희박다항분포확률에 대한 국소최대우도 추정량

백 장선1)

요 약

$\boldsymbol{p}=(p_1, p_2, \cdots, p_k)^T$의 확률벡터를 가진 다항분포로부터 관측된 칸 돗수(cell frequency) 벡터가 $\boldsymbol{N}=(N_1, N_2, \cdots, N_k)^T$ 이며 $\sum_{j=1}^{k} N_j = n$ 이라 하자. 총돗수 n이 칸의 총갯수 k에 비하여 상대적으로 매우 작을 때 이러한 이산형 자료를 희박다항분포자료(sparse multinomial data)라 한다. 이러한 희박다항분포자료의 칸들이 순서화 되어 있을 때 우리는 i번째 칸의 확률 p_i를 돗수 추정량 N_j/n 들을 평활함으로써 추정 할 수 있다. Aerts, et al.(1997)과 Baek(1998) 등에 의해 제안된 국소최소제곱기준에 근거한 국소다항커널추정량은 희박점근일치성의 좋은 성질을 가짐에도 불구하고 확률추정치가 음수값을 가질 수 있는 단점을 내포하고 있다. 본 연구에서는 이러한 단점을 극복하기 위하여 국소최대우도 기준에 근거한 새로운 커널추정량을 제안하고, 그것의 점근적 성질을 연구하였다.

주요 용어 : 희박다항분포, 국소최대우도, 국소다항추정량

1. 서론

순서화된 범주를 갖는 범주형 자료를 분석하고자 할 때 그 자료가 각 범주에 대하여 $\{N_j\}$, $j=1, \cdots, k$ 의 칸 돗수로 구성되어 있다고 하자. 이때 k는 칸의 수이며 총돗수는 $n=\sum_{j=i}^{n} N_j$ 이다. 이러한 다항자료의 첫번째 관심은 칸 확률 $\boldsymbol{p}=\{p_j\}$를 추정하는 것이다. 우리가 추정하려는 칸 확률 p_i는 [0, 1]에서 연속인 함수 $f(\cdot)$에 의해 결정된다고 가정하자. 즉, $p_i = \int_{(i-1)/k}^{i/k} f(u)\, du$, $i=1, \cdots, k$이며 $f(\cdot)$는 $(t+1)$차 미분가능하다고 가정한다. 칸 확률은 일반적으로 돗수 추정량 $\overline{p_i}=N_i/n$ 을 사용하여 왔다. 그러나 이 추정량은 각 칸에서의 관측돗수가 많을 때 유용하며 희박한 다항자료에서는 일치성을 만족하지 못한다.

1) (500-757) 광주광역시 북구 용봉동 300 전남대학교 통계학과 부교수

따라서 인접한 칸의 정보를 차용한 비모수적 추정량들이 개발되었고 그 성능은 돗수 추정량 $\overline{p_i}$ 보다 나은 결과를 보여준다.

$x_l=(l-1/2)/k$, $l=1,\cdots,k$ 를 등간격을 갖는 각 구간의 중심점이라 하면, Aerts, et al.(1997)과 Baek(1998)에서 사용한 커널 추정량은 $(x_l, \overline{p_l})$, $l=1,\cdots,k$ 를 평활하여 구한 커널 추정량이었다. 그들이 사용한 기준은 회귀분석에서 즐겨 사용하는 국소최소제곱법이며, 따라서 i번째 칸 확률에 대한 추정량은 $\overline{p_i}$ 들을 t차 다항 함수로 평활시킨

$$\sum_{j=1}^{k}[\ \overline{p_j}-\beta_0-\ \cdots\ -\beta_t(x_i-x_j)^t\ ]^2K_h\ (\ x_i-x_j)$$

를 최소화하는 $\hat{\beta}=(\ \hat{\beta}_0,\cdots,\ \hat{\beta}_t)$중에서 상수항 $\hat{\beta}_0$이다. 이때 K는 커널함수이며 h는 평활모수이다. 그러나 이러한 국소최소제곱법에 의한 국소다항추정량은 회귀오차가 정규분포를 따른다는 회귀분석 상황에 보다 적합한 것이며 다항확률에 적용 됐을 때 음수값을 취할 수 있는 단점을 가지고 있다. 따라서 다항분포 자료에는 최소제곱법보다는 적절한 우도함수에 기반한 비음수의 추정량을 구축하는 것이 바람직 하다.

$\sum_{i=1}^{k}N_i=n$이 주어졌을 때 다항관측벡터 $(N_1,N_2,\cdots,N_k)$는 k개의 독립적인 포아송확률변수로 간주할 수 있다(Kendall & Stuart 1979, p449). 즉, $N_i\sim Poisson(np_i)$이며 N_i들은 독립적이다. k개의 독립적 포아송 확률변수 N_i의 로그우도함수 L은

$$L\ =\ \prod_{i=1}^{k}\frac{e^{-(np_i)}(np_i)^{N_i}}{N_i!}$$

이며, 로그우도함수는

$$\ln L=\sum_{i=1}^{k}\ [N_i\ln\ (np_i)\ -\ np_i-\ln\ (N_i!)]$$

이다. 이 로그우도함수에서 상수항을 제거하여도 최대화를 달성할 수 있으므로, 상수항 $\sum_{i=0}^{k}\ln(N_i!)$를 제거한 새로운 로그우도함수를

$$l=\sum_{i=1}^{k}\ [N_i\ln\ (np_i)\ -\ np_i\ ] \tag{1}$$

라고 하자.

모수적 일반화선형모형에서는 종종 회귀함수 $\mu(x)=E(Y|X=x)$의 변환함수

$\eta(x)=g(\mu(x))$를 선형으로 모형화한다. 이때 g를 링크함수라 한다. 특히 $g=(b')^{-1}$ 를 만족하는 g를 정준링크(canonical link)라 부른다.

실제상황에서 우도함수는 모르지만 평균과 분산의 관계는 알 수 있는 경우가 종종 존재한다. 이 경우 로그우도함수 $\ln q_{Y|X}(y)$ 대신 준우도함수(guasi-likelihood function) $Q(\mu(x),y)$를 이용하여 $\mu(x)$에 관한 추론을 수행한다. 만약 조건부 분산이 알려진 양의 함수 V에 대하여 $Var(Y|X=x)=V(\mu(x))$로 모형화 될 때, 이에 대응하는 준우도함수 $Q(w,y)$는

$$\frac{\partial}{\partial w} Q(w,y) = \frac{y-w}{V(w)}$$

를 만족하는 함수이다(McCullagh and Nelder 1989.) 준우도방법은 우도방법과 비슷하며 특히 우도함수가 알려져 있지 않을 때 대체하여 사용할 수 있는 방법이다. 또한 단일모수 지수족의 로그우도함수는 $V=b''\circ(b')^{-1}$인 준우도함수의 특별한 경우가 된다. 따라서 반응변수의 분포가 지수족에 속하고, 분산함수가 정확하게 규정되면 준우도추정방법은 최우추정법과 동일하다는 것을 알 수 있다.

자료 $(X_1, Y_1),\cdots,(X_k, Y_k)$를 이용하여 $\mu(X)$를 추정하기 위해 최대화 시켜야 하는 준우도함수는

$$\sum_{i=1}^{k} Q(\mu(X_i), Y_i) = \sum_{i=1}^{k} Q(g^{-1}(\eta(X_i)), Y_i)$$

이다.

포아송 반응변수 y의 경우 $Q(w,y) = y\ln w - w$이며, 정준링크는 $g(u)=\ln(u)$임을 확인할 수 있다. i번째 칸의 중심점 X_i와 그 칸의 돗수 N_i로 이루어진 우리의 자료 $\{(X_i,N_i)\}_{i=1}^{k}$에서, 반응변수 $Y_i=N_i$는 평균 $\mu(X_i)=np_i$를 가진 포아송분포를 따른다. 이 때의 정준링크는 $\eta(x)=g(\mu(x))=\ln(np)$형태를 갖는다. 그러므로 최대화를 위한 준우도함수는

$$l = \sum_{i=1}^{k} Q(g^{-1}(\eta(X_i)), N_i) = \sum_{i=1}^{k}[N_i\ln(np_i)-np_i] \qquad (2)$$

이다. (2)식은 로그우도함수 (1)과 정확히 일치한다.

2. 희박다항분포확률에 대한 국소최대우도 추정량

이제 $X=x$의 중심점을 갖는 칸에서의 칸 확률 p를 추정하기 위해 먼저 $\eta(x)=\ln(np)$에 대한 비모수적 추정량을 구한 뒤 그것을 역변환하는 전략을 취하기로 한다. 우리는 x점에서 국소적으로 t차 다항함수로서 $\eta(x)$를 추정하려고 한다. 이를 위해서는 자료를 x에 대해 중심화 시키고 x근처의 자료에 더 많은 중요성을 줄 수 있도록 우도함수를 가중화함으로써 가능하다. 즉, 커널 $K_h(X_i-x)$를 이용하여 가중화를 한다면 국소로그우도함수는

$$l^* = \sum_{i=1}^{k} Q\left(g^{-1}(\beta_0+\cdots+\beta_t(X_i-x)^t), N_i\right) K_h(X_i-x)$$
$$= \sum_{i=1}^{k}\left[N_i\{\beta_0+\cdots+\beta_t(X_i-x)^t\}-\exp\{\beta_0+\cdots+\beta_t(X_i-x)^t\}\right]K_h(X_i-x) \quad (3)$$

이다. $\hat{\beta}=(\hat{\beta}_0,\cdots,\hat{\beta}_t)$가 위의 국소로그우도함수를 최대화시키는 추정모수라 할 때, $\eta(x)=\ln(np)$에 대한 국소최대우도 커널추정량은 $\hat{\eta}(x\,;t,h)=\hat{\beta}_0$이며, 따라서 p에 대한 추정량은 $\hat{p}(x\,;t,h)=(1/n)\exp(\hat{\beta}_0)$이다. 따라서 추정치는 비음수를 보장한다.

p에 대한 추정량 $\hat{p}(x;t,h)=(1/n)\exp(\hat{\beta}_0(x;t,h))$에 대한 점근적 정규성은 **[정리 1]**과 같이 규명할 수 있다.

[정리 1] $t>0$이 홀수이고, $k\to\infty$에 따라 $h=h_k\to 0$, $kh^3\to\infty$이라고 하자. $x\in[h,\ 1-h]$일 때, 국소로그우도 추정량을 $\hat{p}(x;t,h)$라 하면,

$$\frac{\sqrt{nkh}}{\sqrt{p\int_A K_t(z\,;A)^2\,dz}}\left[\hat{p}(x;t,h)-p\right.$$
$$\left.-\left\{\int z^{t+1}K_t(z)\,dz\right\}\left\{\frac{\frac{\partial^{(t+1)}}{\partial x^{(t+1)}}\ln(np)}{(t+1)!}\right\}ph^{t+1}\{1+O(h)\}\right] \xrightarrow{d} N(0,1) \quad (4)$$

이다. 만약 x가 경계점일 때는 식 (4)의 $\int_A K_t(z\,;A)^2\,dz$와 $\int z^{t+1}K_t(z)\,dz$를 각각 $\int_{D_{x,h}} K_t(z\,;D_{x,h})^2\,dz$와 $\int_{D_{x,h}} z^{t+1}K_t(z:D_{x,h})^2\,dz$로 대체하면 역시 성립한다.

$t>0$가 홀수일 때 점근적 편의는 내부와 경계에서 동일한 차수의 크기를 갖는다. 점근적평균평방오차(AMSE ; asymptotic mean squared error)는 편의의 제곱과 분산의 합이므로 $x = x_i \in [h, 1-h]$에서의 추정량 $\hat{p}(x;t,h)$에 대한 AMSE는

$$\text{AMSE}(\hat{p}(x;t,h))$$

$$= \left\{\int z^{t+1} K_t(z)\, dz\right\}^2 \left\{\frac{S_{(t+1)}(f, f^{(1)}, \cdots, f^{(t+1)})}{(t+1)!\, k}\right\}^2 h^{2t+2}$$

$$+ \frac{\left\{\int_A K_t(z;A)^2\, dz\right\} f(x)}{nk^2h}$$

이다. $x = x_i$가 경계점일 때, 즉 $x = x_i \in [0, h) \cup (1-h, 1]$에서도 점근적 편의와 점근적 분산 모두 내부점일 때 와 동일한 차수를 갖고 있으므로, 경계점에서의 AMSE = $O(h^{2t+2}/k^2) + O((nkh)^{-1})$이다.

내부와 경계에서의 AMSE의 결과들을 통합하여 $\hat{p}(x;t,h)$의 전구간에서의 적합도의 척도인 MSSE(mean sum of squared error)를 다음과 같이 점근적으로 구할 수 있다. 내부구간을 $I = [h, 1-h]$, 경계구간을 $B = [0, h) \cup (1-h, 1]$이라 할 때, $\hat{\boldsymbol{p}} = (\hat{p}(x_1), \cdots, \hat{p}(x_k))$에 대하여

$$\text{MSSE}(\hat{\boldsymbol{p}})$$

$$\sim \frac{h^{2t+2}}{k} \left\{\int z^{t+1} K_t(z)\, dz\right\}^2 \frac{\int S_{(t+1)}(f(u), , \cdots, f^{(t+1)}(u))\, du}{\{(t+1)!\}^2}$$

$$+ \frac{1}{nkh} \left\{\int_A K_t(z;A)^2\, dz\right\}$$

이다.

한편 국소최소제곱법에 의한 칸 확률에 대한 국소다항커널 추정량의 MSSE는 점근적으로

$$\frac{h^{2t+2}}{k} \left\{\int z^{t+1} K_t(z)\, dz\right\}^2 \frac{\int (f^{(t+1)}(u))^2\, du}{\{(t+1)!\}^2} + \frac{1}{nkh} \int_A K_t(z;A)^2\, dz$$

이다(Aerts, et al 1997).

최대우도국소다항커널 추정량 $\hat{p}(x;t,h)$의 MSSE 내 점근적 분산 공헌부분은 국소최소제곱법에 의해 추정한 국소다항커널 추정량의 그것과 정확히 일치한다. 또한 점근적 편의 공헌부분은 국소최대우도방법에 의한 것이 조금 더 복잡하지만 두 방법 모두 동일한 차수 $O\left(\frac{h^{2t+2}}{k}\right)$를 갖고 있음을 확인할 수 있다.

$t>0$이 홀수인 경우 MSSE ($\hat{p}$)를 최소화 시키는 최적평활모수 h_{opt}는

$$h_{opt} = \left[\frac{\{(t+1)!\}^2 \int_A K_t(z;A))^2 \, du}{(2t+2)\left\{\int z^{t+1} K_t(z)\, dz\right\}^2 \left\{\int S_{(t+1)}(f(u), \cdots, f^{(t+1)}(u))\, du\right\} n} \right]^{1/(2t+3)}$$

이다. 따라서 h_{opt}을 사용했을 때 MSSE ($\hat{p}(\cdot\,;t,h_{opt})$) $= O\left(k^{-1}n^{-\frac{2t+2}{2t+3}}\right)$이다. 이것은 Hall and Titterington(1987)에 의해서 밝혀진 다항 칸 확률들이 $(t+1)$차까지 미분가능할때의 추정량에 대한 MSSE의 최적 수렴율과 동일하다. 그러므로 우리의 국소최대우도 추정량은 최적 MSSE 수렴율을 달성함을 알 수 있다.

참고문헌

[1]. Aerts, M. A. , Augustyns, I. and Janssen P.(1997), "Smoothing Sparse Multinomial Data using Local Polynomial Fitting, " *Nonparameteric Statistics*, 8. 127-147.

[2]. Baek, J. (1998), "A Local Linear Kernel Estimator for Sparse Multinomial Data," *Journal of the Korean Statistical Society*, 27, 515-529.

[3]. Fan, J., Heckman, N. E. and Wand, M.P.(1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141-150.

[4]. Hall, P, and Titterington, D. M. (1987), "On Smoothing Sparse Multinomial Data," *Australian Journal of Statistics*, 29, 19-37.

[5]. Kendall, M. and Stuart, A. (1979), *The Advanced Theory of Statistics*, Vol. 2, 4th ed. Charles Griffin & Company, London.

[6]. McCullagh , P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.