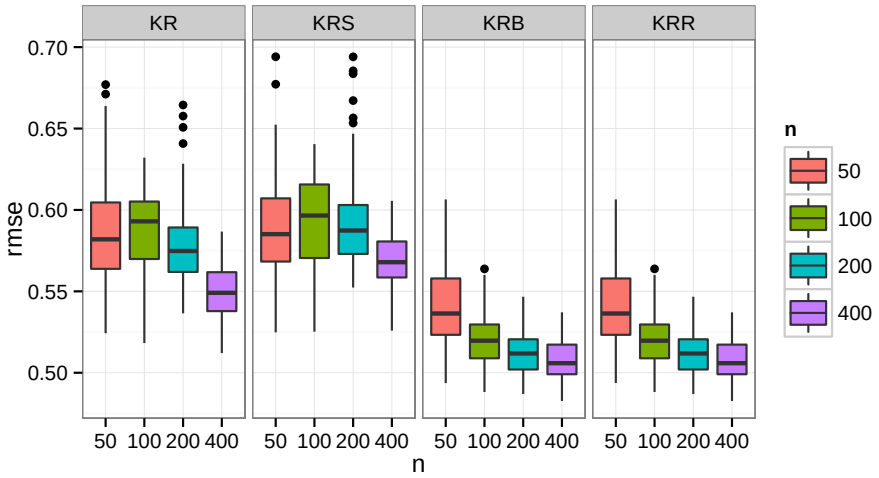


그림 2: 설명변수 $p = 2$이고 n에 따른 RMSE 비교그림.

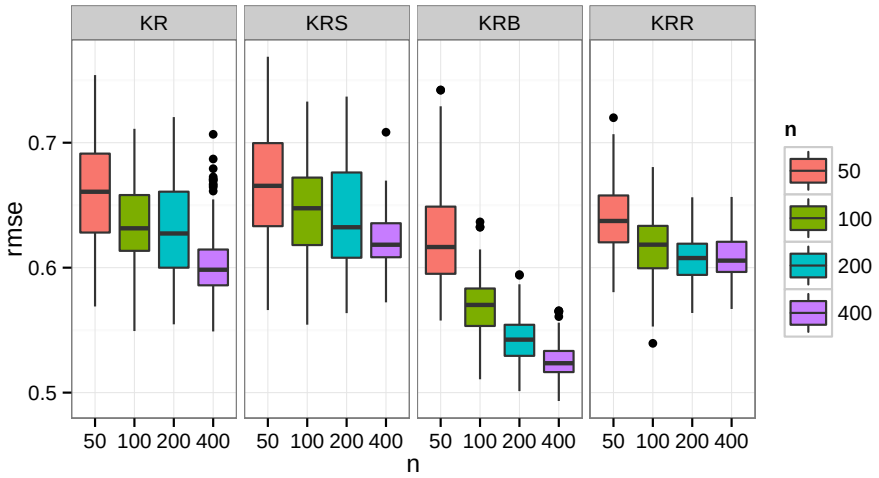


그림 3: 설명변수 $p = 5$이고 n에 따른 RMSE 비교그림.

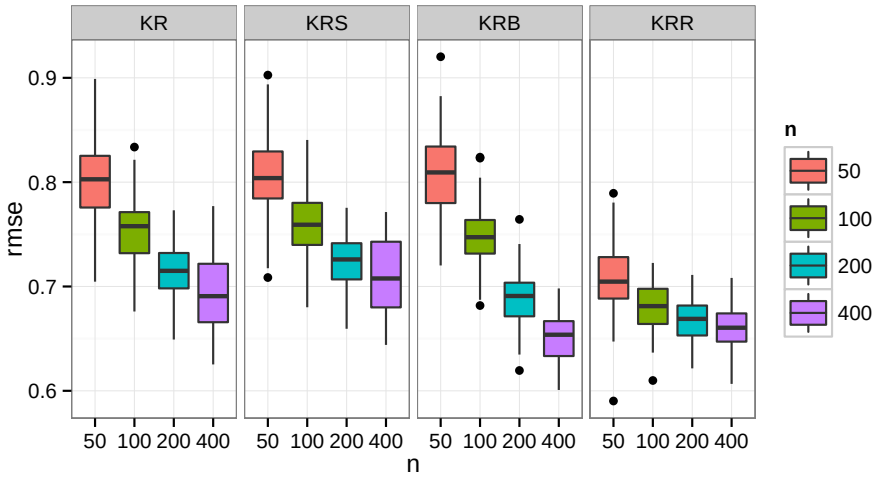
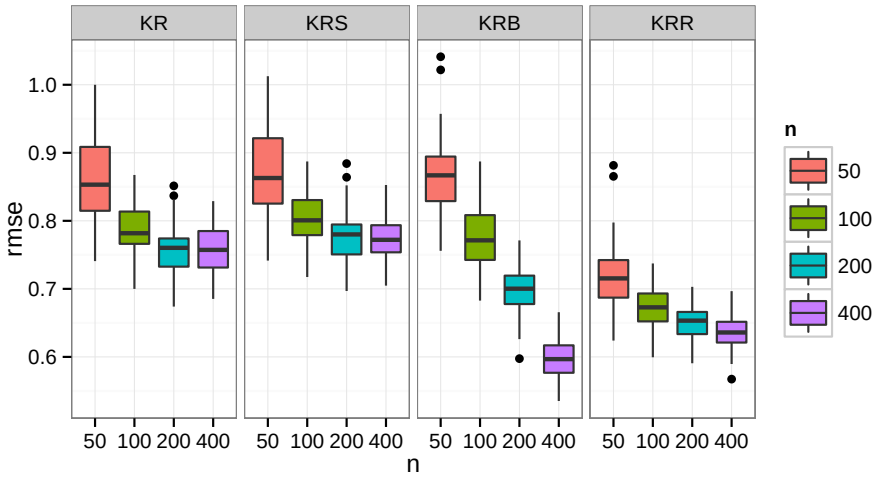


그림 4: 설명변수 $p = 10$이고 n에 따른 RMSE 비교그림.



가로축은 KR, KRS, KRB, KRR의 각각의 방법을 나타내고 각 방법당 관측치 n을 50, 100, 200, 400으로 진행하였기 때문에 4개씩 존재한다. 세로축은 실제값과 추정값의 차이인 RMSE를 나타낸다.

대체로 설명변수 p의 개수와 상관없이 모든 모의실험 상황에서 각 방법의 관측치의 수가 증가함에 따라 RMSE의 평균값이 낮아지고 박스의 크기가 작아지는 경향을 보인다. RMSE의 평균값이 낮다는 것은 실제값과 추정값 사이의 차이가 작다는 것을 의미하기 때문에 추정한 모델이 실제 데이터를 잘 설명한다고 할 수 있다. 또한, 박스의 크기가 작다는 것은 RMSE의 분산이 작

음을 의미하기 때문에 더 신뢰할만한 예측값을 얻었다고 할 수 있다. 따라서, 훈련자료의 관측치가 많아질수록 모델의 적합에 사용할 수 있는 자료가 많아져 보다 정확한 예측이 가능하고 신뢰성이 높은 추정치를 얻을 수 있다. 또한, 모든 모의실험 상황에서 기존에 활용되고 있는 방법인 KR, KRS의 RMSE의 평균값보다 KRB, KRR의 RMSE 평균값이 더 낮음을 알 수 있고 박스의 크기가 작아지는 경향을 볼 수 있다. 따라서 본 논문에서 제안하는 앙상블 기법(ensemble method)을 이용한 커널능형회귀(kernel ridge regression)가 예측성과 신뢰성 측면에서 더 우수한 성능을 지니고 있다고 할 수 있다.

다음은 설명변수 p를 1, 2, 5, 10개로 고정하고 관측치 n을 50, 100, 200, 400의 경우로 바꿔가며 얻은 결과를 나타낸 표이다.

표 1: 설명변수 $p = 1$이고 n에 따른 RMSE 비교 표.

n		KR	KRS	KRB	KRR
50	mean	0.5853	0.5887	0.5397	0.5397
	sd	0.0307	0.0304	0.0243	0.0243
100	mean	0.5879	0.5929	0.5207	0.5207
	sd	0.0244	0.0282	0.0155	0.0155
200	mean	0.5785	0.5919	0.5117	0.5117
	sd	0.0254	0.0288	0.0134	0.0134
400	mean	0.5499	0.5690	0.5075	0.5075
	sd	0.0164	0.0176	0.0124	0.0124

표 2: 설명변수 $p = 2$이고 n에 따른 RMSE 비교 표.

n		KR	KRS	KRB	KRR
50	mean	0.6593	0.6650	0.6223	0.6390
	sd	0.0442	0.0451	0.0405	0.0281
100	mean	0.6359	0.6453	0.5693	0.6165
	sd	0.0343	0.0393	0.0244	0.0278
200	mean	0.6316	0.6421	0.5424	0.6083
	sd	0.0380	0.0419	0.0197	0.0201
400	mean	0.6051	0.6215	0.5259	0.6081
	sd	0.0306	0.0214	0.0151	0.0196

표 3: 설명변수 $p = 5$이고 n에 따른 RMSE 비교 표.

n		KR	KRS	KRB	KRR
50	mean	0.8017	0.8067	0.8089	0.7082
	sd	0.0385	0.0380	0.0402	0.0296
100	mean	0.7516	0.7589	0.7481	0.6808
	sd	0.0313	0.0326	0.0290	0.0217
200	mean	0.7134	0.7230	0.6878	0.6680
	sd	0.0257	0.0256	0.0242	0.0189
400	mean	0.6930	0.7094	0.6509	0.6601
	sd	0.0341	0.0360	0.0226	0.0203

표 4: 설명변수 $p = 10$이고 n에 따른 RMSE 비교 표.

n		KR	KRS	KRB	KRR
50	mean	0.8621	0.8719	0.8638	0.7159
	sd	0.0629	0.0639	0.0517	0.0426
100	mean	0.7869	0.8037	0.7754	0.6712
	sd	0.0390	0.0382	0.0411	0.0301
200	mean	0.7603	0.7787	0.6975	0.6502
	sd	0.0354	0.0353	0.0309	0.0238
400	mean	0.7574	0.7742	0.5962	0.6368
	sd	0.0353	0.0297	0.0284	0.0231

표에서도 박스 그림의 결과와 동일하게 대체로 설명변수 p의 개수와 상관

없이 모든 모의실험 상황에서 각 방법의 관측치의 수가 증가함에 따라 RMSE의 평균값이 낮아지고 분산이 작아지는 경향을 보인다. 따라서 훈련자료의 관측치가 많아질수록 모델의 적합에 사용할 수 있는 데이터가 많아져 보다 정확한 예측이 가능하고 신뢰성이 높은 추정치를 얻을 수 있다. 또한, 모든 모의실험 상황에서 기존에 활용되고 있는 방법인 KR, KRS의 RMSE의 평균값보다 KRB, KRR의 RMSE의 평균값이 더 낮음을 알 수 있고 분산이 작아지는 경향을 볼 수 있다. 따라서 본 논문에서 제안하는 앙상블 기법(ensemble method)을 이용한 커널능형회귀법(kernel ridge regression)이 예측력과 신뢰성 측면에서 더 우수한 성능을 지니고 있다고 할 수 있다.

5.2 실증 분석

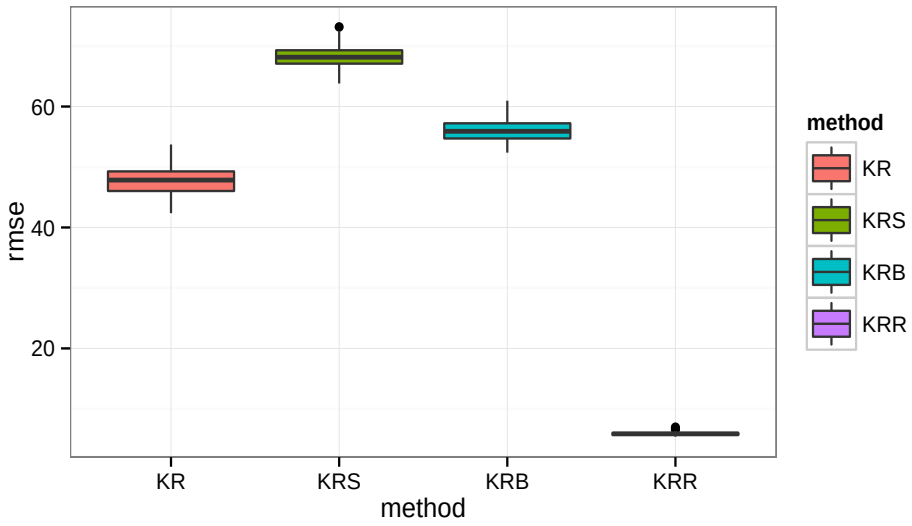
실증분석은 총 8개의 실제 자료를 가지고 KR, KRS, KRB, KRR 4가지 방법의 성능을 비교 분석 하였다. 7개의 실제 데이터 중 커널능형회귀(kernel ridge regression)에 사용되는 능형 모수(ridge parameter) λ가 5-fold cross validation 결과 최적의 λ로 선택되어 4가지 방법 모두 동등한 조건에서 비교 분석한 데이터는 Airfoil self-noise data, Concrete compressive strength data, Energy efficiency data로 총 3개이다. 반면, 1개 또는 2개의 방법에서 5-fold cross validation 결과 최적의 λ가 너무 작은 값이라서 선택되지 않는 경우 동등한 조건에서 비교 분석하지 못했다. 하지만 최적의 λ가 무시해도 될 정도의 아주 작은 값이라는 것은 모델이 L_2-penalty에 영향을 받지 않는 것

을 의미하여 임의로 적절한 λ를 부여해 분석하였다. 이렇게 분석한 데이터는 Computer hardware data, Auto MPG data, Yacht Hydrodynamics data, Housing data 4개이다.

- Airfoil self-noise data

이 자료는 NASA의 데이터로 총 1503($= N$)개의 관측치를 가지고 있고, 1개의 반응변수와 5($= p$)개의 설명변수로 구성되어 있다. 반응변수는 날개에 미치는 압력의 수치이고 연속형 변수이다. 설명변수는 바람의 속도, 불어오는 각도와 같은 날개의 압력에 영향을 주는 요소로 이루어져 있고, 모두 연속형 변수이다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.5(n/N)인 752개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.2($= p^{-1}$)이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 5: Airfoil self-noise data 분석 결과 RMSE 비교 그림.



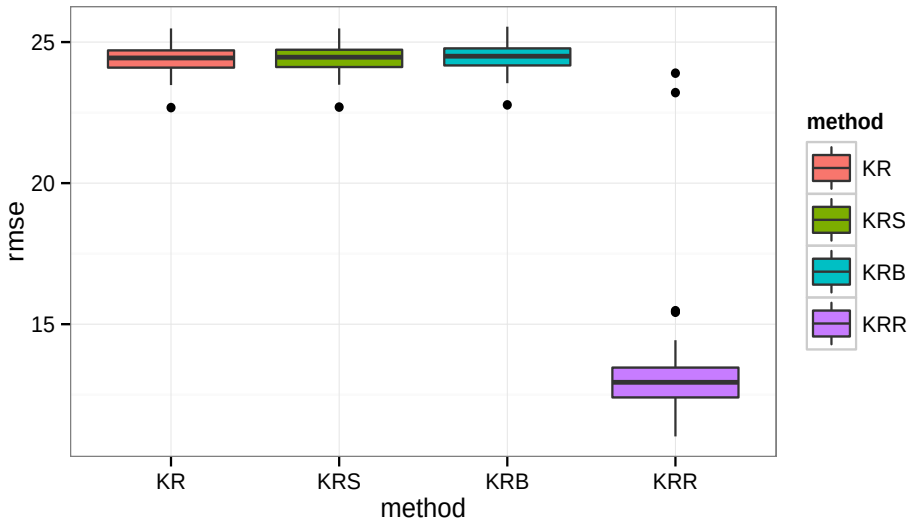
Airfoil self-noise data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 또한, KRR의 박스의 크기도 가장 작다. 따라서 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge regression)이 예측력과 신뢰성 측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

- Auto MPG data

이 자료는 자동차 연비에 관한 데이터로 총 398($= N$)개의 관측치를 가지고

있고, 1개의 반응변수와 8개의 설명변수로 구성되어 있다. 반응변수는 연비를 의미하고 연속형 변수이다. 연비에 영향을 주는 요소인 8개의 설명변수 중 문자형 변수 1개를 제외한 7($= p$)개의 변수를 분석에 사용하였다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.75(n/N)인 299개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.15($= p^{-1}$)이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 6: Auto MPG data 분석 결과 RMSE 비교 그림.

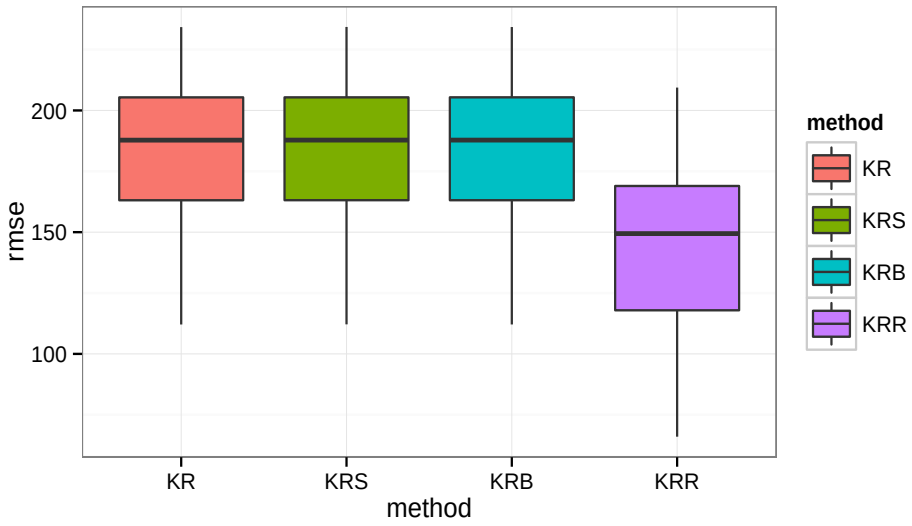


Auto MPG data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 하지만 RMSE의 분산을 나타내는 박스의 크기는 4가지 방법 모두 비슷해 보인다. 이것은 최적의 능형 모수 λ를 선택하는 과정에서 1개 또는 2개의 방법에서 너무 작은 λ가 선택되어 모델에서 penalty가 의미 없는 방법이 존재하여 최적의 λ를 적용하지 못해 동등한 조건에서 비교 분석하지 못했다. 그런데도 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤 포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge regression)이 예측력 측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

- Computer hardware data

이 자료는 컴퓨터 하드웨어 부품인 CPU의 성능에 관한 데이터로 총 209($= N$)개의 관측치를 가지고 있고, 2개의 반응변수와 8개의 설명변수로 구성되어 있다. 2개의 연속형 반응변수 중 CPU의 상대적인 성능을 나타내는 PRP 변수를 반응변수로 선택하였다. CPU의 성능에 영향을 주는 요소인 8개의 설명변수 중 문자형 변수 2개를 제외한 6($= p$)개의 연속형 설명변수를 분석에 사용하였다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.75(n/N)인 279개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.17($= p^{-1}$)이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 7: Computer hardware data 분석 결과 RMSE 비교 그림.



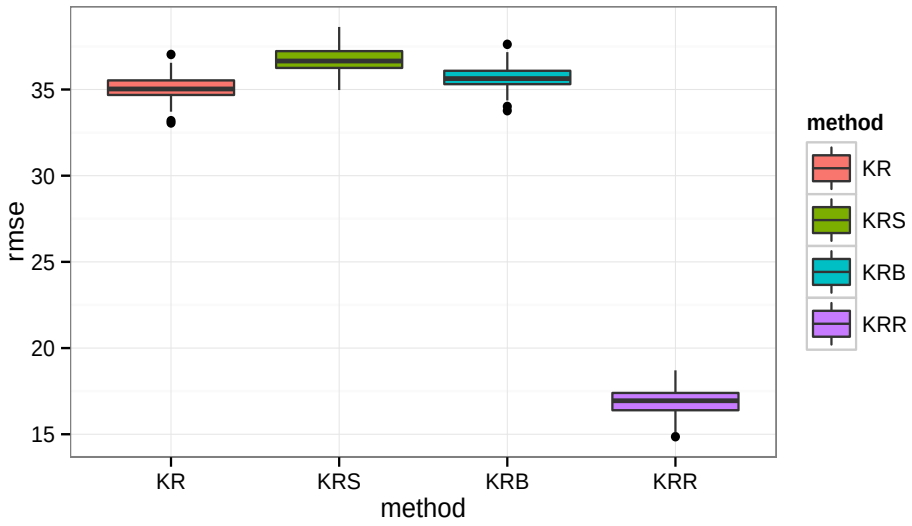
Computer hardware data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 하지만 RMSE의 분산을 나타내는 박스의 크기는 4가지 방법 모두 비슷해 보인다. 이것은 최적의 능형 모수 λ를 선택하는 과정에서 1개 또는 2개의 방법에서 너무 작은 λ가 선택되어 모델에서 penalty가 의미 없는 방법이 존재하여 최적의 λ를 적용하지 못해 동등한 조건에서 비교분석하지 못했다. 그런데도 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge

regression)이 예측력 측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

- Concrete compressive strength data

이 자료는 건축자재인 콘크리트의 내압 강도에 관한 데이터로 총 $103(= N)$개의 관측치를 가지고 있고, 1개의 반응변수와 $8(= p)$개의 설명변수로 구성되어 있다. 반응변수는 콘크리트의 내압 강도이고 연속형 변수이다. 설명변수는 콘크리트의 내압 강도에 영향을 주는 요소로 이루어져 있고, 모두 연속형 변수이다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 $0.75(n/N)$인 78개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 $0.125(= p^{-1})$이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 8: Concrete compressive strength data 분석 결과 RMSE 비교 그림.



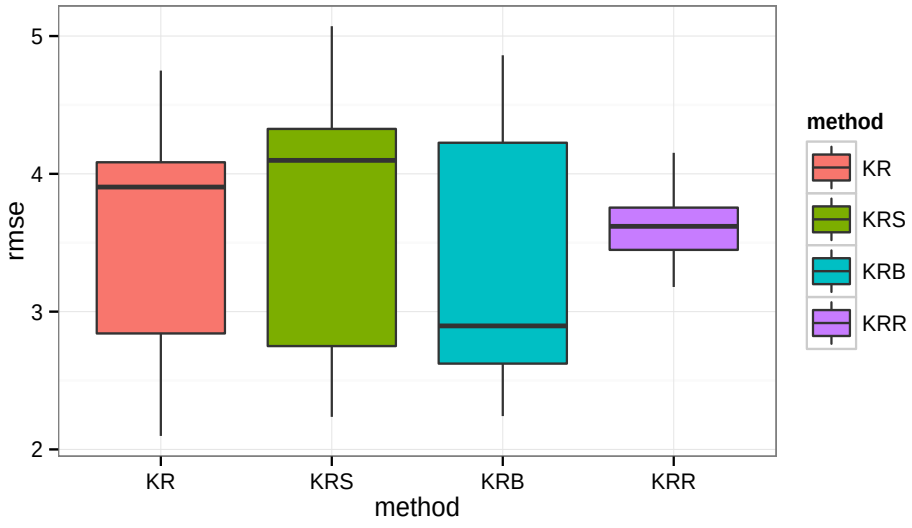
Concrete compressive strength data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 또한, KRR의 박스의 크기도 가장 작다. 따라서 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge regression)이 예측력과 신뢰성 측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

- Energy efficiency data

이 자료는 다양한 형태의 건물 구조에 따른 에너지 효율성에 관한 데이터

로 총 768(= N)개의 관측치를 가지고 있고, 2개의 반응변수와 8(= p)개의 설명변수로 구성되어 있다. 2개의 연속형 반응변수 중 시원함 측면의 효율성을 반응변수로 선택하였다. 건물의 구조를 나타내는 8개의 설명변수는 1개의 범주형 변수와 7개의 연속형 변수가 있는데 모두 연속형 변수로 취급하여 분석하였다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.75(n/N)인 576개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.125(= p^{-1})이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 9: Energy efficiency data 분석 결과 RMSE 비교 그림.

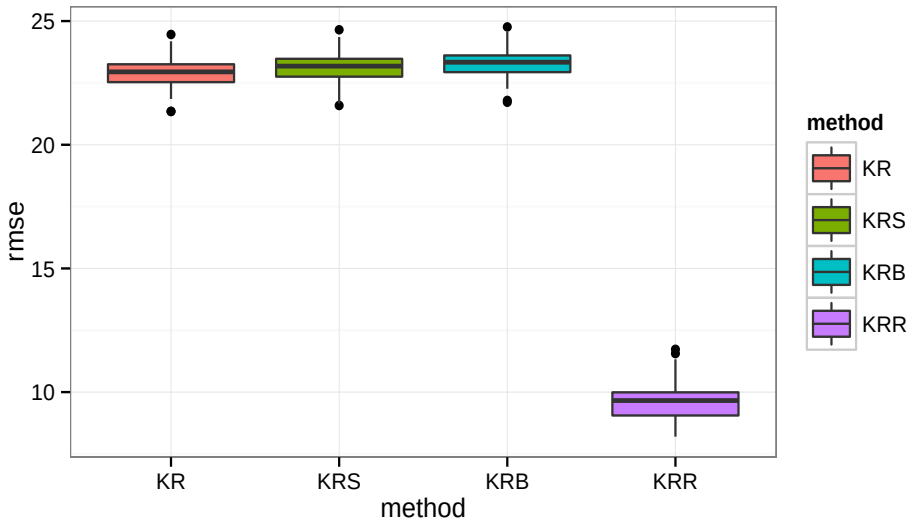


Energy efficiency data의 실증분석 결과 KRB에서 RMSE의 평균값이 가장 작음을 알 수 있다. 하지만 RMSE의 분산을 나타내는 박스의 크기는 KRR가 가장 작다. 실제 자료를 정확하게 예측하는 측면에서는 배깅(bagging)기법을 이용한 커널능형회귀법(kernel ridge regression)이 좋지만, 예측치의 신뢰성 측면에서는 랜덤포레스트(random forests)를 이용한 커널능형회귀법이 더 우수하다. 따라서 기존에 활용되는 방법보다 본 논문에서 제안하는 앙상블 기법(ensemble method)를 이용한 커널능형회귀법이 더 우수함을 입증하였다.

- Housing data

이 자료는 Boston 지역의 집값에 관한 데이터로 총 506($= N$)개의 관측치를 가지고 있고, 1개의 반응변수와 13($= p$)개의 설명변수로 구성되어 있다. 반응변수는 집값을 의미하고 연속형 변수이다. 설명변수는 집값에 영향을 미치는 요소로 1개의 범주형 변수와 12개의 연속형 변수로 이루어져 있다. 범주형 변수는 연속형 변수로 간주하여 분석에 사용하였다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.75(n/N)인 380개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.076($= p^{-1}$)이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 10: Housing data 분석 결과 RMSE 비교 그림.



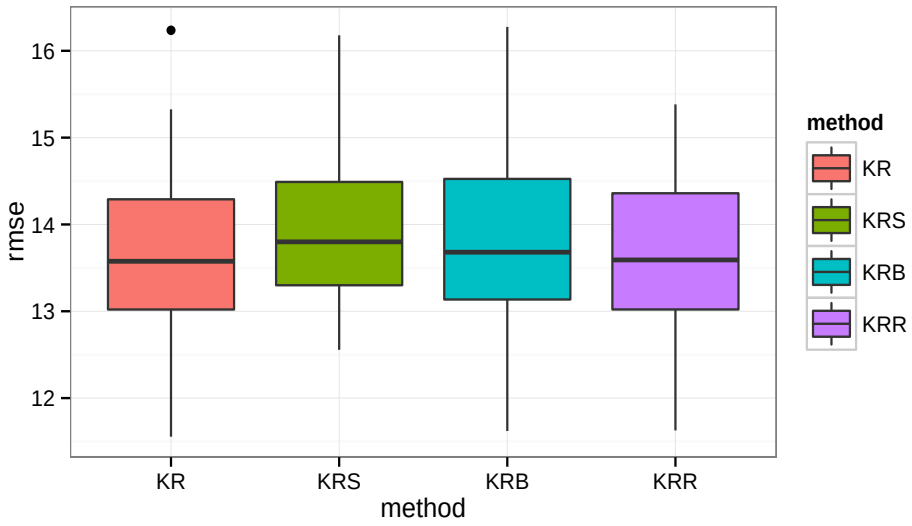
Housing data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 하지만 RMSE의 분산을 나타내는 박스의 크기는 4가지 방법 모두 비슷해 보인다. 이것은 최적의 능형 모수 λ를 선택하는 과정에서 1개 또는 2개의 방법에서 너무 작은 λ가 선택되어 모델에서 penalty가 의미 없는 방법이 존재하여 최적의 λ를 적용하지 못해 동등한 조건에서 비교 분석하지 못했다. 그런데도 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge regression)이 예측력

측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

- Yacht hydrodynamics data

이 자료는 요트의 유체역학에 관한 데이터로 총 308($= N$)개의 관측치를 가지고 있고, 1개의 반응변수와 6($= p$)개의 설명변수로 구성되어 있다. 반응변수는 요트에 미치는 저항력이고 연속형 변수이다. 설명변수는 요트 선체의 기하학적 요소로 이루어져 있고 모두 연속형 변수이다. KRS에서 모델 적합에 사용한 Sub-sample의 개수는 전체 데이터의 0.75(n/N)인 231개이다. KR, KRS, KRB, KRR에서 사용된 가우시안 커널(gaussian kernel)의 parameter σ는 0.17($= p^{-1}$)이다. 다음은 100번 반복하여 얻은 RMSE의 박스 그림이다.

그림 11: Yacht hydrodynamics data 분석 결과 RMSE 비교 그림.



Yacht hydrodynamics data의 실증분석 결과 KRR에서 RMSE의 평균값이 가장 작음을 알 수 있다. 하지만 RMSE의 분산을 나타내는 박스의 크기는 4가지 방법 모두 비슷해 보인다. 이것은 최적의 능형 모수 λ를 선택하는 과정에서 1개 또는 2개의 방법에서 너무 작은 λ가 선택되어 모델에서 penalty가 의미 없는 방법이 존재하여 최적의 λ를 적용하지 못해 동등한 조건에서 비교분석하지 못했다. 그런데도 기존에 활용되는 KR, KRS보다 본 논문에서 제안하는 랜덤포레스트(random forests)를 이용한 커널능형회귀법(kernel ridge

regression)이 예측력 측면에서 더 우수한 성능을 지니고 있음을 입증하였다.

다음은 7개의 실제 자료를 통해 얻은 RMSE의 평균과 표준편차를 나타낸 표이다.

표 5: 실제 자료의 분석 결과 RMSE 비교 표.

		KR	KRS	KRB	KRR
Airfoil	mean	47.7373	68.1578	55.9388	5.8809
	sd	2.2111	1.7915	1.7501	0.2504
Auto	mean	24.3895	24.4389	24.4697	13.1233
	sd	0.4756	0.4781	0.4665	1.7227
Computer	mean	184.1630	184.1874	184.1827	145.7496
	sd	27.6172	27.6127	27.6147	35.0485
Concrete	mean	35.0953	36.6955	35.7132	16.8888
	sd	0.6918	0.7070	0.6654	0.7328
Energy	mean	3.5976	3.6818	3.3003	3.6047
	sd	0.7472	0.8300	0.7872	0.1993
Housing	mean	22.9011	23.1445	23.2945	9.5934
	sd	0.5927	0.5780	0.5763	0.7361
Yacht	mean	13.6947	13.9310	13.7989	13.6649
	sd	0.8815	0.7634	0.8759	0.8175

대체로 모든 실제 자료에서 기존에 활용되고 있는 방법인 KR, KRS보다 KRB, KRR에서 RMSE의 평균값과 분산이 더 낮음을 알 수 있다. 따라서 본 논문에서 제안하는 앙상블 기법(ensemble method)를 이용한 커널능형회귀법(kernel ridge regression)이 예측력과 신뢰성 측면에서 더 우수한 성능을 지니고 있다고 할 수 있다.

다음은 자료의 정보를 나타낸 표이다.

표 6: 실제 자료에 관한 정보 표.

	N	p	σ	비고
Airfoil	1503	5	0.2	• 연속형 설명변수 5개 • Sub sample : 752개 • 결측치 없음
Auto	398	7	0.15	• 연속형 설명변수 4개, 범주형 설명변수 3개 • 반응변수 1개 제거 • Sub sample : 299개 • 결측치 없음
Computer	209	6	0.17	• 연속형 설명변수 6개 • 문자형 변수 2개, 반응변수 1개 제거 • Sub sample : 279개 • 결측치 없음
Concrete	103	8	0.125	• 연속형 설명변수 8개 • Sub sample : 78개 • 결측치 없음
Energy	768	8	0.125	• 연속형 설명변수 6개, 범주형 설명변수 2개 • 반응변수 1개 제거 • Sub sample : 576개 • 결측치 10개 제거
Housing	506	13	0.076	• 연속형 설명변수 12개, 범주형 설명변수 1개 • Sub sample : 380개 • 결측치 없음
Yacht	308	6	0.17	• 연속형 설명변수 6개 • Sub sample : 231개 • 결측치 없음

6 결론

본 연구를 통해, 앙상블 기법을 이용한 커널능형회귀법이라는 새로운 회귀분석 방법을 제시하였다. 일반적으로 자료가 복잡한 비선형 구조를 가지고 있는 경우 해석과 넓은 분석 활용 범위 등 다양한 측면을 고려하여 단순한 선형

구조로 변환하기를 원할 때 커널회귀법을 사용한다. 하지만 자료의 차원이 커질수록 다중공선성 문제로 추정치의 분산이 커져 왜곡된 예측 결과를 가져온다. 변수 간의 강한 상관성으로 발생하는 다중공선성 문제를 해결하기 위해 추정치의 분산을 안정시키고, 변수선택법으로 차원을 축소하는 목적으로 배깅(bagging)과 랜덤포레스트(random forests) 기법을 적용한 커널능형회귀법을 제시하였다. 제시한 방법론을 다양한 상황을 가정하여 모의실험을 진행하였고 그 성능을 입증하였다. 또한, 실제 자료를 이용하여 실증적 실험을 통해 본 논문에서 제시하는 방법론이 기존 방법론보다 우수함을 입증하였다.

참고문헌

강기훈 (2009). 통계학 개론(엑셀을 이용한 실습). 자유 아카데미.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Element of Statistical Learning*. Springer.

Huh, M (2015). Kernel-trick regression and classification. *Communications for Statistical Applications and Methods* **22**, 201–207.

Schölkopf, B., Smola, A. J. (2002). *Learning with Kernels*. MIT Press.