The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 743-758

Computation and Smoothing Parameter Selection in Penalized Likelihood Regression

Young-Ju Kim[1)]

Abstract

This paper consider penalized likelihood regression with data from exponential family. The fast computation method applied to Gaussian data(Kim and Gu, 2004) is extended to non Gaussian data through asymptotically efficient low dimensional approximations and corresponding algorithm is proposed. Also smoothing parameter selection is explored for various exponential families, which extends the existing cross validation method of Xiang and Wahba evaluated only with Bernoulli data.

Keywords : Cross-validation, Kullback-Leibler, Penalized likelihood, Smoothing parameter

1. 서론

스무딩 방법은 회귀분석에서의 함수 추정법 중 하나로서 추정함수의 굴곡성을 최대한 허용하기 위하여 추정함수의 거친 정도에 대한 벌점도를 함께 주는 방법이다. 이 논문에서는 지수족 자료 $Y_i \sim \exp\{(y\theta_i - b_i(\theta))/a(\phi) + c(y,\phi)\}$ 에 대한 회귀분석에서의 스무딩 스플라인을 다루고자 한다. 반응변수 Y 에 대하여 마이너스 로그 우도함수를 $l(\eta; Y)$로 두자. 이 때, $\eta(x)$는 공변량 x 에 의존하는 함수이다. 벌점우도 추정법은 다음과 같은 벌점우도 범함수를 최소화하는 해로 함수 η를 추정한다.

$$\frac{1}{n}\sum_{i=1}^{n} l_i(\eta(x_i); Y_i) + \frac{\lambda}{2} J(\eta) \tag{1.1}$$

여기서 $J(\eta)$는 η의 거친 정도에 대한 벌점도 범함수이고, λ는 평활(smoothing) 모수이다. 위 식(1.1) 의 첫째항은 η의 자료에 대한 적합도를 장려하는 역할을 하는 반면, 두 번째 항은 η의 거친 정도에 대해 벌점을 준다. 이러한 두 가지의 상반된 목적을 조절하는 역할을 하는 것이 평활 모수 λ이다. 주로 무한 차원 공간 $H \subseteq \{\eta; J(\eta) < \infty\}$에 존재하는 위 식(1.1)의 최소해를 스무딩 스

1) Full-Time Instructor, Department of Information Statistics, Kangwon National University, Chuncheon 200-701, Korea
E-mail : ykim7stat@kangwon.ac.kr

플라인이라고 하고 η_λ로 표시한다. 예를 들어, 벌점도함수가 $\int \ddot{\eta}^2 dx$일 때 계산된 최소해를 큐빅 스무딩 스플라인(cubic smoothing spline)이라고 한다.

지수족 분포를 따르는 자료에 대한 벌점우도회귀는 O'Sullivan, Yandell, and Raynor(1986), Silverman(1978), Green and Yandell(1985)에 의해 정립되고 연구되었다. Cox and O'Sullivan (1990), Gu and Qiu(1994), Gu and Kim(2002) 등은 최소해 η_λ 의 점근적 수렴률에 대하여 연구하였고, Gu(1990,1992), Xiang and Wahba(1996), Gu and Xiang(2001)은 평활 모수의 선택방법과 최소해의 계산방법에 대하여 논하였다.

공변량 x가 다차원인 경우에 이러한 최소해를 계산하기 위하여 $O(n^3)$ 알고리듬이 요구된다. 특히 n이 커지는 경우, 즉 대용량의 자료인 경우에는 최소해 계산이 현실적으로 힘들어지게 된다. 이 논문의 목적은 벌점우도회귀에서의 최소해를 빠르고 효율적으로 계산하는 방법과 그에 대응하는 적절한 알고리듬을 제시하는 것이다. Gu and Kim(2002)는 (1.1)의 최소해를 적당한 정수 $q \ll n$ 에 대하여 q-차원의 함수공간 H_q로 근사시켰을 때의 근사해가 정확해와 같은 점근적 수렴률을 가진다는 것을 보였다. 이 때, q의 오더는 $O(n^{2/(pr+1)+\varepsilon})$, $p \in [1,2]$, $r > 1$, $\forall \varepsilon > 0$, 으로 적당히 낮게 조절되어야 한다. $p \in [1,2]$는 참해 η의 평활도(smoothness)에 따라 결정되며, 평활도가 높을수록 2에 가깝다.(참고로, $\int (\eta^{(4)})^2 dx < \infty$이면 $p=2$로 둔다.) r은 벌점도 함수 $J(\eta)$에 의해 결정되는 평활도를 나타낸다.(큐빅 스플라인인 경우 r=4로 둔다.) 이러한 q-차원 근사해의 계산은 $O(nq^2)$로 낮아진다. 특히, $q \asymp n^{2/9}$일 때 q-차원의 근사해의 계산은 $O(n^{13/9})$가 된다. 최근에 Kim and Gu(2004)은 이러한 저차원 근사해를 이용한 빠른 계산방법을 가우시안 자료에 적용시켰다. 이 논문은 비가우시안 자료에 대하여 이러한 저차원 근사해를 이용한 계산방법을 개발하고 적용시키고자 한다.

스무딩 스플라인에서 추정함수의 성능을 결정하는 것은 평활 모수이다. 따라서 평활 모수의 선택방법은 스무딩 스플라인에서 아주 중요한 문제가 된다. 비가우시안 자료에 대한 기존의 평활 모수의 선택방법으로는 Gu(1992)가 제시한 performance-oriented iteration이라고도 불리는 indirect cross-validation 방법과 Xiang and Wahba(1996)와 Gu and Xiang(2001)이 제시한 direct cross-validation 방법이 있다. 이 논문에서는 두 가지의 평활 모수 계산방법 중 좀더 효율적이고 직접적인 계산방법으로 알려진 Gu and Xiang(2001)의 direct cross-validation 방법을 이용하여 제시된 AGACV(Alternative Generalized Alternative Cross-Validation) score를 사용한다. AGACV score는 Xiang and Wahba(1996)와 Gu and Xiang(2001)에 의해 Bernoulli 자료에 대하여 그 성능이 입증되었다. 그러나 그 외의 다른 지수족 자료에 대한 평활 모수의 선택방법의 검증과 개발의 필요성이 여전히 남아 있다. 이 논문은 흔히 사용되는 지수족 자료에 대한 direct cross-validation 방법의 검증과 적절한 수정 및 개발 방법을 제시하고 경험적 성능을 보이고자 한다.

2. 벌점우도회귀의 계산

2.1 RK와 해의 표현

내적 $\langle \cdot, \cdot \rangle$에 대하여 $\langle R(x, \cdot), f(\cdot) \rangle = f(x)$와 같은 성질을 만족하는 비음 정치 함수를 RK(Reproducing Kernel)이라고 하는데, 이러한 RK를 가지는 (1.1)의 최소해의 함수공간을 RKHS(Reproducing Kernel Hilbert Space)라고 한다.

벌점도 함수 $J(\eta)$의 영공간을 $N_J = \{\eta : J(\eta) = 0\}$으로 두면, RKHS는 $H = N_J \oplus H_J$ 와 같은 텐서 합 분해로 나타나게 된다. H_J는 $J(\eta)$를 제곱노름으로 갖는 RKHS이다. RKHS의 닫힌 부분공간의 성질을 이용하면 벌점우도 범함수를 최소화하는 정확해는 무한차원 공간이 아닌 다음과 같은 유한차원 공간에서 표현된다.

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R_J(x_i, x) \tag{2.1}$$

이 때, $\{\phi_\nu\}$는 벌점도 함수의 영공간의 기저함수들이고 R_J 는 H_J 에서의 RK이다.

예를 들어, χ=[0,1]에서 벌점도 범함수 $J(\eta) = \int \ddot{\eta}^2 dx$로 두면, 영공간 N_J =span $\{1, k_1(x)\}$, $H_J = \left\{\eta : \int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0, J(\eta) < \infty\right\}$인 큐빅 스플라인을 얻을 수 있다. 이 때 H_J 에서의 RK는 $R_J(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(x_1 - x_2)$가 된다. ($k_1(x) = x - 0.5$, $k_\nu = B_\nu / \nu!$ 이고 B_ν는 Bernoulli 다항식이다.)

2.2 반복 가중최소제곱

지수족 분포를 따르는 자료인 경우에는 밀도함수가 $\exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$ 형태이므로 우도함수는 $l(\eta; Y) = -\{Y\theta(\eta) - b(\theta(\eta))\}$가 된다. 여기서 θ는 정준 모수이고 $a(\phi)$는 산포 모수이다. 그리고 η는 θ의 단조 변환으로 두어 추정함수 η의 범위를 제한하지 않는다. $c(Y, \phi)$은 η에 의존하지 않으므로 생략할 수 있고 $a(\phi)$는 λ에 흡수되어진다.

$u(\eta; Y) = dl/d\eta$, $w(\eta; Y) = d^2 l/d\eta^2$로 두면 η_0은 참함수일 때 $E[u(\eta_0; Y)] = 0$ 이고 $E[u^2(\eta_0; Y)] = \sigma^2 E[w(\eta_0; Y)]$ 가 되는 것을 쉽게 보일 수 있다. 여기서 σ^2는 상수이다.

고정된 평활 모수 λ에 대하여 (1.1)의 최소해는 Newton iterations을 이용하여 구한다. $\tilde{\eta}(x_i)$에서 (1.1)의 로그우도함수 $l(\eta(x_i); Y_i)$의 이차근사를 구하면 다음과 같이 벌점가중최소제곱 범함수 형태로 나타난다.

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{w}_i (\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{2.2}$$

여기서 $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i / \tilde{w}_i$, $\tilde{u}_i = u(\tilde{\eta}(x_i); Y_i)$ $\tilde{w}_i = w(\tilde{\eta}(x_i); Y_i)$ 이다. Newton iteration을 이용하여 벌점가중최소제곱 범함수 형태의 최소해 $\eta_{\lambda, \tilde{\eta}}$ 로 $\tilde{\eta}$을 업데이트 시켜 나간

다. 이렇게 수렴되는 추정함수가 구하는 해가 된다.

Gu and Kim(2002)는 크기가 $q \le n$인 랜덤 부분집합 $\{z_j, j=1,\ldots,q\} \subseteq \{x_i, i=1,\ldots,n\}$ 에 대하여 다음과 같은 H_q에서의 q-차원의 근사해를 제시하였다.

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_j(z_j, x) = \phi^T d + \xi^T c \tag{2.3}$$

여기서 $\{\phi_\nu\}$는 영공간 N_f의 기저이고 $\xi_j(x) = R_j(v_j, x)$는 RK이다. 이 표현식을 (2.2)에 대입하여 간단한 계산을 하면

$$(\tilde{Y} - Sd - Rc)^T \tilde{W}(\tilde{Y} - Sd - Rc) + n\lambda c^T Qc \tag{2.4}$$

이 때, $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, $\tilde{W} = diag(\tilde{w}_1, \ldots, \tilde{w}_n)$, S 는 (i,ν)번째 원소가 $\phi_\nu(x_i)$인 $n \times m$ 행렬이고, R은 (i,j)번째 원소가 $R(z_j, x_i)$인 $n \times q$ 행렬, 그리고 Q는 (j,k)번째 원소가 $R(z_j, z_k)$인 $q \times q$ 행렬이다. $Y_w = \tilde{W}^{1/2}\tilde{Y}$, $\tilde{Y}_w = \tilde{W}^{1/2}(Sd + Rc)$로 두면, $\hat{\tilde{Y}}_w = A_w(\lambda) Y_w$가 되고 A_w은 평활 행렬이다. (2.4)의 계산은 Kim and Gu(2004)의 최소제곱에서와 마찬가지로 Cholesky 분해와 전진/후방 대입에 의해 할 수 있다.

3. 평활 모수의 선택

앞 절에서 제시된 최소해의 계산 알고리듬은 고정된 평활 모수 λ에 대하여 전개되었다. 평활 모수와 최소해를 계산하는 방법은 두 개의 중첩된 반복 루프로 이루어지는데, 안쪽 루프는 고정된 평활 모수에 대하여 최소해를 계산하고 바깥쪽 루프는 cross-validation을 이용하여 평활 모수를 선택하게 된다. 이 절에서는 먼저 일반적인 평활 모수의 선택방법에 대하여 논한 후, 각 지수족에 대하여 고려한다.

3.1 Cross-Validation

Kullback-Leibler 거리는 최소해(2.1) 또는 저차원 근사 최소해(2.3) η_λ와 참해 η_0의 차이를 측정한다. 평활 모수의 선택은 이러한 Kullback-Leibler 거리를 최소화하는 해를 선택하는 방법으로 결정된다. 지수족에 대한 Kullback-Leibler 거리는

$$KL(\eta_\lambda, \eta_0) = \{\mu(\eta_0)(\theta(\eta_0) - \theta(\eta_\lambda)) - (b(\theta(\eta_0)) - b(\theta(\eta_\lambda)))\}/a(\phi), \quad \mu = db/d\theta = E(Y).$$

관련이 없는 항들은 제거하고 표본점들에 대하여 평균을 취하면 다음과 같은 Relative Kullback-Leibler 거리를 구할 수 있다.

$$RKL(\eta_\lambda, \eta_0) = \frac{1}{n}\sum_{i=1}^{n}\{-\mu(\eta_0(x_i))\theta(\eta_\lambda(x_i)) + b_i\{\theta(\eta_\lambda(x_i))\}$$

$\mu(\eta_0(x_i))\theta(\eta_\lambda(x_i))$를 $Y_i\theta(\eta_\lambda^{[i]}(x_i))$로 교체하고 RKL의 cross validation 추정치를 계산하면

$$\frac{1}{n}\sum_{i=1}^{n} l_i(\eta_\lambda(x_i); Y_i) + \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\theta(\eta_\lambda(x_i)) - \theta(\eta_\lambda^{[i]}(x_i)))\} \tag{3.1}$$

가 되고, 여기서 $\eta_\lambda^{[k]}$ 는 (1.1)의 delete-1 버전의 최소해이다. 그러나, $\eta_\lambda^{[k]}$의 계산이 쉽지 않기 때문에 이러한 cross validation score(3.1)를 계산하는 것이 실제적으로 용이하지 않다. Gu and Xiang(2001)는 $\eta_\lambda^{[k]}$대신에 벌점가중최소제곱 범함수(2.2)의 delete-1 버전의 최소해 $\eta_{\lambda,\eta_\lambda}^{[k]}$으로 대체한 후 간단한 계산을 통하여 다음과 같은 AGACV(Alternative Generalized Approximate Cross-Validation) score를 제시하였다.

$$V^*(\lambda) = \frac{1}{n}\sum_{i=1}^{n} l_i(\eta_\lambda(x_i); Y_i) + \frac{tr(A_w \tilde{W}^{-1})}{n - tr(A_w)}\frac{1}{n}\sum_{i=1}^{n} h_i(-\tilde{u}_i) \tag{3.2}$$

여기서 $h_i = Y_i(d\theta/d\eta|_{\eta_\lambda(x_i)})$, A_w는 평활 행렬이다. (3.2)와 같은 score를 최소화하는 평활 모수를 최적 평활 모수라고 한다. Xiang and Wahba(1996)과 Gu and Xiang(2001)은 이러한 방법을 Bernoulli 자료에만 적용시켜 그 경험적 성능을 조사하였다.

다음 절에서는 이러한 AGACV score를 Bernoulli 자료뿐만 아니라 다른 지수족에도 적용시켜 그 경험적 성능을 조사한다. 저평활(undersmoothing)문제가 발생할 때에는 1보다 큰 상수 α를 두 번째 항에 삽입하여 저평활 문제를 해결하려고 한다. AGACV score(3.2)는 마이너스 로그우도함수와 대각합(trace)의 항으로 이루어져 있다. 평활모수 λ가 작아질수록 (3.2)의 첫째 항인 마이너스 로그우도함수는 감소하는 반면 두 번째 항은 증가하므로 적당한 $\alpha \geq 1$을 대각합의 항 앞에 삽입하여 좀더 평활도가 높은 해를 얻을 수 있게 된다. 이렇게 수정된 AGACV score의 성능이 좋지 않을 경우에는 좀더 효율적이고 적절한 score를 제시하여 각 지수족 분포에 맞는 평활 모수의 선택방법을 제시하고 모의실험을 통해 그 성능을 비교한다. 이 논문에서 다루게 될 지수족 분포로 Binomial, Poisson, Gamma 분포로서 일반적으로 널리 고려되고 사용되고 있는 분포들이다.

3.2 Binomial 족

Binomial 자료 $Y_i \sim Bin(m_i, p(x_i))$에 대하여, $\eta = \theta = \log\{p/(1-p)\}$이고 마이너스 로그우도함수는 $l_i(\eta; Y) = -Y_i\eta + m_i\log(1+e^\eta)$가 된다. 이 때 $p_i = e^{\eta(x_i)}/(1+e^{\eta(x_i)})$이고 $h_i = Y_i$이며 $u_i = m_i p_i - Y_i$, $w_i = m_i p_i(1-p_i)$이다. 실제로 Binomial 자료는 서로 독립인 Bernoulli 자료의 합이므로 Binomial 자료에 대한 delete-1 cross-validation을 생각할 때 Bernoulli 자료에

대하여 하나의 자료가 삭제되는 것이 아니라 실제로 각각 m_i만큼의 자료가 삭제되는 결과를 낳는다. 즉 delete-1 cross-validation이 아니라 delete-m 이 되는 것이다. 그러므로 이 논문에서는 Binomial 자료의 delete-1 cross-validation을 Bernoulli 자료의 관점에서 생각하여 실제로 하나의 자료가 삭제되도록 유도하였다. 간단한 계산 결과 다음과 같은 score를 얻을 수 있다.

$$V(\lambda) = \frac{1}{N}\sum_{i=1}^{n} l_i(\eta_\lambda(x_i); Y_i) + \frac{tr(A_w M \tilde{W}^{-1})}{N - trA_w}\frac{1}{N}\sum_{i=1}^{n} Y_i(1-\tilde{p}_i) \tag{3.3}$$

여기서 $N = \sum_{i=1}^{n} m_i$이고 A_w와 $\tilde{W}$는 각각 $N \times N$ 행렬로서 각 표본점 x_i에 대응하는 원소들은 크기 m_i의 블록을 이룬다는 것을 알 수 있다.

3.3 Poisson 족

$Y_i \sim Poisson(\lambda(x_i))$에 대하여 $\eta = \theta = \log\lambda$이고 $l_i(\eta; Y_i) = -Y_i\eta + e^{\eta}$이다. 그리고 $u_i = e^{\eta(x_i)} - Y_i$, $w_i = e^{\eta(x_i)}$이다. 기존의 AGACV score(3.2)를 Poisson 자료에 적용시켜 보면 그리 좋지 않은 성능을 보여주는 것을 관찰할 수 있었다. 이러한 현상은 특히 (3.2)가 최소치를 갖게 되는 평활모수 근처에서 상대적으로 평평하게 나타나는 경우에서 많이 나타났는데, 이로 인하여 최적 평활모수가 제대로 선택되지 않았다. 이것은 Poisson 자료에 대한 반복가중최소제곱의 발산 또는 근사된 cross-validation 방법 자체의 성능과 관계가 있을 것으로 추측된다. AGACV의 나쁜 성능을 극복하기 위한 여러 가지 가능한 대체방안들 중에서, Poisson 회귀를 밀도추정으로 해석하면 밀도추정에서 유도되는 delete-1 cross-validation을 이용할 수 있다는 사실을 고려하였다. Poisson 우도함수를 (1.1)에 대입하여 간단한 계산을 하면 η와 관계되는 항들은 $-\sum_{i=1}^{n} Y_i\{\eta(x_i) - \log\int e^{\eta}\} + \frac{n\lambda}{2}J(\eta)$ 형태가 된다. 이것은 각 x_i에서의 bin size Y_i를 가지는 binned 자료로 생각하면 이산도메인에서의 밀도함수 $e^{\eta}/\int e^{\eta}$를 추정하는 문제와 동일해진다. 이것은 KL loss 함수의 CV 근사문제를 밀도함수문제의 관점에서 재해석한 결과이다. 벌점밀도추정문제를 다룬 Gu and Wang(2003)의 cross-validation score

$$V(\lambda) = \frac{1}{N}\sum_{i=1}^{n} l(\eta_\lambda(x_i); Y_i) + a\frac{tr(P_y \tilde{R} H + \tilde{R}^T P_y^T)}{N(N-1)} \tag{3.4}$$

를 이용하여 평활 모수를 선택하되, 적절한 상수 $a \geq 1$를 대입하여 저평활 문제를 간단하게 해결하였다. 여기서 $\tilde{R} = (S, R)$, $P_y = (I - \tilde{y}\,\tilde{y}^T/N)diag(\tilde{y})$, $\tilde{y} = (\sqrt{Y_1}, \ldots, \sqrt{Y_n})^T$,

$$H=\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi}+(n\lambda)Q/N \end{pmatrix},$$

$V_{\xi,\phi}$는 $q\times m$ 행렬로서 (j,ν)번째 원소가

$$\frac{1}{N}\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\xi_j(x_i)\phi_\nu(x_i)-\frac{1}{N}\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\xi_j(x_i)\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\phi_\nu(x_i)$$

이며 다른 V행렬들도 비슷하게 정의될 수 있다. 여기서 $\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}=N$ 이다. 3.5절의 모의실험 결과 α의 값은 1.4 정도가 가장 적당한 것으로 나타났다.

3.4 Gamma 족

$Y_i\sim Gam(\nu,\beta(x_i))$인 자료의 밀도함수는 $\{\beta^\nu\Gamma(\nu)\}^{-1}y^{\nu-1}e^{-y/\beta}I_{\{y>0\}}$, $\theta=-1/\mu$, 그리고 $\eta=\log\mu$, $\mu=\nu\beta=E(Y)$, $l_i(\eta;Y_i)=Y_ie^{-\eta}+\eta$, $u_i=-Y_ie^{-\eta(x_i)}+1$, $w_i=Y_ie^{-\eta(x_i)}$, $h_i=Y_ie^{-\eta(x_i)}$이다. 기존의 AGACV score(3.2)를 여러 테스트 함수들을 가지고 적용시켜 본 결과 몇몇 반복표본들에 대하여 저평활이 나타났다. 3.1절에서 논한 바와 같이, AGACV의 두 번째 항에 1보다 큰 상수 α를 삽입하는 간단한 수정을 가하였을 때 그 성능이 크게 향상되는 것을 확인할 수 있었다. α의 값으로 1.4 정도가 가장 적당하였다.

3.5 모의실험

이 절에서는 각 지수족에 대하여 새롭게 유도된 score들과 기존의 score(3.2)의 경험적 성능을 비교하려고 한다. 각 지수족에 대하여 다음과 같은 서로 다른 세 가지 테스트 함수들과 서로 다른 세 가지의 SNR(signal to noise ratio)를 가지고 각각 크기가 $n=100, 500$인 자료를 생성하였다.

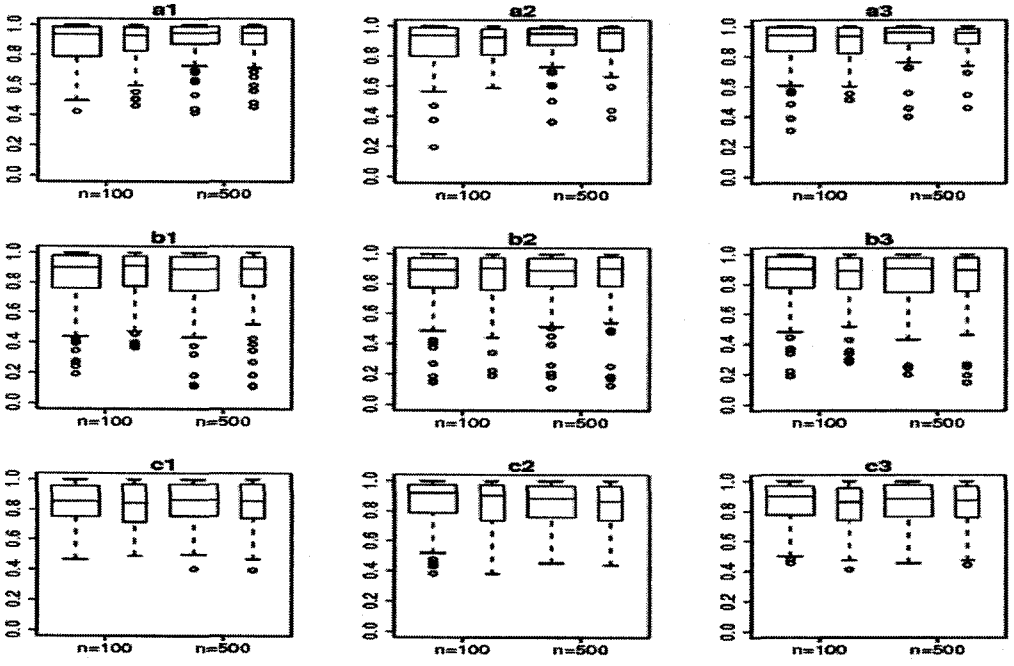
$$\eta_1(x)=1980x^7(1-x)^3+858x^2(1-x)^{10}-2$$
$$\eta_2(x)=2\sin(2\pi x)+0.1$$
$$\eta_3(x)=e^{-(x-0.5)^2}$$

각 테스트 함수에 대하여 각 표본의 크기 n=100, 500 에 대하여 $x_i=(i-0.5)/n$ 를 사용하였다. Binomial 인 경우, $p\in[0.03,0.97]$이 되도록 각 테스트 함수들을 조절하고 각 $m_i=3,5,7$인 Binomial 분포에서 자료를 생성하였다. Poisson인 경우, $\lambda_i\in[0.2,l]$이고 l 은 3,6,9 중 하나가 되도록 테스트 함수들을 조절하였다. Gamma인 경우 형태모수(shape parameter) $\nu=2,3,4$에 대하여 평균 $\mu\in[0.2,3]$이 되도록 테스트 함수들을 조절하였다. 서로 다른 분포, n의 크기와

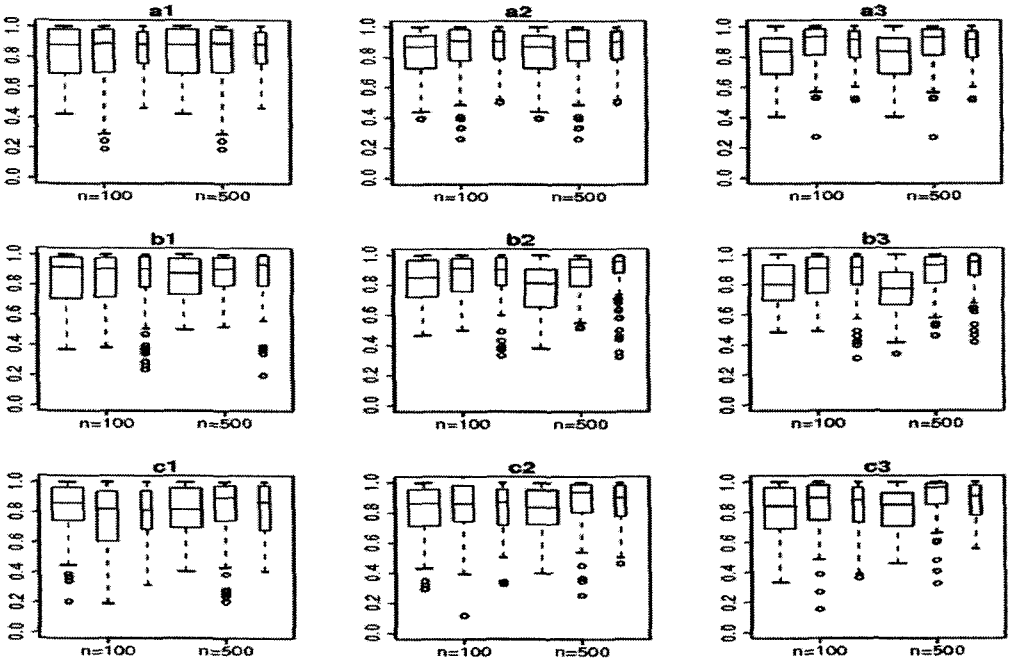
SNR를 가진 각 테스트 함수에 대하여 각각 100개의 반복표본을 생성하였다. 각 표본에 대하여 $q=n$일 때의 큐빅 스플라인 η_λ를 계산하였다. 이 때, 평활 모수는 기존의 score(3.2)와 각 분포에 대하여 새롭게 제시된 scores, 그리고 Kullback-Leibler 손실을 최소화하는 최적 모수를 선택하여 각각 $\lambda_1, \lambda_m, \lambda_0$로 표시하였다. 그림 3.1, 3.2, 3.3은 각각 Binomial, Poisson, Gamma에 대하여 세 테스트 함수들을 이용한 모의실험의 결과를 요약하였다. 새로운 scores의 경험적 성능을 비교하기 위하여 상대적 효율성(relative efficacy)

$$Eff(\eta_\lambda)=\frac{\min KL(\eta,\eta_{\lambda_0})}{KL(\eta,\eta_\lambda)}$$

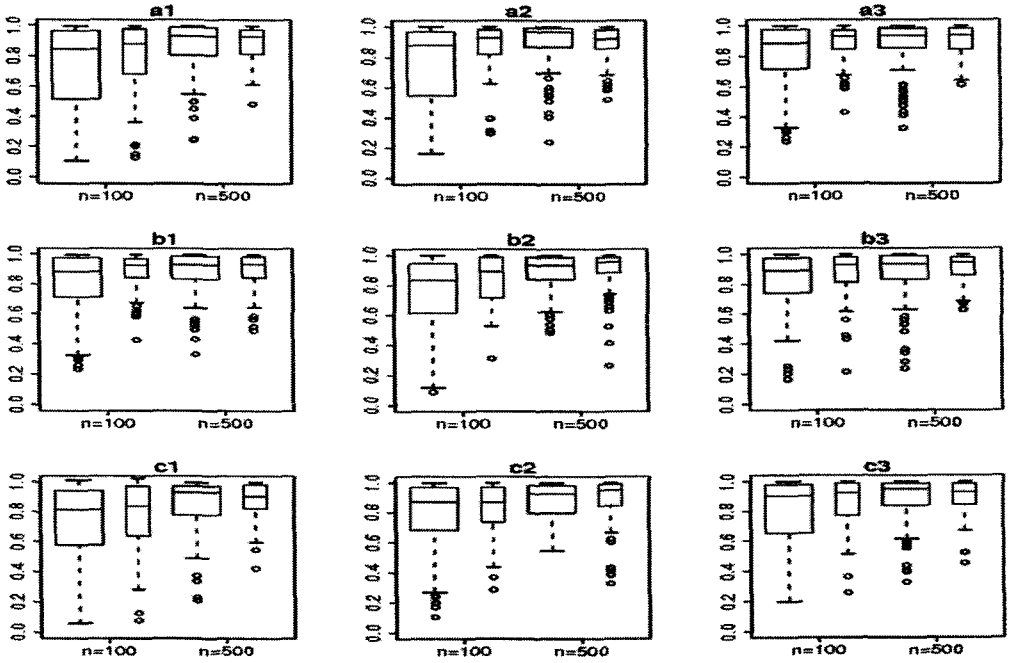
을 계산하여 상자그림으로 나타내었다. 그림 3.1은 각 테스트 함수들에 대하여 표본크기 $n=$ 100, 500에 대하여 Binomial 자료를 생성하여 각각 (3.2)를 최소화하는 평활모수를 가지고 계산한 최소해(두꺼운 상자)와 (3.3)을 최소화하는 평활모수를 가지고 계산한 최소해(얇은 상자)의 상대적 효율성을 나타내고 있다. a1-a3은 η_1에 대하여 각각 m_i=3,5,7, 그리고 b1-b3과 c1-c3은 각각 η_2와 η_3에 대하여 l=3,6,9일 때와 ν=2,3,4를 이용하여 계산한 결과이다. 모의실험 결과 앞서 제시한 score(3.3)가 기존의 score(3.2)보다 열등하지 않다는 것을 알 수 있었다. 다변수 함수를 포함한 다른 여러 종류의 테스트 함수를 가지고 실시된 모의실험들의 결과도 위의 결과와 질적으로(qualitatively) 일관되게 나타났다.



<그림 3.1>Binomial 자료에 대한 (3.3)의 성능



<그림 3.2> Poisson 자료에 대한 (3.4)의 성능



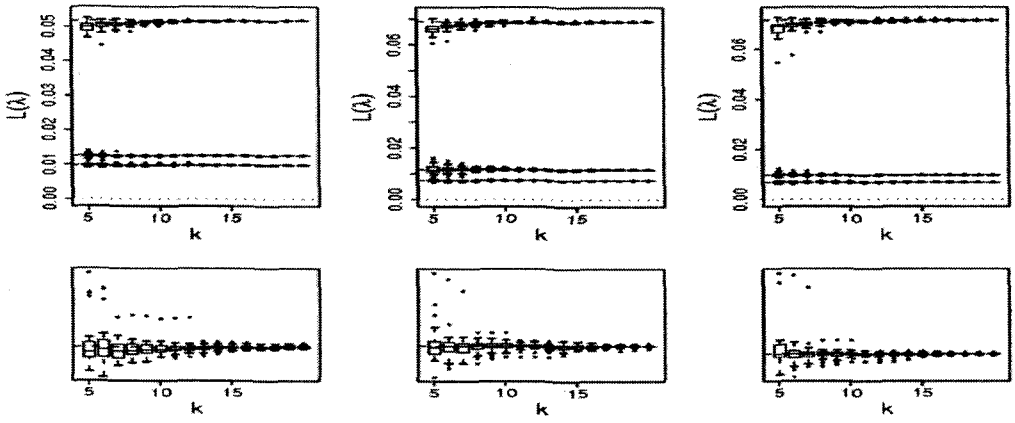
<그림 3.3> Gamma 자료에 대한 (3.2)의 성능

그림 3.2는 Poisson 자료에 대하여 score(3.4)의 성능을 기존의 AGACV(3.2)에 비교하기 위하여 실시한 모의실험의 결과를 요약하였다. 그림 3.1과 마찬가지로 각 세 개의 테스트 함수들과 서로 다른 SNR를 이용하여 각 표본크기 $n=100, 500$에 대하여 score(3.2)와 (3.4)의 상대적 효율성을 보여주고 있다. 각 상자그림의 첫 번째 두꺼운 상자는 score(3.2), 두 번째로 두꺼운 상자는 α=1인 score(3.4), 그리고 세 번째 얇은 상자는 α=1.4인 score(3.4)의 상대적 효율성을 나타낸다. 다른 많은 모의실험들의 결과 delete-1 cross-validation(3.4)이 기존의 score(3.2)보다 나은 성능을 보이며, 특히 α=1.4일 때의 score(3.4)가 전반적으로 제일 좋은 성능을 보여주는 것을 확인할 수 있었다. 그림 3.3은 Gamma 자료에 대한 모의실험의 결과를 요약하였다. 그림 3.1과 마찬가지로 세 개의 테스트 함수들과 각각의 SNR를 가지고 각 표본크기 $n=100, 500$에 대하여 α=1(두꺼운 상자)과 α=1.4(얇은 상자)일 때의 score(3.2)를 이용하여 계산된 상대적 효율성을 나타낸다. α=1.4일 때 저평활 문제가 개선되는 것을 확인할 수 있다. 일련의 모의실험에서 α의 값은 1.2에서 1.4 정도로 결정할 수 있다.

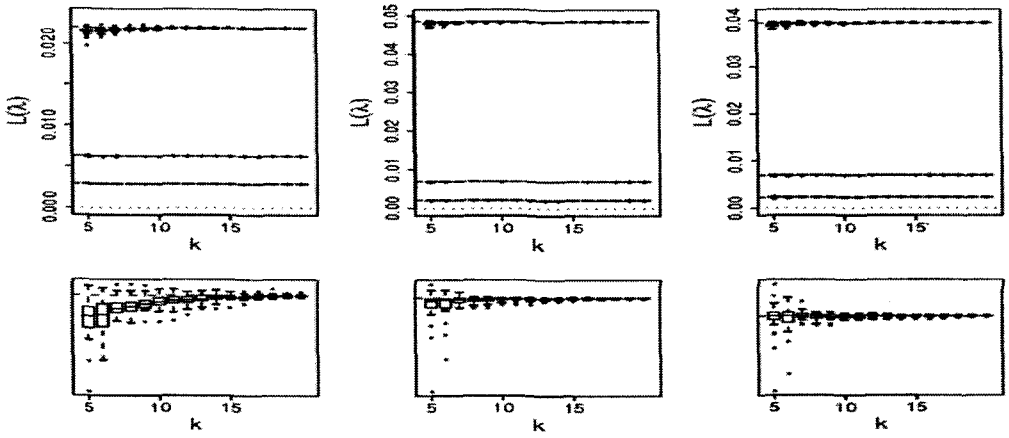
4. q의 선택

Gu and Kim(2002)은 q의 오더가 $O(n^{2/(pr+1)+\varepsilon})$이면 q-차원의 근사해(2.3)가 정확해와 같은 점근적 수렴률을 가진다는 것을 보였다. 실제로 $q=kn^{2/(pr+1)}$으로 두면 모의실험을 통하여 k의 값을 결정할 수 있다. 큐빅 스플라인에 대하여 $r=4$이며 참함수들은 대부분 충분히 매끄럽다는 가정 아래 $p=2$로 둔다. 3절에서 다루었던 지수족 분포들에 대하여 다음과 같이 모의실험을 실시하였다. 각 지수족에 대하여 3.5절에서 사용한 각 테스트 함수들을 이용하여 표본크기 n=100, 300, 500에 대하여 표본들을 추출했다. 각 표본과 각 $k=$ 5(1)15 에 대하여 크기가 $q=kn^{2/9}$인 30개의 서로 다른 랜덤 표본 $\{z_j\}\subset\{x_i\}$을 추출하여 3절에서 제시한 평활 모수의 선택방법들을 이용하여 저차원 근사해를 계산하였다. 또한 $q=n$에 대하여 정확해인 큐빅 스플라인을 계산하였다. 그림 4.1, 4.2, 4.3은 각 η_1, η_2, η_3에 대하여 각각의 SNR를 가지고 생성한 Binomial자료에 대하여 계산된 큐빅 스플라인들에 대하여 계산한 KL 손실들을 상자그림으로 나타내었다. 각 그림의 상단에 있는 3개의 그래프의 상자그림들은 위에서 아래로 n=100 ,300, 500에 대한 것이다. 아래의 3개는 n=500일 때의 상자그림의 선명도를 높인 것이다. 점선은 $q=n$일 때의 최소해를 나타낸다. 그림 4.4, 4.5, 4.6은 Poisson 자료에 대하여, 그림 4.7, 4.8, 4.9는 Gamma 자료에 대하여 같은 방법으로 계산한 KL 손실들을 상자그림으로 나타낸 것이다.

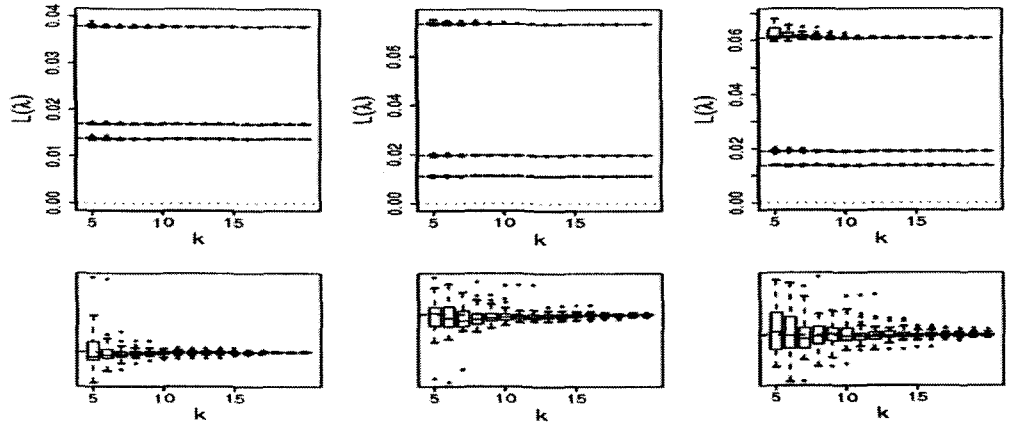
각 그래프들에서 보이는 것과 같이 k가 커질수록 상자의 높이가 줄어들면서 점선으로 가까워지는 것을 볼 수 있다. 그리고 k의 값이 약 10 정도가 되면 상자들의 높이가 안정적이 되는 것을 확인할 수 있다. 그러므로 상자그림들로부터 k가 10 정도일 때가 충분히 안정적으로 사용할 수 있는 값이 된다. 같은 크기의 다른 랜덤 부분집합들을 추출하여 근사해를 계산하여 안정성을 확인할 수 있다. 참고로, 임의로 거칠음이 많은 참함수를 가정한다면 p=1이 되고 $q=kn^{2/5}$ 가 되지만 모의실험의 결과는 질적으로 거의 동일하게 나타났다.



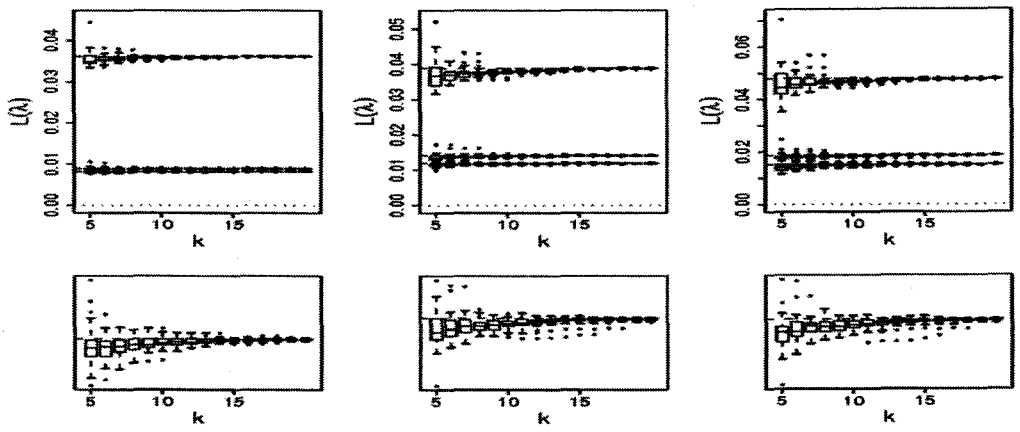
<그림 4.1> q 의 값이 근사해의 일관성에 미치는 영향: Binomial 자료 with η_1



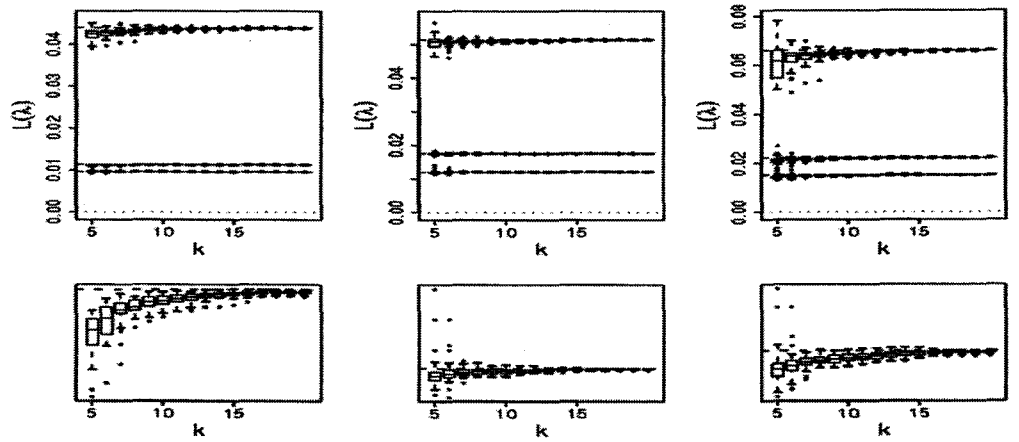
<그림 4.2> q 의 값이 근사해의 일관성에 미치는 영향: Binomial 자료 with η_2



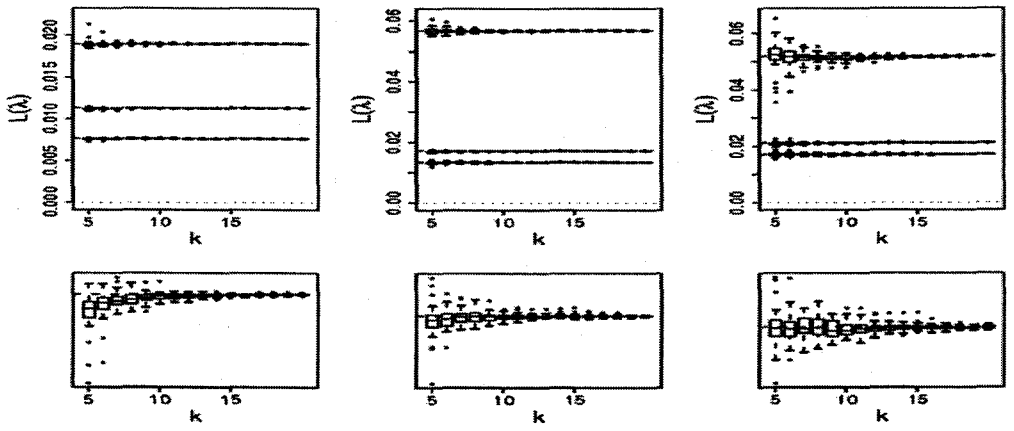
<그림 4.3> q 의 값이 근사해의 일관성에 미치는 영향: Binomial 자료 with η_3



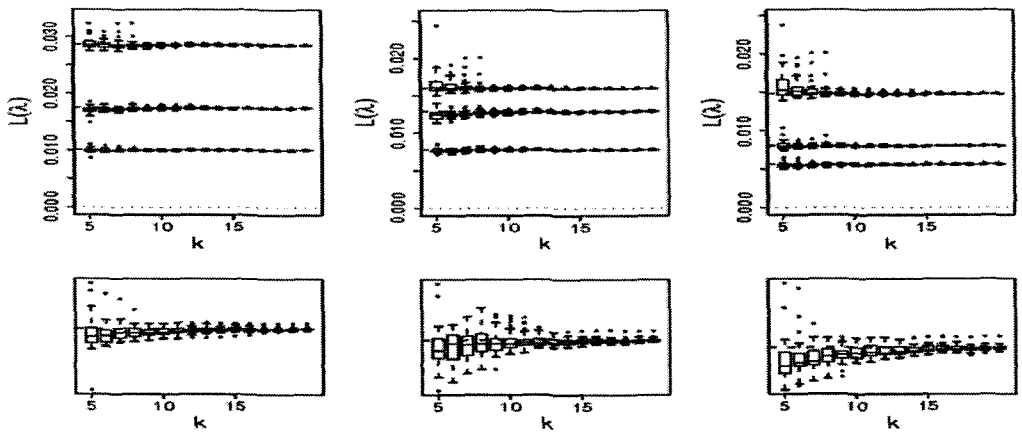
<그림 4.4> q 의 값이 근사해의 일관성에 미치는 영향: Poisson 자료 with η_1



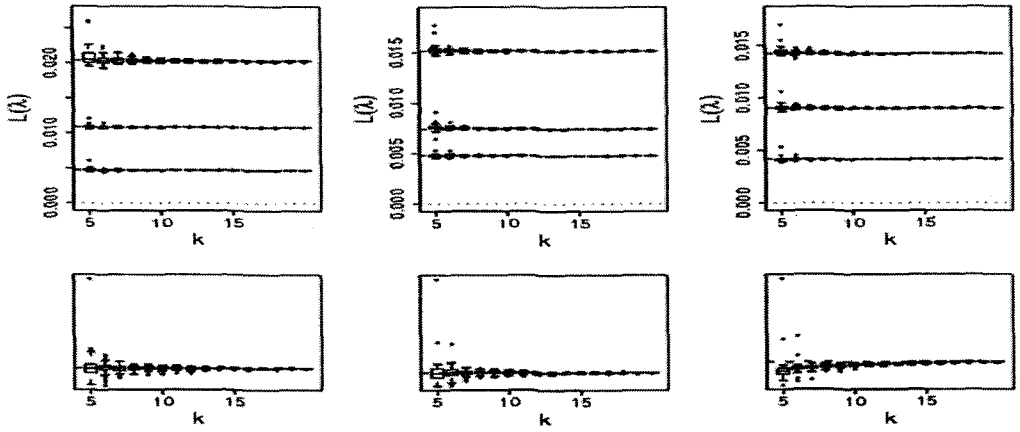
<그림 4.5> q 의 값이 근사해의 일관성에 미치는 영향: Poisson 자료 with η_2



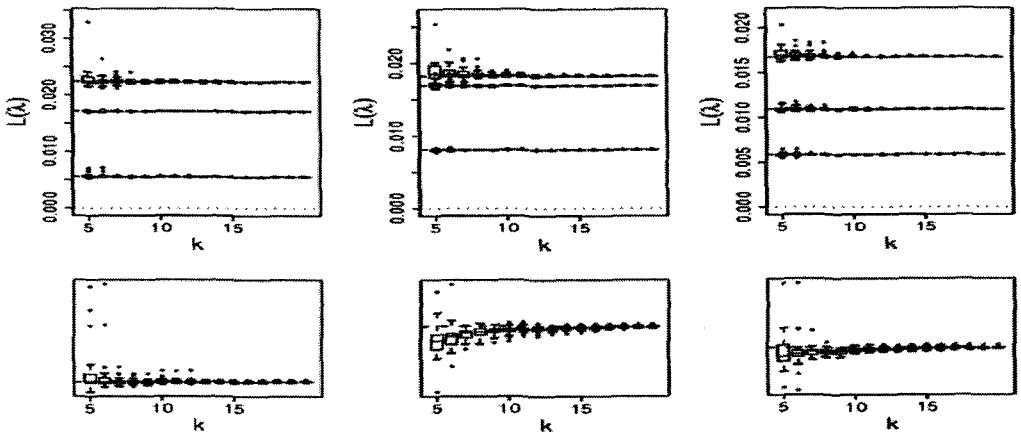
<그림 4.6> q 의 값이 근사해의 일관성에 미치는 영향: Poisson 자료 with η_3



<그림 4.7> q 의 값이 근사해의 일관성에 미치는 영향: Gamma 자료 with η_1



<그림 4.8> q 의 값이 근사해의 일관성에 미치는 영향: Gamma 자료 with η_2



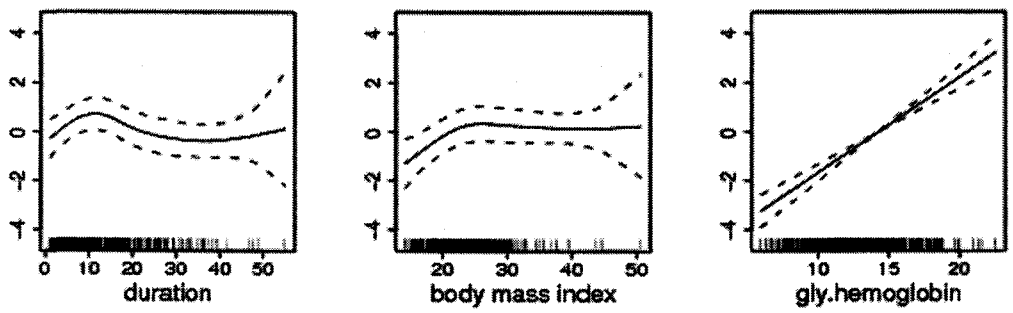
<그림 4.9> q 의 값이 근사해의 일관성에 미치는 영향: Gamma 자료 with η_3

5. 예제

이제 앞 절들에서 제시한 방법들을 실제 자료들에 적용시켜 본다. 이 절은 실제 자료의 분석이 아니라 이 논문에서 제시한 q-차원의 근사해와 $q = n$의 정확해를 비교하는 데에 목적이 있다. 계산시간은 Kim and Gu(2004)에서 사용했던 같은 머신(workstation with Athlon MP 2800+ and 3 Gbytes random access memory)에서 측정되었다.

5.1 Diabetic Retinopathy 자료

이 자료는 Wisconsin Epidemiological Study of Diabetic Retinopathy(WESDR)로서, 1980년부터 시작된 남위스콘신 지역의 11개 카운티에서 진찰을 받고 있는 환자들의 코호트에 대한 역학 연구 자료이다. 이 중 부분자료인 699명의 환자들에 대한 첫 번째 follow-up에서 비만의 지속성(년), 글리코실화된 헤모글로빈(퍼센트), 체지방 인덱스, 그리고 망막증(Retinopathy)의 진전에 대한 binary indicator 등이 측정된 자료를 사용하였다. 망막증의 진전을 반응변수로 두고 Binomial regression으로 분석할 때 가장 적합한 모델은 additive 모델이었다.(Gu, 2002)



<그림 5> WESDR 자료: 실선은 fitted terms, 점선은 95% 베이지안 신뢰구간

그림 5에서 $q = n$ =669(옅은 색)일 때와 q=43 ≈(10)669^{2/9}(짙은 색)일 때 각각의 스무딩 스플라인은 실선으로, 95%의 베이지안 신뢰구간은 점선으로 표시하였다. 이 때 평활 모수는 3.2절에서 제시한 cross-validation score(3.3)를 최소화하는 최적 평활 모수로 하였다. 계산 시간은 각각 CPU시간으로 $q = n$=669일 때는 170s, q=43일 때는 약 6.99s이 걸렸다. 참고로, 전체 자료의 랜덤 부분집합을 이용하므로 같은 크기의 다른 부분집합을 사용할 때, 그리고 다른 머신을 사용할 때 계산시간이 약간씩 달라질 수 있다.

6. 결론

이 논문은 가우시안 자료에 대해 개발된 계산방법(Kim and Gu, 2004)에 병행하여, 비가우시안

자료, 특히 지수족 자료에 대하여 점근적 효율성을 유지하는 저차원의 근사해를 통하여 스무딩 스플라인을 계산하는 방법을 제시하였다. 또한 Bernoulli 자료에만 국한되었던 평활 모수의 선택방법을 Bernoulli 자료 외에 다른 지수족 자료에 대하여도 고려하였다.

저차원 근사해를 이용하는 아이디어는 새로운 것은 아니지만, 이러한 아이디어를 스무딩 스플라인에 적용시켜 점근적 분석과 함께 수치적 모의실험을 통하여 근사해 공간의 차원을 구체적으로 결정하는 방법을 제시한 것은 새로운 것이며(Kim and Gu, 2004), 이 논문은 비가우시안 자료에 대하여 이러한 방법을 적용시켰다. 좀더 복잡하고 세심한 $\{z_j\}_{j=1}^{q}$의 선택은 보다 작은 q의 크기를 결정할 수 있겠지만, 이 논문이 제시하는 랜덤 선택은 항상 쉽고 보편적으로 사용할 수 있다는 장점이 있다.

가우시안 자료와는 달리 비가우시안 자료에 대한 스무딩 스플라인의 계산에서는 평활 모수를 계산하는 방법이 폭넓게 연구되어 있지 않다. 기존에 제시되었던 cross-validation score인 AGACV score는 오직 Bernoulli 자료에 대하여서만 응용되어 그 성능이 확인되어 있었다. 그러나 이 score의 이론적 정당화(theoretical justification)는 아직 밝혀지지 않고 있는 것으로 알고 있다. 그럼에도 불구하고 평활모수의 대역적 자동선택방법인 AGACV의 경험적 성능은 다양한 자료에 대하여 확인될 필요가 있다. 이 논문에서는 기존의 score의 응용범위를 다른 지수족으로 확장시켜서 그 성능을 확인하였다. 그 결과, 간단한 수정 또는 새로운 접근이 요구되어, 각 지수족에 알맞은 AGACV의 가능한 변형된 scores를 제시하였다. 그리고 일련의 모의실험과 예제 자료를 통하여 제시된 scores들의 성능을 확인하였다. 실제로 더 많은 다양한 테스트 함수들과 SNRs를 가지고 많은 모의실험을 수행하였고 결과들은 일관되게 나왔다.(Kim, 2003) 여기서 각 score들에 적용된 α를 삽입하는 간단한 수정법에서 제시된 α의 값을 결정하는 경험적 방법은 이론적인 정당화와 좀더 일반적인 방법의 개발이 후속연구의 주제가 될 수 있다. 하지만 이 논문은 충분히 많은 모의실험을 통하여 경험적 α의 값을 구체적으로 제시하고 있다.

이 논문이 제시하는 계산방법과 평활 모수의 결정방법을 다른 지수족 분포 자료뿐만 아니라 중도절단 생존 자료에 응용하는 것도 후속연구의 주제가 될 수 있다.

참고문헌

[1] Cox, D.D. and O' Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annuls of Statistics* 18, 124-145

[2] Green, P. J. and Yandell, B. (1985). Semi-parametric generalized linear models. In R. Gilchrist, B. Francis, and J. Whittaker (Eds.), *Proceedings of the GLIM85 Conference*, pp. 44-55. Berlin: Springer-Verlag.

[3] Gu, C. (1990). Adaptive spline smoothing in non Gaussian regression models. *Journal of American Statistical Association* 85, 801-807.

[4] Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational Graphics and Statistics* 1, 169-179.

[5] Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

[6] Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression: General formulation and efficient approximation. *Canadian Journal of Statistics* 30, 619-628.

[7] Gu, C. and Qiu, C. (1994). Penalized likelihood regression: A simple asymptotic analysis. *Statistical Sinica* 4, 297-304.

[8] Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation. *Statistical Sinica* 13, 811-826.

[9] Gu, C. and Xiang, D. (2001). Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational Graphics and Statistics* 10, 581-591.

[10] Kim, Y.-J. (2003), Smoothing splines regression: Scalable computation and cross-validation, Ph.D. thesis, Purdue University, West Lafayette, IN, USA

[11] Kim, Y.-J. and Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of Royal Statistical Society Series B* 66, 337-356.

[12] O' Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of American Statistical Association* 81, 96-103.

[13] Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *Applied Statistics* 27, 26-33.

[14] Wahba, G. (1990). *Spline Models for Observational Data*, Vol 59 of CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM.

[15] Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistical Sinica* 6, 675-692.

[2005년도 8월 접수, 2005년도 10월 채택]