The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 23–32

Penalized Likelihood Regression with Negative Binomial Data with Unknown Shape Parameter

Young-Ju Kim[1)]

Abstract

We consider penalized likelihood regression with data from the negative binomial distribution with unknown shape parameter. Smoothing parameter selection and asymptotically efficient low dimensional approximations are employed for negative binomial data along with shape parameter estimation through several different algorithms.

Keywords: Negative binomial; penalized likelihood; shape parameter; smoothing parameter.

1. 서론

음이항분포 자료 $Y \sim NB(\nu, p)$에 대하여 $\eta(x) = \log\{p(x)/(1-p(x))\}$로 두면 확률밀도함수 $\{\Gamma(\nu+y)/(y!\Gamma(\nu))\}p^{\nu}(1-p)^{y}$에 대하여 음의 로그우도함수 $l(\eta; Y) \propto (\nu + Y)\log(1+\exp(\eta)) - \nu\eta$를 가진다. 여기서 ν는 알려지지 않은 형태모수이며 $\eta(x)$는 공변량 x에 의존하는 함수이다. 우리는 비모수 함수 추정법인 벌점우도회귀를 이용하여 $\eta(x)$를 추정하고자 한다. 벌점우도회귀는 함수의 굴곡성 (flexibility) 을 최대한 허용하는 방법으로서 추정함수를 고차원 또는 무한차원의 함수공간으로 허용하는 비모수 추정법이다. 벌점우도회귀법은 다음과 같은 벌점우도 범함수를 최소화하는 해로 함수 η를 추정한다.

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

이 때, $J(\eta)$ 는 η의 거친 정도에 대한 벌점도 범함수이고, λ는 η의 자료에 대한 적합도 (goodness of fit) 와 평활도 (smoothness) 의 상충관계 (trade-off) 를 조절하는 평활 모수 (smoothing parameter) 이다. 벌점도함수가 $\int \ddot{\eta}^2 dx$일 때 (1.1)의 최소해를 큐빅 스무딩 스플라인 (cubic smoothing spline) 이라고 한다 (여기서 $\ddot{\eta}$는 η의 2번째 도함수이다.). (1.1) 의 최소해는 무한 차원 공간 $H \subseteq \{\eta : J(\eta) < \infty\}$ 에 존재한다.

1) Assistant Professor, Department of Information Statistics, Kangwon National University, 192-1 Hyoja-Dong, Chuncheon, Kangwon-do 200-701, Korea.
E-mail : ykim7stat@kangwon.ac.kr

기존의 벌점우도회귀에서 최소해의 계산방법은 자료의 수만큼의 기저를 요구하기 때문에 대용량 자료에 대하여 최소해 계산의 현실적 어려움이 있다. 이러한 최소해 계산의 현실적 어려움을 극복하기 위한 방안으로 Gu와 Kim (2002), Kim과 Gu (2004), Kim (2005) 는 벌점우도회귀에서의 최소해를 저차원 함수공간으로의 근사시키는 방법을 제안하고 각각 가우시안 자료와 세 가지 지수족 자료에 대하여 적용시켰다. 이러한 저차원 근사해의 차원 q 는 작아질수록 계산의 효율성은 높아지지만 비모수적 성질을 잃지 않기 위하여 충분히 큰 q가 동시에 요구된다. 우리는 음이항분포 자료에 대하여 이러한 저차원 함수공간으로의 근사해를 계산하고자 한다.

벌점우도회귀에서 추정함수의 평활도를 결정하는 평활 모수 (smoothing parameter)의 선택방법은 추정함수의 성능을 결정하므로 중요한 문제가 된다. Kim (2005) 에서 제시한 세 가지 지수족 자료에 대한 평활모수 선택방법과 더불어 Bernoulli 자료에만 적용되었던 Gu와 Xiang (2001) 의 방법이 음이항분포에 대하여도 작동하는지를 확인할 필요가 있다. 특히 음이항분포의 형태모수 (shape parameter) 가 알려지지 않는 경우 형태모수의 추정을 함께 고려한다.

2. 벌점우도회귀

벌점우도회귀법을 이용한 함수추정에서 식 (1.1) 의 최소해의 함수공간 H는 힐버트 공간 (Hilbert space) 이며 유한차원 m인 영공간 $N_J = \{\eta : J(\eta) = 0\}$ 을 가진다고 가정된다. H에서의 내적 $< \cdot, \cdot >$ 에 대하여 $< R(x, \cdot), f(\cdot) >= f(x)$를 만족하는 비음정치 함수 $R(\cdot, \cdot)$를 RK (Reproducing Kernel) 라고 하고 이러한 비음정치 함수를 가지는 함수공간 H를 RKHS (Reproducing Kernel Hilbert Space) 라고 한다 (Gu, 2002).

무한차원 공간 H에서의 최소해 계산은 현실적으로 불가능하다. 이를 극복하는 방안으로 RKHS의 성질을 이용하여 식 (1.1) 의 최소해가 유한차원 함수공간인 $H_n = N_J \oplus \text{span}\{R_J(x_i, \cdot), i = 1, \ldots, n\}$, R_J는 RK, 에서 표현된다는 사실이 잘 알려져 있으며 $O(n^3)$의 계산이 요구된다. Gu와 Kim (2002) 는 식 (1.1) 의 최소해를 적당한 정수 $q \leq n$에 대하여 $q-$차원의 함수공간 H_q로 근사시킨 근사해의 점근적 수렴률이 정확해와 같아진다는 것을 보였다. 이 때 q의 오더는 $O(n^{2/(pr+1)+\epsilon})$, $p \in [1, 2]$, $r > 1$, $\forall \epsilon > 0$, 으로 적당히 낮게 조절되어야 한다. $p \in [1, 2]$는 η의 평활도 (smoothness) 에 따라 결정되며, 평활도가 높을수록 2에 가깝다. 예를 들어, $\int (\eta^{(4)})^2 dx < \infty$이면 $p = 2$로 둔다 (여기서 $\eta^{(4)}$는 η의 4번째 도함수이다.). r은 벌점도 도함수 $J(\eta)$에 의해 결정되는 평활도를 나타내는 값으로서 큐빅 스플라인에 대하여 $r = 4$가 된다 (Gu와 Kim, 2002, 2절의 조건 2 참조). 벌점도 함수 $J(\eta) = \int \ddot{\eta}^2 dx$이고 추정함수의 평활도가 충분히 높다고 가정할 때 q의 오더는 $O(n^{2/9})$가 되며 $q-$차원 근사해의 계산은 $O(nq^2)$이 된다. 크기가 $q \leq n$인 랜덤 부분집합 $\{z_j, j = 1, \ldots, q\} \subseteq \{x_i, i = 1, \ldots, n\}$에 대하여 $H_q = N_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$에서의 최소해는 다음과 같이 표현될 수 있다.

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_J(z_j, x). \tag{2.1}$$

여기서, $\{\phi_\nu\}_{\nu=1}^m$는 영공간 N_J의 기저이고, $R_J(\cdot,x)$는 RK이며 $d=(d_1,\ldots,d_m)^T$와 $c=(c_1,\ldots,c_q)^T$는 실계수벡터이다. $q=n$일 때 식 (2.1) 은 식 (1.1) 의 정확해가 된다. 예를 들어, $\chi=[0,1]$에서 $J(\eta)=\int\ddot{\eta}^2dx$일 때, $N_J=\text{span}\{1,k_1(x)\}$, $k_1(x)=x-0.5$인 $H=N_J\oplus H_J$에서의 큐빅 스플라인을 얻을 수 있다. $H_J=\{\eta:\int_0^1\eta dx=\int_0^1\dot{\eta}dx=0,J(\eta)<\infty\}$에서의 RK는 $R_J(x_1,x_2)=k_2(x_1)k_2(x_2)-k_4(x_1-x_2)$가 된다. 여기서 $k_\nu=B_\nu/\nu!$이고 B_ν는 Bernoulli 다항식이다 (Gu, 2002).

벌점우도회귀에서 평활 모수는 최소해의 성능을 결정하므로 평활모수의 선택방법은 중요한 문제가 된다. 기존의 Gu와 Xiang (2001) 의 방법은 Bernoulli 자료에만 적용되었으며 Kim (2005) 는 세 가지 지수족 자료에 대한 평활모수 선택방법을 제시하였다. 음이항분포 자료의 경우, 특히 형태모수 (shape parameter) 가 알려지지 않는 경우에는 Gu와 Xiang (2001) 의 방법을 적용할 때 형태모수의 추정을 함께 고려하여야 한다.

3. 계산

저차원 근사해 식 (2.1) 의 계산문제는 크게 세 가지로 나누어 생각할 수 있다. 첫째는 최적 평활모수의 선택방법이고 두 번째는 음이항분포의 형태모수의 추정문제이다. 아래에서 제시될 계산 알고리듬에서 형태모수의 추정은 최적 평활모수의 계산과 관련되므로 이 문제들을 동시에 다루고자 한다. 마지막으로 저차원 근사최소해의 함수공간 H_q의 차원을 결정하는 문제이다. 근사최소해의 계산은 $O(nq^2)$이므로 최소해를 적절한 크기의 $q\le n$에 대한 H_q로 제한하면 최소해의 계산력을 높일 수 있다. 본 논문에서는 벌점도 함수 $J(\eta)=\int\ddot{\eta}^2dx$를 사용하여 큐빅 스플라인을 계산하였다.

3.1. 평활모수와 형태모수

벌점우도회귀에서 근사 최소해 식 (2.1) 의 계산은 Kim (2005) 와 같이 로그우도함수의 이차근사를 이용하면 가우시안의 로그우도함수 형태로 나타나므로 Kim과 Gu (2004) 의 방법을 이용할 수 있다. 이 때 고정된 평활 모수에 대하여 Newton iteration을 이용하여 근사최소해를 구하고, 참해와 근사최소해의 Kullback-Leibler 거리의 cross-validation인 Gu와 Xiang (2001) 의 AGACV (Alternative Generalized Approximate Cross-Validation) score를 최소화하는 최적 평활 모수를 찾는다.

$u(\eta;Y)=dl/d\eta$, $w(\eta;Y)=d^2l/d\eta^2$로 두면, 고정된 평활 모수 λ에 대하여 $\tilde{\eta}(x_i)$에서 로그우도함수 $l(\eta(x_i);Y_i)$의 이차근사는 다음과 같이 벌점가중최소제곱 범함수 형태로 나타난다.

$$\frac{1}{n}\sum_{i=1}^n\tilde{w}_i(\tilde{Y}_i-\eta(x_i))^2+\lambda J(\eta). \tag{3.1}$$

여기서 $\tilde{Y}_i=\tilde{\eta}(x_i)-\tilde{u}_i/\tilde{w}_i$이고, $\tilde{u}_i=(\nu+Y_i)\tilde{p}_i-\nu$, $\tilde{w}_i=(\nu+Y_i)\tilde{p}_i(1-\tilde{p}_i)$, $\tilde{p}_i=e^{\tilde{\eta}(x_i)}/(1+e^{\tilde{\eta}(x_i)})$이다. 식 (3.1) 에 η의 저차원 표현식 (2.1) 을 대입하면 다음과 같은 식을 얻을 수 있으며,

$$(\tilde{Y}-Sd-Rc)^T\tilde{W}(\tilde{Y}-Sd-Rc)+n\lambda c^TQc. \tag{3.2}$$

이 때, $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^T$, $\tilde{W} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$, S는 (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, 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, $\hat{Y}_w = A_w(\lambda)Y_w$가 되고 A_w은 평활 행렬이다. 식 (3.2) 을 d와 c에 대하여 최소화시켜서 큐빅 스플라인 $\eta_{\lambda,\tilde{\eta}}$을 계산한다. Newton iteration을 이용하여 벌점가중최소제곱 범함수 형태의 최소해 $\eta_{\lambda,\tilde{\eta}}$로 $\tilde{\eta}$을 업데이트시켜 나간다. 이렇게 수렴되는 추정함수가 구하는 해가 된다 (Kim, 2005). 그리고 AGACV score $V^*(\lambda)$를 최소화하는 평활모수로 최적평활모수를 구한다.

$$V^*(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} l_i(\eta_\lambda(x_i); Y_i) + \frac{\mathrm{tr}(A_w\tilde{W}^{-1})}{n - \mathrm{tr}(A_w)}\frac{1}{n}\sum_{i=1}^{n} h_i(-\tilde{u}_i). \tag{3.3}$$

여기서 $h_i = -Y_i\tilde{p}_i$, $\tilde{W} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$, A_w는 평활 행렬이다 (Gu와 Xiang, 2001).

음이항분포의 형태모수 ν가 알려지지 않은 경우에는 적절한 방법을 이용하여 ν를 추정해야 한다. 고정된 평활모수에 대하여 최소해의 Newton 업데이트 시 형태모수가 관여되므로 형태모수의 추정은 중요해진다. 음이항분포에서 형태모수를 알지 못할 때 형태모수를 추정하는 일반적인 방법으로 적률법과 최대우도법 등이 있다 (Van de Ven, 1993). 우리는 음이항분포의 형태모수의 최대우도추정방법을 사용하여 형태모수를 추정하는 몇 가지 계산알고리듬을 고려하였다.

- *알고리듬 1.* 고정된 평활모수에 대하여 형태모수와 η_λ를 반복 추정한다. 고정된 형태모수에 대하여 η의 Newton update를 계산하고 η의 현재 추정치를 이용하여 ν에 대한 로그우도함수를 최대화하는 ν로 추정하는 방법을 반복한다. 고정된 평활모수에 대하여 η_λ와 ν의 계산은 inner loop로, AGACV score 식 (3.3) 를 이용한 최적 평활모수의 계산은 outer loop로 이루어진다.

- *알고리듬 2.* 고정된 형태모수 ν에 대하여 AGACV score 식 (3.3) 를 최소화하는 최적 평활모수를 구하여 $\hat{\eta}$를 계산하는 방법과 현 추정치 $\hat{\eta}$에서 형태모수 ν의 최우추정량을 구하는 방법을 반복한다.

- *알고리듬 3.* 적당한 범위 안에서 형태모수 ν를 고정시킨 후 식 (3.3) 을 이용하여 $\hat{\eta}$를 구한다. 주어진 $\hat{\eta}$에 대하여 ν에 대한 로그우도함수 $l(\nu; Y, \hat{\eta}) = \log\Gamma(\nu + Y) - \log\Gamma(\nu) + (Y + \nu)\log(1/(1 + \exp(\hat{\eta})) - \nu\hat{\eta}$을 최대화하는 ν로 형태모수의 추정치를 구한다. 이 때 $\hat{\eta}$는 고정된 형태모수에 의존하므로 ν의 로그우도함수에 포함되었다.

알고리듬 1은 고정된 평활모수 안에 형태모수의 추정이 nested 되어 있고 알고리듬 2는 최적평활모수에 대한 $\hat{\eta}$와 형태모수 ν의 반복추정법이다. 알고리듬 2는 Newton iteration을 이용하여 ν를 추정하는 반면 알고리듬 3은 grid search를 이용하므로 상대적으로 사용이 쉽다.

위에서 제시한 알고리듬의 성능을 보기 위하여 Kim (2005) 에서와 마찬가지로 서로 다른 세 가지 테스트 함수와 세 가지 SNR (signal to noise ratio) 을 이용하여 생성된 자

료를 가지고 다음과 같은 모의실험을 수행하였다.

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

각 테스트 함수에 대하여 각 표본의 크기 $n = 100$에 대하여 $x_i = (i-0.5)/n$, $i = 1, \ldots, n$을 사용하였다. 각 ν에 대하여 평균 $\mu \in [0.2, l]$이 되도록 테스트 함수들을 조절하였다. l은 각각 5, 10, 20 중 하나를 선택하여 생성된 100개의 반복표본을 이용하였다. 3.1절에 제시한 서로 다른 알고리듬을 이용하여 얻은 형태모수의 추정치와 $\hat{\eta}$에 대한 KL (Kullback-Leibler) 손실을 계산하였다.

$$\begin{aligned} L(\lambda) =& KL(\eta_0, \eta_\lambda) \\ =& \frac{\nu}{n}\sum_{i=1}^{n}\{(1+e^{-\eta_0(x_i)})(\log(1+e^{\eta_\lambda(x_i)}) - \log(1+e^{\eta_0(x_i)})) - \eta_\lambda(x_i) + \eta_0(x_i)\}. \end{aligned}$$

그림 3.1은 위에서 아래로 각각 테스트 함수 η_1, η_2, η_3에 대하여 모의실험을 통한 알고리듬 1, 2, 3의 성능이 요약되었다. 그림 3.1의 왼쪽 세 그래프는 형태모수 $\nu = 2$ 일 때 세 가지 SNR에 대하여 알고리듬 1 (굵은 상자), 알고리듬 2 (중간 굵기의 상자) 와 알고리듬 3 (얇은 상자) 으로부터 계산된 형태모수의 추정치의 상자그림들이다. 가로 실선은 형태모수의 참값을 나타낸다. 알고리듬 3을 이용한 모의실험에서 형태모수 ν의 범위는 0.5(0.2)3.5 로 선택하였다. 형태모수의 범위는 사용자가 임의로 설정하여도 좋다. 그림 3.1의 오른쪽 세 그래프는 각 알고리듬에서 계산된 형태모수의 추정치를 이용하여 얻은 $\hat{\eta}$의 KL 손실을 상자그림으로 표현하였다.

그림 3.1의 왼쪽 세 개의 그래프에서 보는 바와 같이 알고리듬 1은 형태모수 ν를 과소 추정하지만 산포는 작게 나타났다. 알고리듬 2는 전반적으로 과대추정의 경향을 띠며 분산도 크게 나타났다. 알고리듬 3은 위의 두 알고리듬에 비하여 상대적으로 참값에 가까운 추정치를 제공하였다. 반면에 알고리듬 3은 알고리듬 1에 비하여 추정치의 분산이 크게 나타났다. 여기서는 보이지 않았지만 $\nu = 4$일 때도 $\nu = 2$일 때와 질적으로 동일한 결과를 얻었고 특히 알고리듬 3은 $\nu = 4$일 때 알고리듬 2에 비해 분산이 작게 나타났다. 따라서 사용자의 판단 하에 알고리듬 1이나 알고리듬 3을 사용하여 형태모수의 추정치를 얻을 수 있다. 그림 3.1의 오른쪽 세 그래프는 서로 다른 알고리듬에 따른 $\hat{\eta}$의 성능을 나타내며 형태모수의 추정 성능과 질적으로 비례하는 것을 볼 수 있었다. 참고로, 여기에서는 생략되었지만, 계산의 용이함 때문에 선호되는 형태모수의 적률 추정량은 표본분산이 표본평균보다 큰 경우에 주로 사용되는데 여기서는 심각한 과소추정 경향이 나타났으므로 적절한 추정방법이 될 수 없었다.

3.2. q의 계산

참함수는 충분히 매끄럽다는 가정 아래 $J(\eta) = \int \ddot{\eta}^2 dx$로 두고 Gu와 Kim (2002) 에서 제시한 q의 오더 $O(n^{2/9})$를 이용하여 저차원 근사해 (2.1) 의 차원을 결정하고자 한

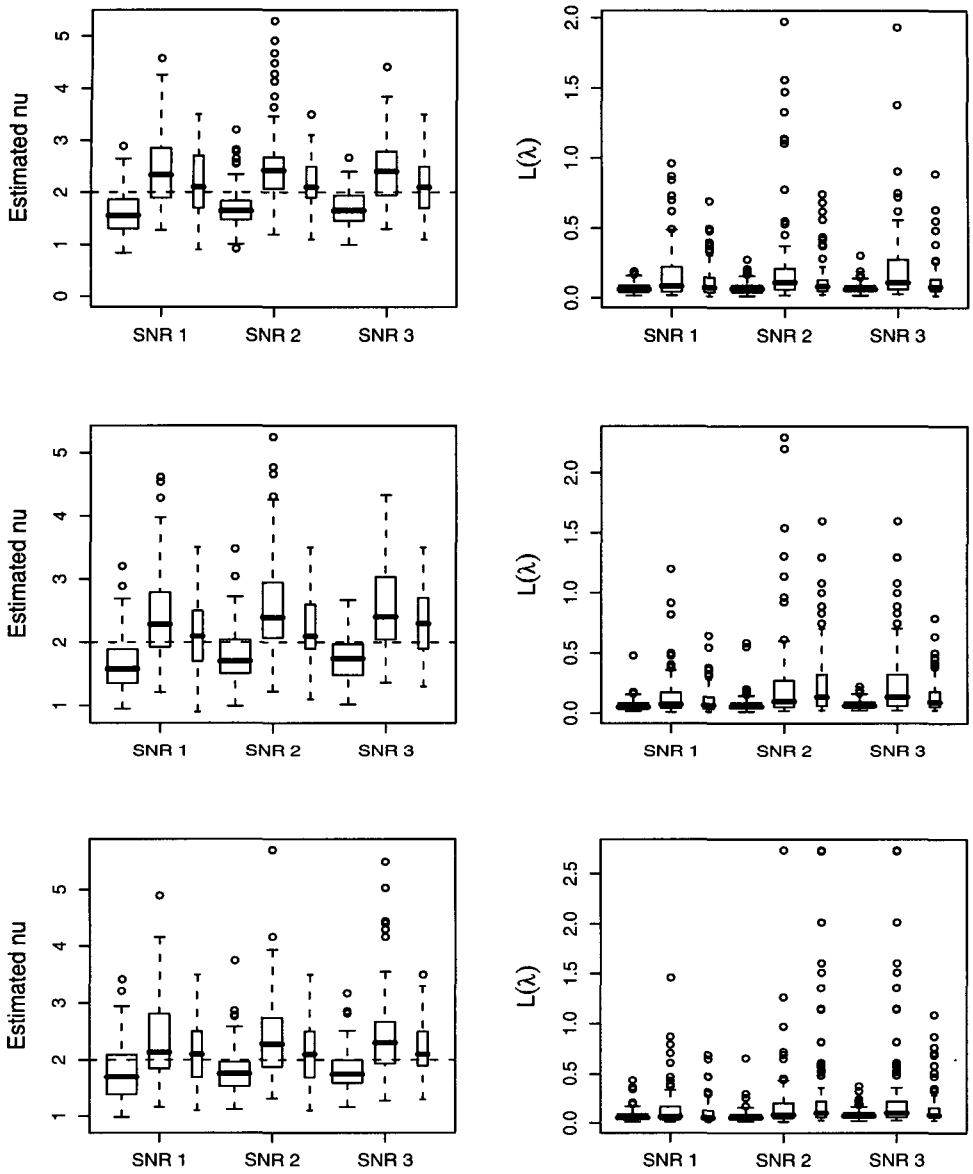

그림 3.1: 위에서부터 아래로 η_1, η_2, η_3에 대한 형태모수의 추정치 (왼쪽) 와 $\hat{\eta}$의 KL손실 (오른쪽) 을 통한 알고리듬의 성능비교: 각 η의 서로 다른 SNR, $\nu = 2$를 이용한 알고리듬 1 (굵은 상자), 알고리듬 2 (중간굵기의 상자) 와 알고리듬 3 (얇은 상자) 의 성능

다. $q = kn^{2/9}$으로 두고 모의실험을 통하여 k의 값을 결정할 수 있다. 모의실험을 위하여 테스트함수 η_1와 $\nu = 2$에 대하여 서로 다른 세 개의 SNR 중 하나를 선택하여 크기가 $n = 100$, $n = 500$인 표본을 각각 추출했다. ν의 값을 알지 못할 경우에는 3.1절의 알고리듬을 이용하여 추정한 값을 사용하였다. 각 표본에 대하여 각 $\nu = 5(1)15$를 사용하여 크기가 $q = kn^{2/9}$인 30개의 서로 다른 랜덤 표본 $\{z_j\} \subset \{x_i\}$을 추출하여 30개의 저차원 근사해 (2.1) 와 $q = n$에 대한 정확해를 계산하고 서로 다른 k의 값에 대하여 Kullback-Leibler 손실을 계산하여 상자그림으로 나타내었다.

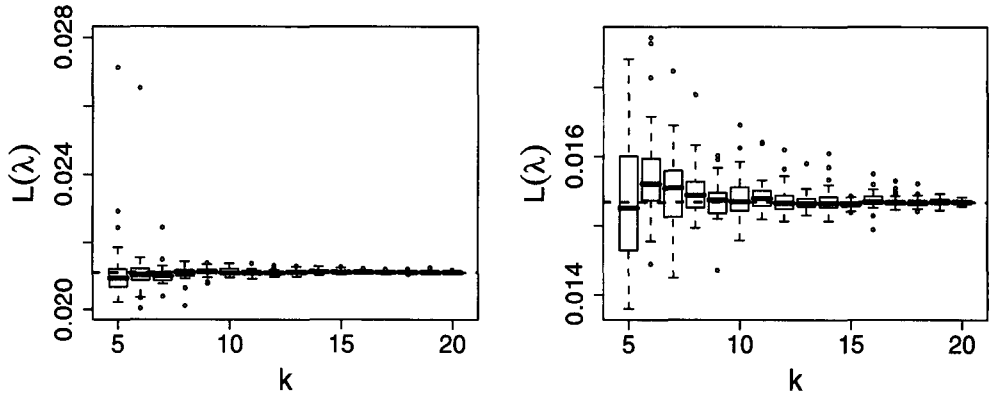

그림 3.2: $\nu = 2$가 알려져 있을 때 q의 값과 근사해의 일관성: $n = 100$ (왼쪽), $n = 500$ (오른쪽)

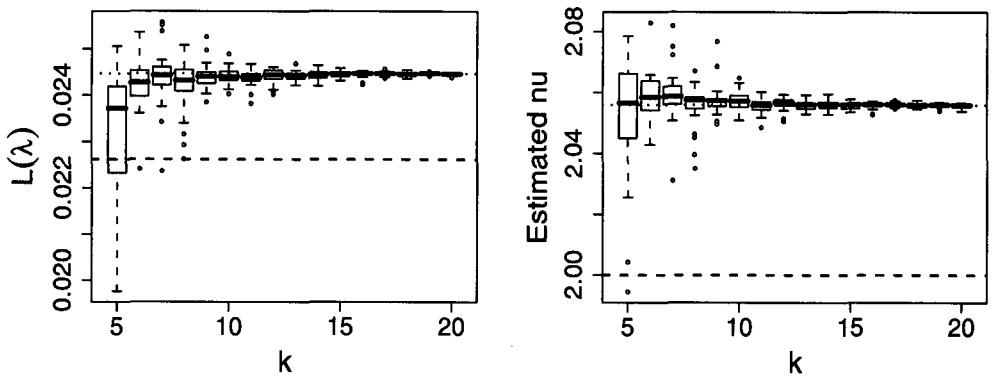

그림 3.3: $\nu = 2$일 때 q의 값 및 알고리듬 1을 이용한 형태모수의 추정치

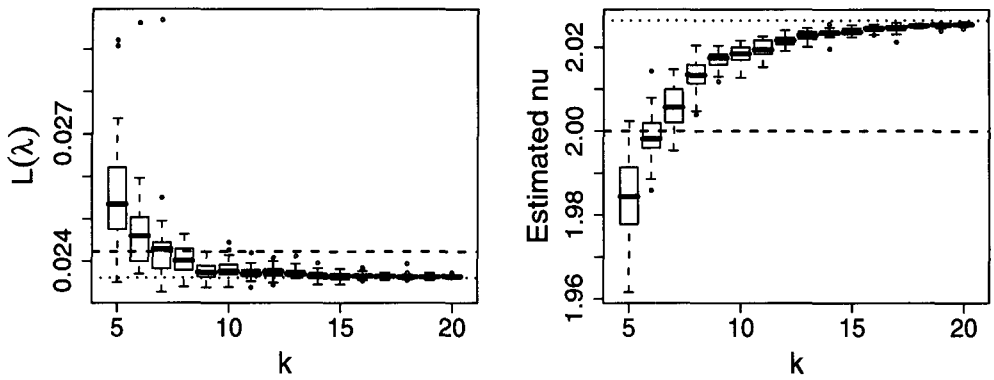

그림 3.4: $\nu = 2$일 때 q의 값 및 알고리듬 2를 이용한 형태모수의 추정치

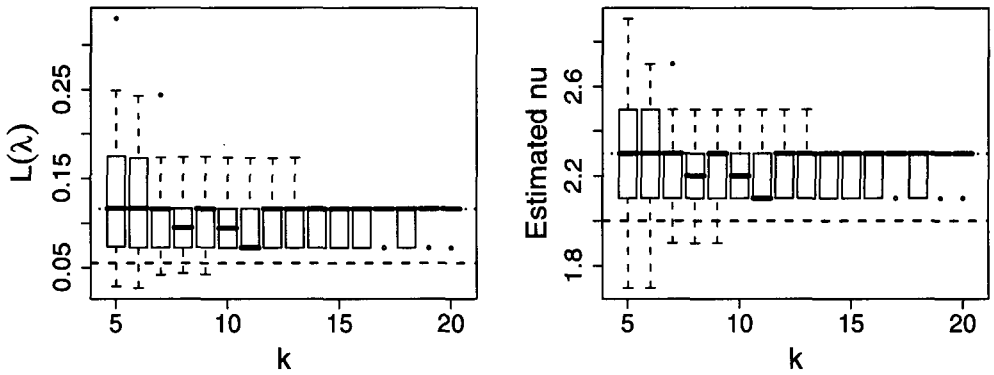

그림 3.5: $\nu = 2$일 때 q의 값 및 알고리듬 3을 이용한 형태모수의 추정치

그림 3.2은 형태모수 $\nu = 2$가 알려져 있는 경우 $n = 100$과 $n = 500$일 때 서로 다른 k의 값에 대하여 계산된 KL 손실을 상자그림으로 나타내었다. 그림 3.3, 그림 3.4, 그림 3.5는 형태모수 $\nu = 2$가 알려져 있지 않은 경우 $n = 100$일 때 서로 다른 ν의 값에 대하여 계산된 KL 손실과 3.1절에서 제시한 서로 다른 알고리듬을 이용하여 계산한 형태모수 추정치의 상자그림을 보여주고 있다. 그림 3.3, 그림 3.4, 그림 3.5의 왼쪽 상자그림은 큐빅 스플라인들에 대한 KL 손실을 보여준다. 두 점선은 $q = n$일 때의 최소해에 대한 KL 손실을 나타낸다; 굵은 점선은 형태모수의 참값 $\nu = 2$일 때의 KL 손실을, 가는 점선은 3.1절의 알고리듬을 이용하여 추정한 $\hat{\nu}$을 이용하여 계산한 KL 손실을 나타낸다.

각 그림의 오른쪽 상자그림은 서로 다른 k에 대하여 3.1절의 알고리듬을 이용하여 계산한 형태모수의 추정치의 상자그림이다. 굵은 점선은 형태모수의 참값 $\nu = 2$를, 가는 점선은 $q = n$일 때 3.1절의 알고리듬을 이용하여 추정한 $\hat{\nu}$을 나타낸다. 각 그림에서 보는 것과 같이 k가 커질수록 KL 손실의 상자의 높이가 줄어들면서 $q = n$일 때의 최소해에 대한 KL 손실 (가는 점선) 로 가까워지는 것을 볼 수 있다. 그리고 k의 값이 $10 \sim 12$ 정도가 되면 상자들의 높이가 안정적이 되는 것을 확인할 수 있다. 또한 형태모수의 추정치 역시 k가 커질수록 안정적으로 나타나는 것을 볼 수 있으며 이러한 경향은 형태모수의 추정 알고리듬에 따라 크게 다르지 않았다. 위의 결과는 테스트함수 η_1와 $\nu = 2$, $l = 5$ 일 때 생성된 하나의 표본에 대한 결과이며, 같은 조건의 다른 랜덤표본을 선택하여 동일한 모의실험을 하면 나타나는 수치는 다를 수 있으나 동질의 결과를 얻을 수 있다. 여기에 나타나지 않은 다른 테스트 함수들과 다른 SNRs, 다른 형태모수, 그리고 다른 $n > 100$에 대하여도 질적으로 동일한 결과를 얻었으므로 여기서는 생략하기로 한다.

4. 결론

이 논문은 벌점우도회귀를 이용한 음이항분포 자료의 적합문제를 다루었다. 형태모수가 알려지지 않는 경우 평활모수의 선택방법과 함께 형태모수의 추정방법을 제시하였다. 점근적 효율성을 유지하는 저차원 근사를 이용한 큐빅 스무딩 스플라인의 빠른 계산법을 적용하였고 이에 따른 형태모수의 추정치도 확인하였다. 형태모수가 알려지지 않은 음이항분포 자료의 저차원 근사해를 이용한 벌점우도회귀분석법은 산림학이나 사고통계, 의학분야에서 나타나는 대용량 자료를 다루는 실제 분석가들에게 더 효율적이고 빠른 계산법으로 제공될 것으로 기대한다.

참고문헌

Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics*, **1**, 169–179.

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

Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression: general formulation and efficient approximation. *The Canadian Journal of Statistics*, **30**, 619–628.

Gu, C. and Xiang, D. (2001). Cross-validating non-Gaussian data: generalized approximate cross-validation revisited. *Journal of Computational and Graphical Statistics*, **10**, 581–591.

Kim, Y.-J. (2005). Computation and smoothing parameter selection in penalized likelihood regression, *The Korean Communications in Statistics*, **12**, 743–758.

Kim, Y.-J. (2006) Bayesian confidence intervals in penalized likelihood regression, *The Korean Communications in Statistics*, **13**, 141–150.

Kim, Y.-J. and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society*, Ser. B, **66**, 337–356.

Van de Ven, R. (1993). Estimating the Shape parameter for the Negative Binomial distribution. *Journal of Statistical Computation and Simulation*, **46**, 111–123.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, Ser. B, **40**, 364–372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society*, Ser. B, **45**, 133–150.

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

[Received July 2006, Accepted November 2006]