大邱曉星가톨릭大學校 基礎科學硏究論集 11. 2, 115~133, 1997
J. Basic Sci. Res. Inst., C.U.T.H. 11. 2, 115~133, 1997

신경망을 이용한 비모수 회귀분석에 관한 소고

황 창 하 · 유 지 영

대구효성가톨릭대학교 자연대학 정보통계학과

On Nonparametric Regression Method Using Neural Networks

Changha Hwang · Jiyoung Yu

Department of Statistical Information, College of Natural Sciences Catholic University of Taegu-Hyosung

ABSTRACT - This paper surveys the context of regression by feed-forward neural network approaches, and compares it with kernel, spline and projection pursuit method. For the comparison study we use the famous French curve as the main example.

1. 서 론

일반적으로 가장 널리 사용되는 통계적 방법은 회귀분석이다. 회귀분석은 종속변수 y와 독립변수 벡터 x의 상호 관련성인 회귀함수를 구하는 통계적 방법이다. 일반적으로 모수적 방법은 자료에 내재하고 있는 오차들이 특정한 형태의 분포를 가진다고 가정하는 반면에 비모수적 방법은 오차들에 대해 특정한 분포를 가정하지 않고 거의 모든 분포에 대해서 자료를 합리적으로 잘 적합할 수 있는 방법이다. 한편, 비모수적 방법은 전적으로 자료에 의해 결정되어지는 분포와 회귀함수를 전제로 하므로 모분포의 형태에 덜 의존한다는 장점이 있지만 모형의 모수의 수가 자료의 크기에 비례한다는 문제점이 있다.

컴퓨터의 급속한 발전으로 인하여 다양한 비모수적 방법들이 제안되고 있다. 비모수적 회귀분석 방법으로는 커널 방법, 스플라인 방법, projection pursuit(PP) 방법 등 여러가지가 있는데, 커널이나 스플라인 방법은 독립변수가 여러 개인 경우에는 차원문제(curse of dimensionality)로 자료의 개수가 적으면 고차원 공간에서 회귀함수를 잘 추정하지 못한다. 한편, 투영(projection)을 이용한 회귀분석 방법은 차원문제를 극복하기 때문에 고차원 공간에서 회귀함수를 추정하기 위해서는 투영방법이 효적인 방법이다. 신경망을 이용한 회귀분석 방법과 PPR(projection pursuit regression)이 이에 속한다. 그런데 3층 전방향 신경망을 이용한 회귀분석 방법은 PPR과 수학적으로 매우 비슷하여 자주 비교, 분석된다.

비모수적 또는 준모수적(semiparametric) 방법으로 분류되어지는 신경망은 좀더 유용한 모형을 만들기 위하여 적응모수(adaptive parameter)의 개수를 증가시킬 수 있는 일반적 함수형태를 취하며 모수의 개수는 자료의 크기와는 독립적으로 변화를 줄 수 있으므로 모수적 방법과 비모수적 방법의 장점들을 동시에 가지고 있다. 공학의 신호처리, 로보틱스, 콘트롤, 문자인식, 컴퓨터 그래

To whom all correspondence should be addressed
TEL. (053)850-3546

픽 등 여러분야에 이용되고 있으며 신경계가 어떻게 정보를 저장하고 다루는가에 관한 지식을 이용하려는 시도로부터 발전되었다. 비교적 간단하면서도 많은 노드들이 상호 연결된 망으로 구성되어 있으며 함수근사를 구현하기 위한 좋은 도구로 사용될 수 있다. 여러 가지 단점을 가지는 일반적인 신경망을 보완하여 이상치에 덜 민감한 로버스트 신경망 알고리즘이 제안되어졌으며, 또한 베이지안 이론을 바탕으로한 적합과 예측을 모두 잘할 수 있는 신경망 알고리즘이 제안되어졌다.

본 논문에서는 커널, 스플라인, PP 및 신경망을 이용한 회귀함수의 추정방법을 이론적으로 비교분석하고, 독립변수가 여러개인 경우에 좋은 추정결과를 보여진다고 이미 알려진 PPR과 신경망 회귀분석 방법을 독립변수가 하나인 경우에 커널 및 스플라인을 이용한 회귀분석 방법과 모의실험을 통하여 비교 분석한다. 그리고 베이지안 기법을 이용한 신경망과 이상치에 덜 민감한 로버스트 신경망을 이용한 회귀분석법을 일반적인 신경망을 이용한 회귀분석법과 비교 분석한다. 모의실험에 사용되는 회귀함수는 Wahba & Wold(1975)에 의해 제안된 French curve이며, 오차의 분포로는 정규분포, 오염된 정규분포, t_4분포를 사용한다. PPR에 대해서는 차원(dimension) M을 변화시켜보고, 신경망은 은닉층의 노드수를 8개로 고정시켜 회귀함수를 추정하는데 사용되었다. 한편, 커널은 평할계수의 크기를 변화시켜 회귀함수 추정을 하였다.

2. 비모수적 회귀분석

2.1. 커널 방법

커널 방법은 국부적 가중합을 연결시킨 것이다. 자료 (x_i, y_i), $i=1,2,\cdots,n$ 가 주어졌을 때 커널함수에 의해 추정된 회귀함수는 다음과 같이 표현된다.

$$\hat{y}_i = \frac{\sum_{j=1}^{n} K(\frac{x_i - x_j}{b}) y_j}{\sum_{j=1}^{n} K(\frac{x_i - x_j}{b})}, \quad i=1,2,\ldots,n$$

여기서 b는 평할계수, K는 커널 함수이다. 한편 커널함수는 다음과 같은 성질을 갖는다.

(1)모든 t에 대하여 $K(t) \geq 0$ 성립한다.

(2) $\int_{-\infty}^{\infty} K(t)dt = 1$

(3)모든 t에 대하여 $K(-t) = K(t)$이 성립한다.

여기서 (1), (2)는 확률밀도함수 특성이다.

커널의 모양은 회귀함수추정에 큰 영향을 주지는 않지만 다음과 같은 여러 종류의 커널함수가 있다.

(1) box

$$K_{box}(t) = \begin{cases} 1, & |t| \leq 0.5 \\ 0, & |t| > 0.5 \end{cases}$$

(2) Triangle

$$K_{tri}(t) = K_{box}(t) * K_{box}(t)$$

(3) Parzen

$$K_{par}(t) = K_{tri}(t) * K_{box}(t)$$

(4) Normal

$$K_{nor}(t) = \frac{1}{\sqrt{2\pi}\,0.37} e^{\frac{-t^2}{2*(0.37)^2}}$$

y_i변수는 x_i가 x_j에 근접해 있을 때 큰 가중치를 갖게되고, x_j로부터 멀리 떨어져 있으면 작거나 0의 가중치를 갖는다. 평할계수 b는 $K(t/b)$의 폭을 정하는 중요한 역할을 하며 x_i 주변영역 크기를 결정하여 y_i가 상대적으로 큰 가중치를 갖도록 한다. $\hat{y}_i = \sum_{j=1}^{n} \omega_{ij} y_j$이며,여기서 $\omega_{ij} = K(\frac{x_i - x_j}{b}) / \sum_{j=1}^{n} K(\frac{x_i - x_j}{b})$이고

$\sum_{j=1}^{n} \omega_{ij} = 1$이 된다. 따라서 추정량 $\hat{y_i}$는 y_i의 국부적 가중합이 된다.

이 방법은 관측자료의 수가 커질수록 커널함수와 그에 따른 모수의 개수가 많아지므로 계산이 복잡해지며, 회귀함수를 잘 추정하기 위해서는 자료의 개수가 매우 많아야한다는 단점이 있다. 즉, 이 방법은 차원문제를 극복하지 못하므로 자료가 희소한 고차원에서는 회귀함수를 잘 추정하지 못한다. 그리고 자료의 국부적인 정보만으로 회귀함수를 추정하고, 평활계수를 자료로부터 직접 추정하는데는 다소 어려움이 있다는 단점이 있다. 아울러 이 방법은 이상치에 로버스트하지 못하다. 자세한 내용을 위해서는 Simonoff(1996)를 참고하라.

2.2. 스플라인 방법

스플라인은 제도가들이 주어진 자료의 점들을 지나는 매끄러운 곡선을 얻기위해 유연한 나무로 만든 나무자 또는 탄성있는 철재자에 의해 만들어지는 곡선을 이용한데서 유래되었다. 스플라인 함수는 임의의 구간에서 연속조건을 만족하는 부분다항식(Piecewise Polynomial)을 연결시킨 것이다.

자료들이 미지의 함수값을 나타낼 경우 미지의 함수를 추정하여 모든 점들을 통과하도록 만든 곡선을 보간 스플라인(interpolating spline)이라 하며, 오차를 수반한 자료에서 잔차제곱합과 곡률(curvature)의 합이 최소가 되도록 함수를 추정한 곡선을 평활스플라인(smoothing spline)이라 한다. 스플라인 함수에는 여러 가지 차수가 고려되지만 1차, 2차 스플라인 함수는 불연속성의 가능성이 있으며, 4차 이상은 계산이 복잡해지는 등의 여러 가지 이유로 3차 스플라인 함수를 많이 사용한다.

3차 스플라인은 어떤 구간 $[a, b]$에서 독립변수의 자료 $x_1, \cdots, x_n$ 이 $a \le x_1 \le \cdots \le x_n \le b$로 순서화된 마디(knot)로 가정이 될 때 각 구간 $(a, x_1), (x_1, x_2), \cdots, (x_n, b)$ 에서 함수 f는 3차 다항식이고, 연속이며, 각 구간의 경계점인 x_i에서의 함수값과 일차, 이차 미분값이 연속이고 동일하다는 성질을 만족한다. 구간 $[a, b]$의 $a \le x_1 \le \cdots \le x_n \le b$에서 관측값 $y_1, \cdots, y_n$이 주어질 때 평활 스플라인 함수는 다음과 같다.

$$S(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx$$

여기서 λ는 임의의 고정된 상수로써 평활계수(smoothing parameter)의 역할을 하게되며 이 때 $\lambda = 0$이면 보간함수 형태가 된다. 적합의 정도는 λ에 의해 통제된다.

이 방법은 마디의 수가 증가할수록 선형결합에 필요한 항의 수가 증가하며, 마디와 기저함수를 자료로부터 직접 추정할 수 없고, 경험적으로 선택해야 한다는 단점이 있다. 또 차원문제를 극복하지 못하므로 자료가 희소한 고차원에서는 회귀함수를 잘 추정하지 못한다.

2.3.Projection Pursuit Regression

PPR은 차원문제를 극복하기 위해서 Friedman & Stuetzle(1981)에 의해 제안되어졌으며 투영된 입력벡터의 비선형 함수들의 선형결합으로 표현된다. PPR의 기본 개념은 다음과 같다.

$x = (x_1, x_2, \cdots, x_p)^T$ 를 설명벡터, y를 반응변수라 할 때 이들의 관측값은 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$, y_i, $i = 1, \cdots, N$로 표현된다. PPR은, $a_1, a_2, \cdots$를 설명벡터 x가 투영되는 p차원 단위벡터, 즉 방향벡터이고 $\bar{y} = \frac{1}{N} \sum_{m=1}^{N} y_m$ 일 때, 다음의 관계식

$$y \approx \bar{y} + \sum_{m=1}^{M_0} \beta_m \phi_m (a_m x)$$

를 만족하는 M_0, $a_1, a_2, \cdots, a_{M_0}$ 및 $\phi_1, \phi_2,$

$\cdots, \phi_{M_0}$를 찾는 방법이다. 즉, PPR은 투영된 설명벡터의 비선형 함수들의 선형결합이다. 이때 β_m은 가중치이며, ϕ_m은 비선형 변환함수이다. 그리고 β_m, ϕ_m 및 a_i는 자료로부터 회귀함수를 최적으로 추정할 때의 것들이다. M_0가 충분히 크면 임의의 연속함수는 PPR에 의해 원하는 정확도로 근사될 수 있다고 Diaconis & Shahshahani(1984)에 의해 밝혀졌다.

PPR에서 "projection"은 x벡터를 투영의 길이가 $a_m^T x$인 방향벡터 $a_1, \cdots, a_n$ 상으로 투영시키는 것을 의미하며, "pursuit"은 "좋은" 방향벡터 $a_1, \cdots, a_n$를 찾는데 사용하는 최적화기법을 의미한다.

이 방법은 차원 M_0를 사용자가 선택해야 하며, ϕ_m이 비선형 함수형태이므로 추정된 회귀식에 대한 해석이 어렵다는 단점이 있다.

3. 신경망을 이용한 회귀분석

3.1. 다층 전방향 신경망의 구조

다층 전방향 신경망은 퍼셉트론과 같이 하나의 조정층(single-adjustable layer)으로 구성되는 모형들의 한계점 때문에 입력층과 출력층 사이에 한 개 이상의 은닉층(hidden layer)이 존재하는 새로운 모형으로서 1980년대 중반에 제안되었으며 이 신경망의 학습알고리즘으로서는 역전파 알고리즘(backpropagation algorithm)이 주로 사용된다.

망구조는 입력층, 은닉층, 출력층 방향으로 연결되어 있으며, 각 층내의 연결과 출력층에서 입력층으로의 직접적인 연결은 존재하지 않는 전방향(feedforward)네트워크이다. 학습은 입력층의 각 노드에 입력 데이터를 제시하면 이 신호는 각 노드에서 변환되어 중간층에 전달되고 최종적으로 출력층으로 나오게 된다. 이 출력값과 목표값을 비교하여 그 차이를 감소시키는 방향으로 연결강도를 조절하는 것이다.

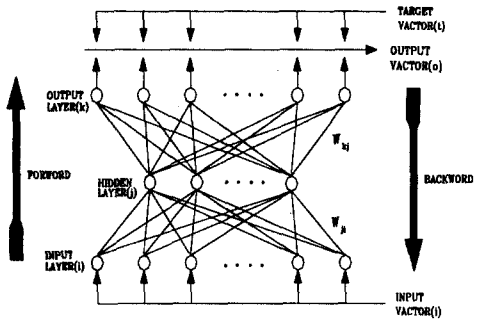


그림1. 다층 전방향 신경망 구조

일반적으로 하나이상의 은닉층을 가질 수 있고, 입력층에서 출력층으로 연결된 "skip-layer"를 가질 수 있다. 만약 모든 노드들이 동일한 활성함수 f_h나 f_0 를 갖는다면 신경망에 의해 추정된 함수는 다음과 같은 식으로 표현된다.

$$y_k = f_0\Big(a_k + \sum_{i \to k} w_{ki} x_i + \sum_{j \to k} w_{kj} f_h\Big(a_j + \sum_{i \to j} w_{ji} x_i\Big)\Big)$$

여기서 w는 상위층의 각노드들과의 연결강도, 즉 가중치이다. 편의항을 a로 두며, 이를 0으로 두면 모든 노드와의 연결이 끊어지므로 삭제가 된는 반면 +1일 경우는 모든 노드와 연결이 이루어진다. 은닉층에서 취해지는 활성함수 f_j로 선형함수, 로지스틱함수($e^x/(1+e^x)$)가 일반적으로 사용된다.

신경망은 적은 수의 기저함수로 미지의 다변량 함수를 추정할 수 있으며 가중치와 각 기저함수 내에 있는 곱의 항의 수는 자료에 최적으로 적합되도록 신경망 학습 알고리즘을 거쳐 적응적으로 결정된다. 한 개의 은닉층을 가진 신경망과 세 개의 은닉층을 가진 신경망은 모두 임의의 연속함수를 원하는 정확도로 근사시킬 수 있다고 밝혀졌고, 후자는 더 유동적(flexible)이다. 또한 신경

망은 국부대 기저함수와 전역대의 기저함수의 보충적 특징을 가지며, 모수들이 적응적으로 결정이 되는 장점이 있다. 그러나 이 방법은 "black-box" 방법이어서 해석상의 어려움은 여전히 있으며, 은닉 노드의 개수를 결정해야 하는 문제가 있다. 은닉 노드 개수의 결정에 관한 자세한 내용을 위해서 Hwang & Kim(1997)을 참고하라.

3.2. 역전파 알고리즘

다층 전방향 신경망을 학습시키기 위한 학습 알고리즘으로 역전파 알고리즘이 널리 이용되어져 왔다. 일반적으로 역전파 알고리즘은 최급강하법을 기본으로 한 매우 유용한 회귀함수추정 방법이다.

역전파 알고리즘의 기본원리는 입력층의 각 노드에 입력패턴을 주면 이 신호는 각 노드에서 변환되어 은닉층에 전달되고 최후에 출력층에서 값을 출력하게 된다. 이 출력값과 기대값을 비교하여 차이를 줄여나가는 방향으로 연결강도를 조절하고, 상위층에서 역전파하여 하위층에서는 이를 근거로 다시 자기층의 연결강도를 조정해 나간다. 지도학습에서는 입력 및 원하는 출력(목표출력) 패턴(벡터)이 신경망에 제시된다. 신경망은 입력층에 주어진 입력패턴이 출력층에 전파되면서 변한 출력패턴을 목표패턴과 비교한다. 신경망에서 출력된 패턴이 목표패턴과 일치하는 경우에는 학습이 일어나지 않는다. 그렇지 않은 경우는 얻어진 출력패턴과 목표패턴의 차이를 감소시키는 방향으로 신경망의 연결강도를 조절하여 학습한다. 신경망에 은닉 노드가 없는 경우에는 델타규칙과 동일하다. w_{ji}는 입력층에서 은닉층으로의 가중치, θ_j는 은닉층 노드의 가상노드를, v_{kj}는 은닉층에서 출력층으로의 가중치를, θ_k는 출력층 노드의 가상노드를, x_{pi}는 입력층의 i번째 노드의 p번째 입력값을, h_{pj}는 은닉층의 j번째 노드에서 p번째 입력벡터의 출력을, $\hat{y}_{pk}$는 출력층의 k번째 노드에서 p번째 입력벡터의 출력을, a_{pj}는 입력층에서 은닉층으로의 가중치와 입력층의 출력값과의 곱에 대한 합을, b_{pk}는 은닉층에서 출력층으로의 가중치와 은닉층의 출력값과의 곱에 대한 합을, g_j는 은닉층의 시그모이드 비선형 활성함수를, g_k는 항등함수로 출력층 활성함수를 나타낸다면 $a_{pj} = \sum_i \omega_{ji} x_{pi}$, $h_{pj} = g_j(a_{pj})$, $b_{pk} = \sum_j v_{kj} h_{pj}$, $\hat{y}_{pk} = g_k(b_{pk})$로 표현할 수 있다. 가중치의 변화량에 대해서는 다음과 같이 정리된다.

$$\begin{aligned} \Delta v_{kj} &= \alpha(y_{pk} - \hat{y}_{pk}) h_{pj} \\ &= \alpha \delta_{pk} h_{pj} \\ \Delta \theta_k &= \beta \delta_{pk} \end{aligned}$$

여기서 α, β는 학습률이다.

3.3. 로버스트 신경망

일반적인 역전파 알고리즘이 고전적인 함수근사 방법에서의 문제점을 해결하는데는 탁월한 성능을 발휘하는 것은 사실이지만, 일반적인 역전파 알고리즘의 문제점은 최급하강법 자체가 가장 낮은 골짜기를 목표로 하는 방법이 아니라 지금 있는 점에서 보아 가장 급경사면을 따라 오차를 조절해가므로 지역극소에 빠질 수 있고, 훈련시간이 길며 이상치에 민감하다는 등의 단점을 들수가 있다. 일반적으로 회귀분석을 위한 자료는 오차를 수반하고 많은 경우에 이상치 또는 이상치로 의심되는 관측치가 포함된다. 따라서 과대오차에 민감하지 않고, 이상치의 영향을 최소화시키는 로버스트 역전파 알고리즘이 제안되어져 왔다.

다음의 내용은 황창하, 김상민, 박희주(1997)에서 인용했다. 로버스트 역전파 알고리즘을 유도하기 위해 통계물리에서는 다음과 같은 일반화된 에너지함수를 사용한다.

$$E(V, W) = \sum_{p=1}^{P} \sum_{j=1}^{m_p} V_p z(y_{pj}, \hat{y}_{pj}) + E_{prior}(V)$$

여기서 $z(y_{pj}, \hat{y}_{pj}) = \frac{1}{2}(y_{pj} - \hat{y}_{pj})^2$이고, V_p는 0또는 1의 값을 가지는 확률변수 $\{V = V_p, p = 1, \cdots, P\}$이다. 즉, 입력자료가 이상치이면 0, 아니면 1의 값을 갖는다. 한편 $E_{prior}(V)$는 V_p의 사전분포에 의해 공헌되어진 에너지의 양을 나타내며, 이것의 일반적인 선택은 다음과 같다.

$$E_{prior}(V) = \eta \sum_{p=1}^{P} (1 - V_p).$$

이것은 $z(y_{pj}, \hat{y}_{pj}) < \eta$ 이면 $V_p = 1$이 되어 주어진 관측치가 표본으로 간주되고, 그렇지 않으면 $V_p = 0$이 되어 이상치로 간주된다.

한 편 η는 지정된 반복회수에 도달할 때마다 계산되는 우측경계값으로 정의한다. 즉, 지정된 반복수에 도달할 때마다 관측치들에 대응되는 $\sum_{j=1}^{m_p} z(y_{pj}, \hat{y}_{pj})$ 값들을 구한후에 정렬하여 삼사분위수(Q_3)와 일사분위수(Q_1)를 계산한다. 그리고 삼사분위수에서 일사분위수를 공제한 값인 사분위범위수(IQR)를 계산하여 우측경계값, $Q_3 + 1.5 * IQR$ 을 η값을 취한다. 목표는 V_p가 이진 값을 가진다는 제약조건하에서 $\{V_p\}$ 와 W에 관해서 $E(V, W)$ 를 최소화시키는 것이다. 그런데 이 문제는 연속형변수와 이산형변수가 혼합된 경우의 최적화 문제이기 때문에 해석적인 문제를 구할 수 없을 뿐아니라 최급하강법을 사용하여 해를 구하는 것도 쉽지않다. 따라서 이런 문제점을 해결하기 위하여 Gibbs분포를 사용하며, 그 분포는 $P[V, W] = \frac{1}{Ze^{-\beta E[V, W]}}$로 정의된다. 이때 Z는 관계식 $\sum_V \int_W P[V, W] = 1$ 을 만족한다. 따라서 $E(V, W)$를 최소화하는 문제는 $P[V, W]$를 최대화하는 문제로 귀착된다. 그러나 이것 또한 연속형변수와 이산형변수가 혼합된 경우의 최적화문제이기 때문에 어려움이 따른다. 따라서 이런 문제에 대한 하나의 해결책으로는 W의 주변분포 $P_{margin}[W]$를 구하여 이것을 최대화 시키는 것이다. 따라서 일반적인 역전파 알고리즘 중에서 delta를 조정하는 부분만을 수정하여 로버스트 역전파 알고리즘을 구현할 수 있다.

(1) 은닉층 노드의 경우

$$\Delta w_{ji} = \alpha \sum_k \delta_{pk} v_{kj} \hat{y}_x \equiv \alpha \delta_{pj} \hat{y}_x,$$
$$\Delta \theta_j = \beta \delta_{pj}$$

(2) 출력층 노드의 경우;

$$\Delta v_{kj} = \alpha \sum_{p=1}^{P} \frac{1}{1 + e^{(\beta(\sum_{j=1}^{m_p} \frac{1}{2}(y_{pk} - \hat{y}_{pk})^2 - \eta))}} * (y_{pk} - \hat{y}_{pk}) * h_{pj} \equiv \alpha \delta_{pk} h_{pj}$$
$$\Delta \theta_k = \beta \delta_{pk}$$

여기서 α, β는 학습률이다. 모수 β는 알고리즘 훈련과정에서 계산되는 학습률을 조절하는 역할을 한다.

3.4. 가중치 감소(Weight Decay)

베이지안 기법을 이용한 가중치 감소법은 과대적합을 피하기 위해 사용된다. N개의 입력패턴 x와 목표값 y로 구성된 관측자료의 집합 D가 주어질 때, 가중치 벡터 w의 분포함수 $p(w|D)$를 찾고자한다. 다음과 같은 베이지안 방법을 사용하여 가중치들의 사후분포를 찾을 수가 있다.

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

조건부 분포 $p(D|w)$는 가중치들의 함수로 간주될수 있으며 우도라고 한다. 전통적인 네트워크 훈련방법은 우도함수를 최대로하는 가중치벡터를

구하는 것이다. 일단 입력값들이 주어지면 베이즈 공식을 이용하여 사전분포를 사후분포로 변환할 수있다. 사후분포를 계산하려면 사전분포 $p(w)$와 우도함수 $p(D|w)$가 주어져야 한다.

w의 사전분포에 대한 간단한 선택중 하나는 평균벡터가 0인 다변량정규분포이며 다음과 같이 표현된다.

$$p(w) = \frac{1}{Z_W(\alpha)} \exp\left(-\frac{\alpha}{2} \| w \|^2\right)$$

여기서 정규화 인수(normalization factor) $Z_W(\alpha)$는 $Z_W(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{W/2}$로 주어진다. W는 가중치의 전체 개수를 의미하며 α는 적당한 상수이다.

다음은 우도함수의 선택에 관한 것으로서 주어지는 입력값들에 대한 목표값들의 분포는 다음과 같이 표현될 수 있다.

$$p(y|x, w) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}\{\hat{y}(x; w) - y\}^2\right)$$

$\hat{y}(x; w)$은 신경망에서 출력된 결과값이다. 분산은 모수 β^{-1}에 의해 통제되어지며 모수 β에 대해서는 알고 있다고 가정을 한다. 입력패턴들이 독립이라고 가정하면 우도함수는 다음과 같이 된다.

$$p(D \mid w) = \prod_{n=1}^{N} p(y_i \mid x_i, w) = \frac{1}{Z_D(\beta)} e^{\left(-\frac{\beta}{2}\sum_{n=1}^{N}(\hat{y}(x_i; w) - y_i)^2\right)}$$

이 때 정규화 인수는 $Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{N/2}$로 주어진다.

신경망에서의 주요 관심사는 새로운 입력값에 대해 출력값을 예측하는 것이다. 이 예측값은 가중치들을 적분함으로써 얻어질 수 있으므로

$$p(y \mid x, D) = \int p(y \mid x, w)\, p(w \mid D)\, dw$$

형태가 되며 사후분포 $p(w \mid D)$가 이를 최대로 하는 가중치벡터 w_{MP}에서 가파른 봉우리를 이루고 있다면 $p(y \mid x, w_{MP})$에 의한 적분값에 수렴할 수 있다. 실질적으로 이런 방법을 이용하려면 w_{MP}를 결정해야한다. 사후확률을 최대로 하는 값을 찾는 것 대신에 일반적으로 오차함수라 불리는 음의 로그함수를 최소화하는 값을 구하는 것이 편리하다. 음의 로그함수가 단조함수이므로 두방법의 결과는 동일하다. 특별한 사전분포와 우도함수에 대하여 무시해도 좋을만큼 작은 상수항이 더해진 음의 로그확률을 다음과 같이 정의할 수 있다.

$$E(w) = \frac{\beta}{2}\sum_{n=1}^{N}\{\hat{y}(x_i; w) - y_i\} + \frac{\alpha}{2}\|w\|^2$$

가중치 감소항을 갖는 이 방법은 가중치 w_{ij}의 평방합을 벌칙(penalty)으로 사용하여 각 단계에서 가중치의 크기를 축소시키고자 하는 것이며 다시 다음과 같이 정규화 형태로 표현할 수있다.

$$E + \lambda\sum_{ij} w_{ij}^2 = E + \lambda C$$

E는 모든 가중치들 오차평방합이다. 가중치 감소법을 사용할 때의 중요한 문제점은 모수 λ를 어떻게 선택해야 하는가이다. 만약 가중치들의 사전분포가 $p(w) \propto \exp -\lambda C(w)$의 형태를 취하면 위의 정규화 형태의 식을 최소화시키는 것은 가중치들의 사후분포를 최대화하는 문제로 귀착된다. 사전분포가 평균이 0이고, 분산이 $1/2\lambda$인 정규분포를 따른다면 입력값들의 로지스틱함수 적합은 ± 3주변이며, 전체 입력값의 표준오차는 2정도가 된다. 만일 입력값들이 [0, 1]사이에 있는 작은 값들이면 가중치들의 표준오차는 5정도를 제안하고 있으며 이 때 $\lambda = 1/50$ 이다. 이

런 논의가 다소 보수적이라는 견해도 있지만 기본적으로 $\lambda \approx 10^{-3} \sim 10^{-1}$를 제안하고 있다. 한편, Ripley는 $\lambda \approx 10^{-4} \sim 10^{-2}$를 제안하고 있다.

4. 모 의 실 험

지금까지 여러 가지 비모수 회귀분석 방법을 이론적으로 비교 분석하였다. 본 절에서는 Wahba & Wold (1975)가 제안한 French curve에 대하여 각 방법이 어떻게 회귀함수를 추정하는지 모의실험을 통하여 비교 분석한다. 모의실험에 사용된 오차항의 분포는 다음 세가지 종류이다.

1. 정규자료

Wahba & Wold(1975)에 의해 제안된 회귀함수

$$y_i = f_{\text{true}}(x_i) + \varepsilon_i, \quad i = 1, \cdots, n$$

여기서 $f_{true}(x) = 4.26(e^{-x} - 4e^{-2x} + 4e^{-3x})$이며, $\varepsilon_i \sim N(0, 0.2^2)$이다. 입력값 x는 실수구간상의 임의의 실수들로서 본 논문에서는 1과 3사이에서 랜덤으로 발생된 100개 값이다.

◆ S-PLUS program

```
> x <-seq( 0, 1, length = 100)
>e <-rnorm( 100, 0, 0.2)
>y <-
  4.26 * ( exp(-x) - 4 * exp(-2*x)
   + 4 * exp(-3*x)) + e
```

2. 오염된 자료

주어진 함수에 오차항을 $N(0, 0.2^2)$대신 $N(0, 9*(0.2)^2)$에서 10개를 생성하여 정규자료에 포함시킨다.

◆S-PLUS program

```
> x.rnd <- runif( 10, 1, 100)
> x.rand <- floor( x.rnd)
  [1] 51 44 65 3 ··· 50
> err.out <-
    rnorm( 10, mean = 0, sd = 2.25)
> y.out <-
  4.26*(exp(-x.rand) - 4* exp(-2*x.rand)
  +4* exp(-3*x.rand)) + err.out
```

3. 과대오차를 갖는 자료

오차항에 대하여 정규분포보다 꼬리가 두터운 t_4-분포로부터 생성한 자료를 적용시킨다.

◆S-PLUS program

```
> err.t <- rt(100, 4)
```

시뮬레이션은 S-PLUS와 MATLAB을 사용하여 구현되었다.

4.1.커널 방법

그림2, 그림3, 그림4는 커널 방법으로 정규자료, 오염된 자료, 과대오차를 갖는 자료에 대해 회귀함수를 추정한 결과이며, 국부적 평균곡선을 연결시킨 것이므로 대체로 추정함수곡선이 매끄럽지 못하다. 커널 방법은 평활계수를 0.05, 0.1, 0.5, 1로 변화시켜 회귀함수추정을 하였으며, 그림2, 그림3, 그림4에서 $b = 0.05$일 때 추정된 회귀함수들은 모두 모든 점들을 다 지나고 있다. 이 것은 잔차를 최소화하지만 새로운 입력자료에 대한 예측에는 좋은 결과를 주지 못하므로 좋은 추정이 아니다. 평활계수 b가 커질수록 추정된 회귀함수는 다소 매끄러워(smooth)지기는 하지만 여전히 자료의 국부적 정보만을 사용하므로 완전하게 매끄럽지는 못하다. 그림2의 $b = 0.5$인 경우와 그림3의 $b = 0.5$인 경우에 이상치에 민감하지 않고 적절한 회귀함수 추정을 보여주며, 그림4에서는 $b = 1$일 때 적절한 회귀함수 추정을 보여

준다. 여기서 어떻게 평활계수를 선택하는지에 관한 문제는 다루지 않고 단지 평활계수가 회귀함수의 모양에 결정적으로 영향을 미치고 커널 방법에 추정된 회귀함수는 자료의 국부적 정보를 이용하기 때문에 매끄럽지 못한 회귀함수를 추정하고 있다는 것을 지적한다. 나중에 신경망의 결과와 비교해 보면 차이점을 발견할 수 있을 것이다. 자세한 내용을 위해서는 Simonoff(1996)를 참고하라.

◆S-PLUS program

```
> k <- ksmooth(x, y, bandwidth = 0.05)
>win.graph()
>plot(x, y)
>lines(x, k$y, lty=2)
```

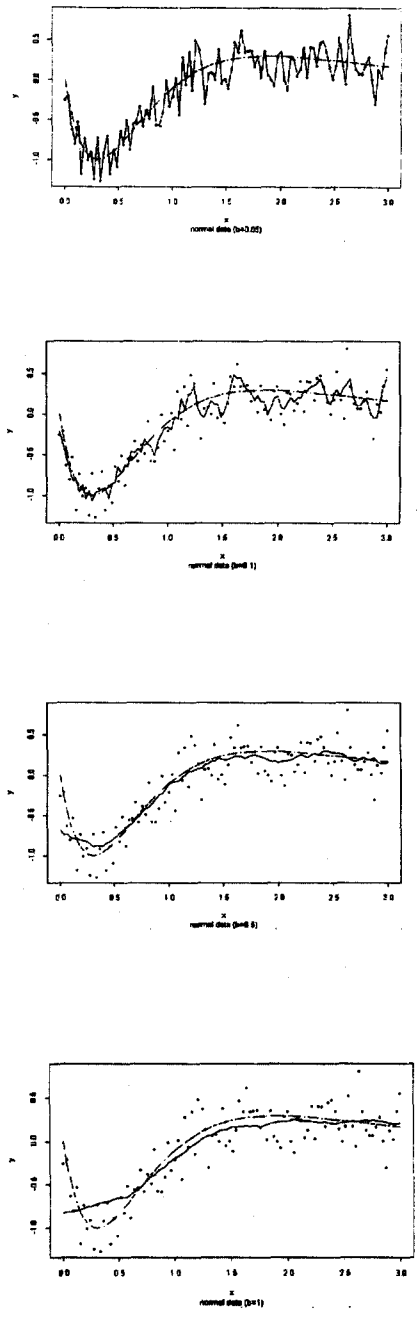


그림 2. 정규자료에 대한 커널 방법

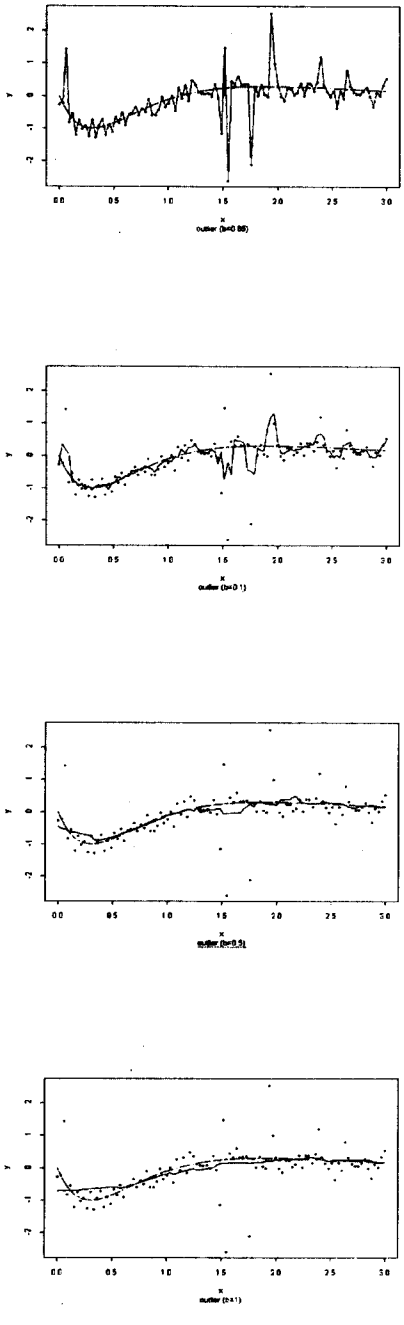


그림 3. 이상치 자료에 대한 커널 방법

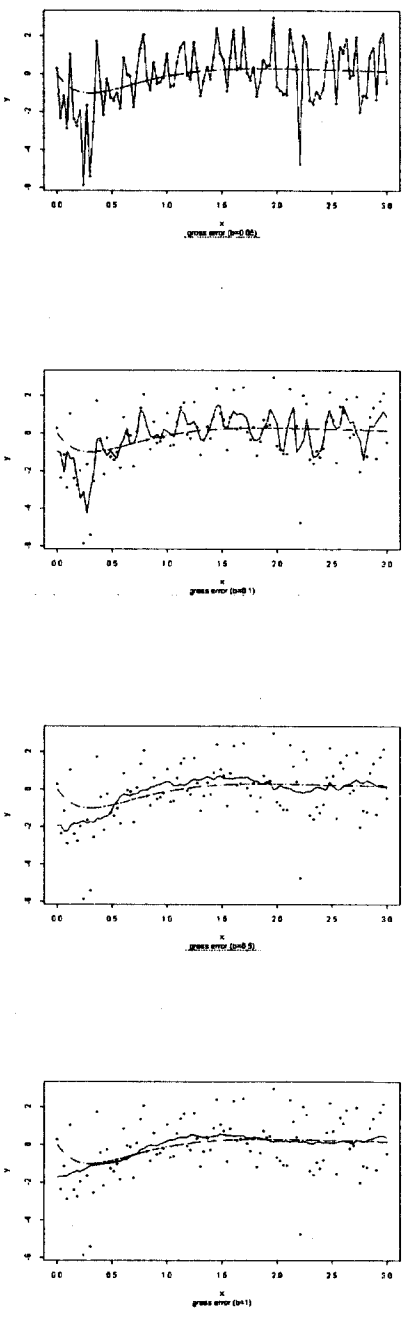


그림 4. 과대오차 자료에 대한 커널 방법

4.2. 스플라인 방법

그림5, 그림6, 그림7은 스플라인 방법으로 각각의 자료에 대해 회귀함수를 추정한 결과이며, 모수 λ를 0.5, 0.1, 0.01, 0.001로 변화시켜 함수를 추정하였다. 그림5에서는 $\lambda = 0.001$일 때 적절한 회귀함수추정을 하며, 그림6을 볼 때 적절한 λ를 선택하여 이상치에 영향을 덜 받는 회귀함수를 추정할 수 있음을 알 수 있다. 대체로 $\lambda = 0.001$일 때 회귀함수를 잘 추정하는 경향이 있다. 그림7에서는 $\lambda = 0.1$일 때 적절한 회귀함수추정을 한다.

◆S-PLUS program

```
>s <- smooth.spline(x, y, spar = 0.001)
>win.graph()
>plot(x, y)
>lines(x, s$y, lty=2)
```

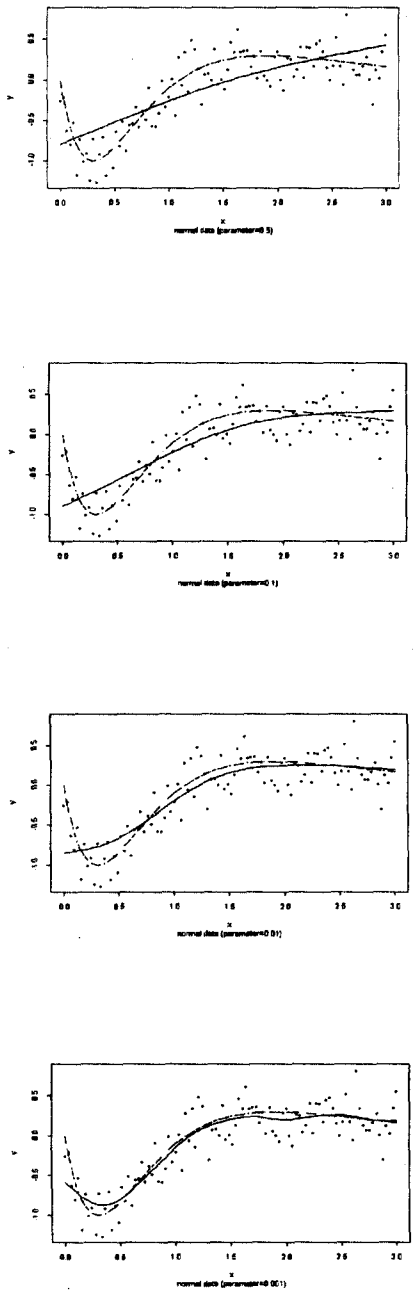
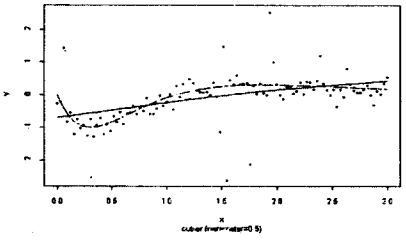
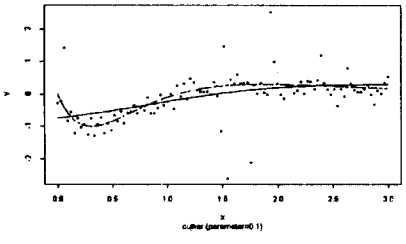
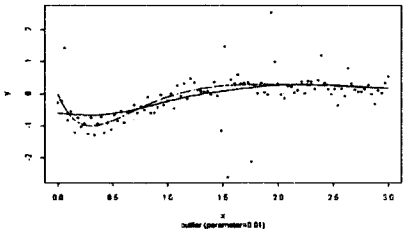


그림 5. 정규자료에 대한 스플라인 방법







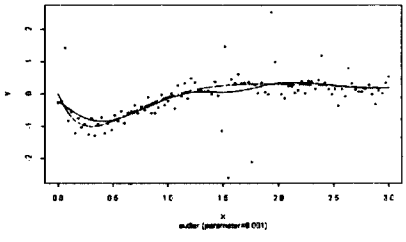
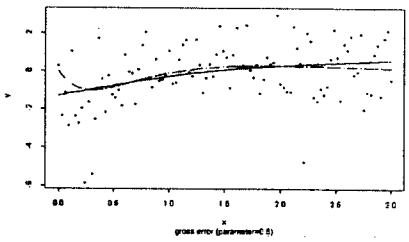
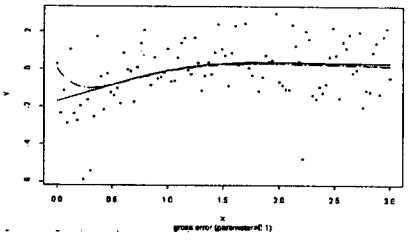
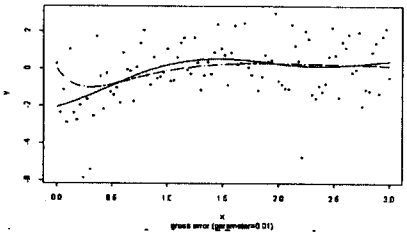


그림 6. 이상치 자료에 대한 스플라인 방법







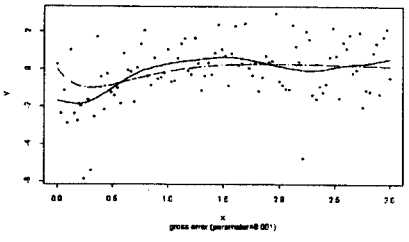


그림 7. 과대오차 자료에 대한 스플라인 방법

4.3. PPR 방법

그림8, 그림9, 그림10은 PPR 방법으로 각각의 자료에 대해 회귀함수를 추정한 결과이다. S-PLUS의 "ppreg()"함수에서는 차원 M_0를 인수 min.term과 max.term항을 두어 조절하고 있다. 따라서 min.term과 max.term항을 조절하여 차원 M_0를 변화시켰다. 그림8에서 보듯이 차원 M_0가 커져도 min.term=1인 경우와 유사한 회귀함수의 추정결과를 보여준다. 독립변수가 하나인 자료이므로 min.term=1인 경우가 적절하다. 그림9, 그림10에서도 동일하게 min.term=1인 경우가 적절하다. PPR에서는 M_0를 결정하는 것이 중요한데 통계적인 결정기준 보다는 다소 경험적인 방법을 사용한다. 원래 PPR은 앞에서도 언급한 바와 같이 독립변수가 여러개일 때 더욱 의미가 있는데 독립변수가 하나인 경우에도 적합을 잘 하는 것으로 보인다.

◆S-PLUS program

```
>p <-  ppreg(x, y, 1, 2, xpred = x)
>win.graph()
>plot(x, y)
>lines(x, p$ypred, lty=2)
```

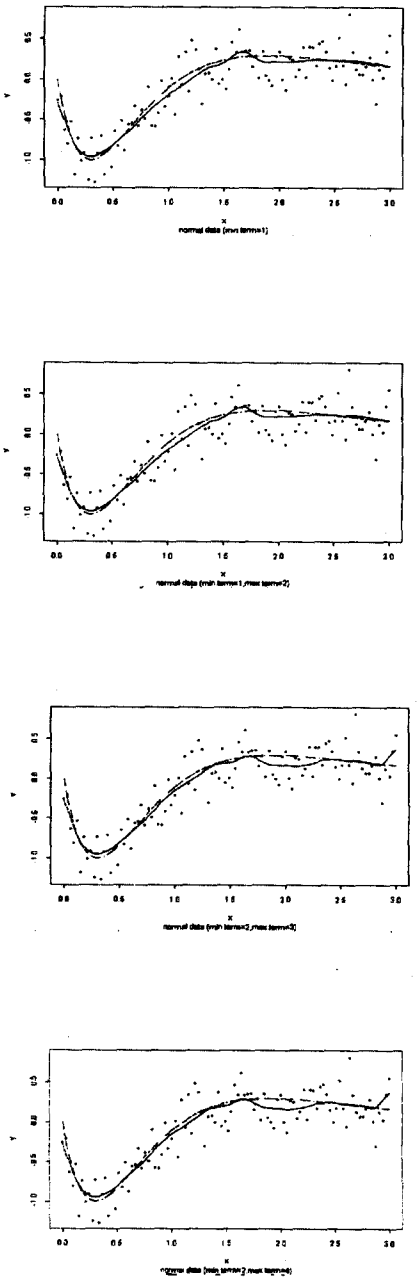
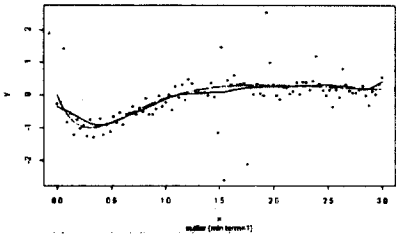
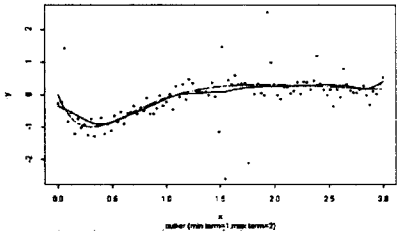
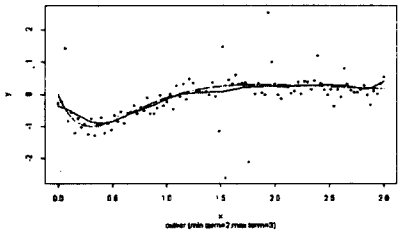


그림 8. 정규자료에 대한 PPR 방법







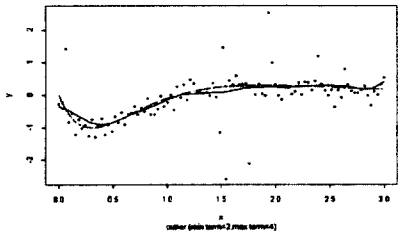
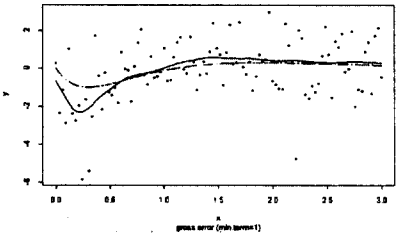
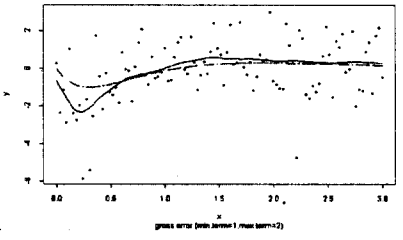
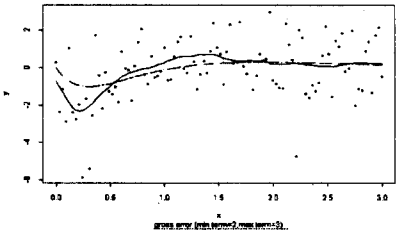


그림 9. 이상치 자료에 대한 PPR 방법







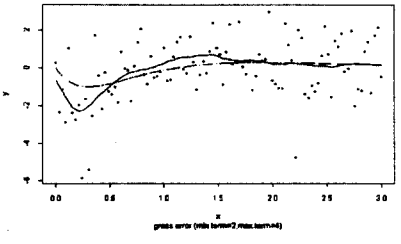


그림 10. 과대오차 자료에 대한 PPR 방법

4.4. 일반적 신경망 방법

회귀함수를 적합시키는데 은닉층의 노드의 개수를 결정하는 것이 매우 중요하다. 본 논문에 사용된 자료와 동일한 자료에 대하여 Ripley(1995)는 다음의 결과를 얻었으며, 각 기준에 근거하여 은닉층 노드의 개수는 8이 적당함을 알 수 있다.

노드 개수	λ	SS_q	SS_{qcv}	p^*	NIC
2	0	3.47	4.29	7	4.03
	10^{-3}	3.57	4.78	5	3.98
3	0	4.12	4.81	10	4.92
	10^{-3}	3.40	4.54	6	3.93
	10^{-2}	3.61	4.91	6	4.11
	10^{-1}	5.40	6.20	6	5.94
5	0	3.14	4.57	16	4.42
	10^{-3}	3.21	4.51	17	4.59
	10^{-2}	3.61	4.91	6	4.11
	10^{-1}	5.38	6.18	6	5.92
8	0	2.89	5.35	25	4.89
	10^{-3}	3.21	4.54	18	4.67
	10^{-2}	3.58	4.94	7	4.16
	10^{-1}	5.36	6.14	7	5.90

여기서 SS_q는 평방합을 나타내고, SS_{qcv}는 또 다른 자료집합에 대해 적합된 10-중 CV 예측치를 나타낸다. 그리고, p*는 trace[$K'J^{-1}$]이다. 이때 $J=-E\frac{\partial^2 g(X_i,\theta_0)}{\partial\theta\partial\theta^T}$, $K=Var\frac{\partial g(X_i,\theta_0)}{\partial\theta}$ 이다. 이 척도들은 θ_0를 $\hat{\theta}$로 바꾸고, 기대값을 훈련자료의 평균으로 바꾸었을 때 미분값과 Hessian을 사용하여 추정되어질 수 있다. Moody의 p_{eff}는 p^*이며 AIC의 이런 버전을 NIC라 한다. 그리고 Hwang & Kim(1997)은 NIC가 붓스트랩 모형 선택기준과 동일함을 보였다. 그림11은 일반적인 신경망을 이용하여 추정된 회귀함수이다. 이상치로 보여지는 자료들의 분포가 y축의 0을 중심으로 그 영향을 상쇄하게끔 분포되어 있기때문에 오염된 자료에 대해서는 이상치영향을 덜 받는 것으로 보인다. 그러나 과대오차를 갖는 자료에서는 이상치로 보이는 관측값의 영향을 받고 있다. 이것으로 일반적인 신경망이 이상치에 민감다는 것을 알 수 있다.

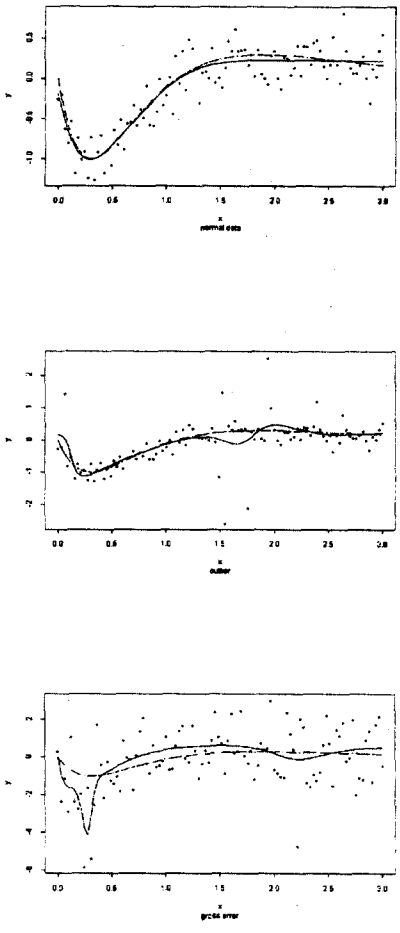


그림 11. 일반적 신경망 방법

◆S-PLUS program
>library(nnet)
>n<-nnet(x, y, , 8, linout=T)
>win.graph()
>plot(x, y)
>lines(x, n$fitted.value, lty=2)

4.5. 로버스트 신경망 방법

그림12는 이상치에 민감한 일반적인 신경망의 단점을 보완한 로버스트 신경망을 이용하여 회귀함수를 추정한 결과를 보여준다. 이상치 자료와 과대오차를 갖는 자료에 대해서는 회귀함수를 잘 추정하지만, 정규분포의 오차를 갖는 자료에 대해서는 일반적인 신경망이 회귀함수를 더 잘 추정하느것으로 생각된다. MATLAB을 사용하여 모의실험을 한 후에 S-PLUS로 그림을 그렸다.

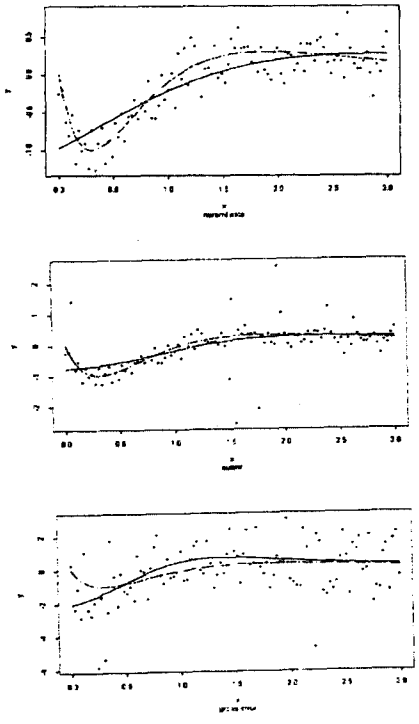
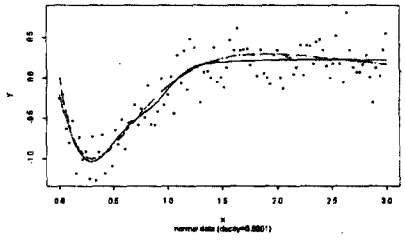


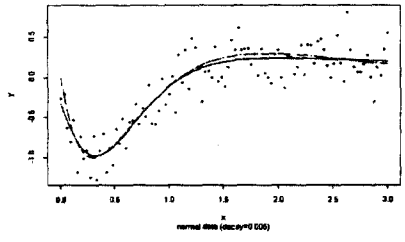
그림 12. 로버스트 신경망 방법

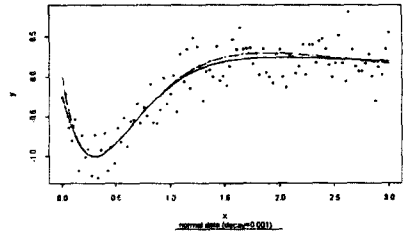
4.6. 가중치 감소법

그림13, 그림14, 그림15는 가중치 감소법으로 각각의 자료에 대하여 회귀함수추정한 결과이다. 그림13에서는 $\lambda = 0.0001$ 일 때 , 그림14에서는 $\lambda = 0.001$ 일 때 적절한 회귀함수추정을 하며, 그림15에서는 일반적인 신경망 방법과 유사하게 이상치로 오인된 관측값의 영향을 받는 것으로 보이며, $\lambda = 0.01$ 일 때 회귀함수추정이 적절한 것으로 보인다. Ripley(1996)는 여기서 사용된 벌칙항 대신에 스플라인 방법에서 사용된 것과 같은 벌칙항을 사용하여 비교 분석하는 것이 의미있다고 말하고 있다.

◆S-PLUS program
>library(nnet)
>wd <- nnet(x, y,, 8, linout=T, decay=0.001)
>win.graph()
>plot(x, y)
>lines(x, wd$fitted.value, lty=2)







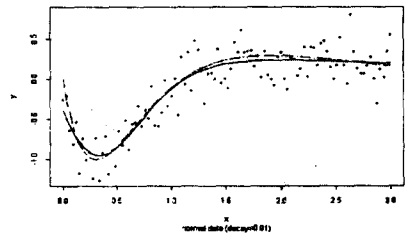
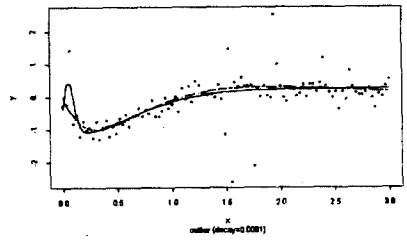
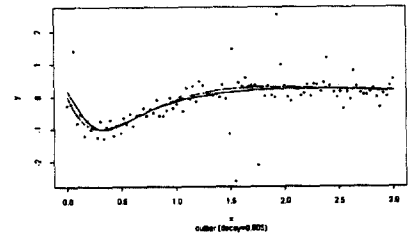
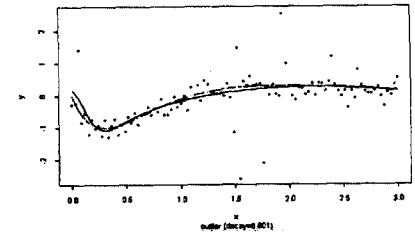


그림 13. 정규자료에 대한 가중치 감소법







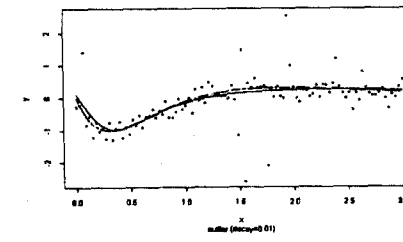
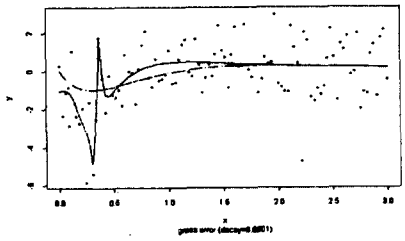
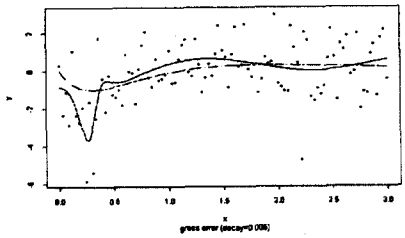
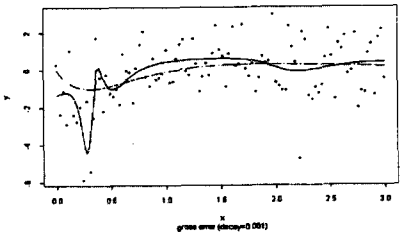


그림 14. 이상치 자료에 대한 가중치 감소법







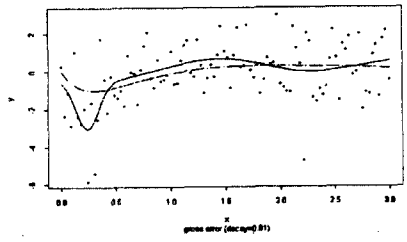


그림 15. 과대오차 자료에 대한 가중치 감소법

5. 결 론

주어진 자료에 커널, 스플라인, PPR, 신경망 방법을 적용시킨 결과 커널 방법과 스플라인 방법은 독립변수가 여러개인 경우는 차원문제로 좋은 회귀함수 추정을 못하지만 본 논문에서 사용한 자료는 독립변수가 하나이므로 적절한 회귀함수 추정을 하고 있다. 그러나 커널 방법은 자료의 국부적 정보에 대한 회귀함수 추정 방법이므로 적합결과가 매끄럽지 못하다. 한편, 차원문제를 극복하기위해 제안되었던 PPR 방법은 과대오차와 이상치에 민감하지 않으며 좋은 회귀함수 추정 결과를 보여주고 있으며, 이와 유사한 형태인 일반적인 신경망은 이상치와 과대오차에 민감한 회귀함수 추정 결과를 보여준다. 로버스트 신경망은 이상치에 덜 민감한 회귀함수 추정 결과를 보여주고 있으나, 가중치 감소법이 오히려 과대적합을 피하고 좋은 회귀함수추정을 하고 있는 것으로 보여진다. 판별분석, 문자인식 등과 같은 문제에 위의 방법들을 적용하여 성능을 평가하는 것이 필요하다고 생각되고, Chen & Jain(1994)의 로버스트 신경망을 이용하여 모의실험을 해 보는 것도 향후의 연구과제로서 좋을 것으로 생각된다.

참 고 문 헌

1. 황 창 하, 김 상 민, 박 희 주 (1997). 회귀분석을 위한 로버스트 신경망, 한국통계학회논문집 제4권 2호, 327-332.
2. Bishop, C. M. (1995a). *Bayesian methods for neural networks*, NCGR/95/009, Aston University, UK.
3. Bishop, C. M. (1995b). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.
4. Chen, D. S. & Jain, R. C. (1994). *A robust back propagation learning algorithm for function approximation*, IEEE Transactions on Neural

Networks 5, 467-469.

5. Diaconis, P. and Shahshahani, M.(1984), *On linear functions of linear combinations,* SIAM J. Sci. Stat. Comput, **5**, 175-191
6. Friedman, J. H. and Stuetzle, W.(1981), *Projection pursuit regression,* J. Amer. Statis.. Assoc, **76**, 817-823
7. Hwang, C. & Kim, D. (1997). *Bootstrap model selection criterion for determining the number of hidden units in neural network model,* Korean Communication in Statistics, **4**, 827-832.
8. Ripley, B. D. (1995). *Statistical ideas for selecting network architectures,* In Neural Neyworks: Artificial Intelligence and Industrial Applications, eds B. Kappen & S. Gielen. London: Springer.
9. Ripley, B. D. (1996). *Pattern Recognition and Neural Networks,* Cambridge University Press.
10. Simonoff, J. S. (1996). *Smoothing Methods in Statistics,* Springer.
11. Venables, W. N. & Ripley, B. D. (1994). *Modern Applied Statistics with S-PLUS,* Springer-Verlag.
12. Wahba, G. & Wold, S. (1975). *A completely automatic French curve,* Communications in Statistics, **4**, 1-17.