Journal of Statistical Planning and Inference 142 (2012) 1619–1622

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Short communication

A note on consistent estimation of Kullback–Leibler discrepancy in Poisson regression

Zbigniew Szkutnik*, Łukasz Grzyb

AGH University of Science and Technology, Faculty of Applied Mathematics, Al. Mickiewicza 30, 30-059 Kraków, Poland

ARTICLE INFO

Article history:
Received 18 March 2011
Received in revised form
2 August 2011
Accepted 21 December 2011
Available online 3 January 2012

Keywords:
Kullback–Leibler discrepancy
Poisson regression
Kernel estimator
Bandwidth selection

ABSTRACT

A recent theorem by Hannig and Lee on consistency of their estimator of Kullback–Leibler discrepancy is re-proved under assumptions suitably modified to correct a fault in the original proof.

1. Introduction

In a recent paper Hannig and Lee (2006) have studied direct kernel estimation of a Poisson regression function $f:[0,1)\to\mathbb{R}$, observed on a set of grid points $x_j=j/n$, $j=0,\dots,n-1$. The data consist of independent Poisson counts $y_j\sim\mathcal{P}(f_j)$, with $f_j=f(x_j)$. Sometimes the object of interest is another function, say g, related to f via an operator equation $f=\mathcal{A}g$ and the problem turns into a Poisson inverse problem. In a number of important applications this sort of data may result from the discretization of observations of an inhomogeneous Poisson process with intensity function nf. For more background references on direct and indirect Poisson regression models see, e.g., Reynaud-Bouret (2003), Antoniadis and Bigot (2006) and Szkutnik (2005).

For simplicity reasons, Hannig and Lee (2006) extend periodically both the function f and the data or, equivalently, identify the points 0 and 1 to transform the interval [0,1) to a circle. The kernel estimator in the direct problem takes the form

$$\hat{f}(x)=\frac{\sum_{j=0}^{n-1}K_h(x-x_j)y_j}{\sum_{j=0}^{n-1}K_h(x-x_j)}, \tag{1}$$

with a symmetric kernel K and with $K_h(\cdot)=h^{-1}K(\cdot/h)$. As usual, the crucial point for the quality of estimation is the choice of the bandwidth h. The method proposed and discussed by Hannig and Lee (2006) is to minimize w.r.t. h a suitably defined, nearly unbiased approximation $\hat{\Delta}_{KL}^{k}(h)$ of the Kullback–Leibler (KL) discrepancy

$$\Delta_{KL}(\hat{f},f)=\frac{1}{n}\sum_{j=0}^{n-1}[f_j-\hat{f}_j+\hat{f}_j(\log\hat{f}_j-\log f_j)]\approx\int_0^1[f(x)-\hat{f}(x)+\hat{f}(x)(\log\hat{f}(x)-\log f(x))]\,dx,$$

* Corresponding author.
E-mail address: szkutnik@agh.edu.pl (Z. Szkutnik).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.12.022

where $\hat{f}_j = \hat{f}(x_j)$. The approximation is defined as follows. First, in order to handle the problem of 'low count data', select an integer lumping parameter k and define

$$y_j^k = \sum_{|m| \le k} y_{j+m}, \quad f_j^k = \sum_{|m| \le k} f_{j+m}.$$

Then, with $w_m = K_h(m/n)/\sum_\ell K_h(\ell/n)$,

$$\alpha_j^k = \left\{ \log \frac{y_j^k}{2k+1} + \frac{1}{2y_j^k} - \frac{1.36177}{(y_j^k)^2} + \frac{2.15204}{(y_j^k)^3} \right\} I_{\{y_j^k > 0\}} - \{\log(2k+1) + 2.10898\} I_{\{y_j^k = 0\}}$$

and

$$\beta_j^k = \left[\frac{y_j^k}{2k+1} \log \frac{y_j^k}{2k+1} - \frac{1}{2(2k+1)} \right] I_{\{y_j^k > 0\}},$$

where I_A stands for the indicator function of the event A, the approximate KL discrepancy is defined as

$$\hat{\Delta}_{KL}^k(h) = \frac{1}{n} \sum_{j=0}^{n-1} \left\{ y_j - \hat{f}_j + \hat{f}_j \log \hat{f}_j - \beta_j^k \sum_{m=-k}^{k} w_m - \alpha_j^k \left(\hat{f}_j - \sum_{m=-k}^{k} w_m y_{j+m} \right) \right\}.$$

For this approximation, Hannig and Lee (2006) claim in their Theorem 1 that if, among other conditions, $nh \to \infty$ but $nh = o(\min\{n^{1/3}, k^2\})$, then

$$\frac{\hat{\Delta}_{KL}^k(h) - \Delta_{KL}(\hat{f}, f)}{\Delta_{KL}(\hat{f}, f)} \xrightarrow{P} 0,$$

which has an important consequence of the KL loss of the estimator (1), with h selected through the minimization of $\hat{\Delta}_{KL}^k(\cdot)$, converging to zero at the same speed as if h were selected through the minimization of $\Delta_{KL}(\hat{f}, f)$.

However, a crucial inequality (39) in the original proof does not seem to be correct. With $l(y) = 1 - y + y \log y$ (and not with $l(y) = y - 1 - \log y$, as in Hannig and Lee, 2006), one has

$$\Delta_{KL}(\hat{f}, f) = n^{-1} \sum_{j=0}^{n-1} f_j \, l(\hat{f}_j / f_j), \tag{2}$$

and it is crucial to lower bound $f_j l(\hat{f}_j / f_j)$. Taylor's formula gives $l(y) = (y-1)^2/(2\xi)$, with some ξ between 1 and y, and, consequently, $f_j l(\hat{f}_j / f_j) = (\hat{f}_j - f_j)^2/(2\xi f_j)$, with some ξ between 1 and $\hat{f}_j / f_j$, which gives

$$f_j l(\hat{f}_j / f_j) \ge C(\hat{f}_j - f_j)^2 / f_j, \tag{3}$$

with $C = 0.5 \cdot \min(1, \min_j f_j / \max_j \hat{f}_j)$, and not with $C = \min_j f_j / (2 \max_j f_j)$, as in Hannig and Lee (2006). Consequently, there is no non-zero lower bound for C, because $\hat{f}_j$ can be arbitrarily large. In effect, the inequality (38) in Hannig and Lee (2006) does not have to be true, which breaks down the proof on p. 904.

In the next section we show that this can be cured at the price of slowing down the rate of increase of nh to $(n \log n)h = o(\min\{(n \log n)^{1/3}, k^2\})$.

2. Results

Although y_j's and $\hat{f}_j$'s may *theoretically* be arbitrarily large, they are *practically* bounded in the sense that extremely large values are extremely unlikely. Our proof will quantify this idea by means of the following lemma.

Lemma 1. *Let $y_j \sim \mathcal{P}(\lambda_j)$ be Poisson random variables with $\lambda_j \le M$ for a constant M and for $j = 1, 2, \ldots$. Then, for any $s > 0$,*

$$P\left(\max_{1 \le j \le n} y_j > M \log n \right) = o(n^{-s}),$$

when $n \to \infty$.

Proof. It follows from the Markov inequality that for any non-negative random variable X and any $t > 0$

$$P(X \ge t) \le \inf_{u > 0} e^{-tu} \psi(u),$$

where $\psi(\cdot)$ is the moment generating function of X (cf., e.g., Billingsley, 1979). For $X \sim \mathcal{P}(\lambda)$, this gives for $t > \lambda$

$$P(X \ge t) \le \exp[t - \lambda - t \log(t/\lambda)].$$

Set $t = c\lambda$. Then, for $c > e^2$,

$$P(X \ge c\lambda) \le \exp[c\lambda(1 - \log c)] \le \exp[-(\lambda/2) c \log c] = 1/c^{c\lambda/2}.$$

Hence, with $X \sim \mathcal{P}(M)$,

$$P\left(\max_{1 \le j \le n} y_j > M \log n\right) \le nP(X > M \log n) \le \frac{n}{(\log n)^{(M \log n)/2}} = \mathrm{O}(n^{-(M \log \log n)/3}) = \mathrm{o}(n^{-s})$$

for any $s > 0$, because

$$\log \frac{n}{(\log n)^{(M \log n)/2}} = \log n - \frac{M}{2} \log n (\log \log n) = \log n \left[1 - \frac{M}{2} \log \log n\right]. \qquad \square$$

We are now ready to prove the corrected version of (the last part of) Theorem 1 of Hannig and Lee (2006).

Theorem 1. *Suppose that f is Lipschitz and bounded away from zero and infinity and that the kernel K is compactly supported, symmetrical, unimodal and square-integrable. Let b be the number of grid points in the support of K_h. If $k < b < n$ simultaneously approach infinity and*

$$b \log n = \mathrm{o}(\min\{(n \log n)^{1/3}, k^2\}), \tag{4}$$

then

$$\frac{\hat{\Delta}_{KL}^{k}(h) - \Delta_{KL}(\hat{f}, f)}{\Delta_{KL}(\hat{f}, f)} \xrightarrow{P} 0.$$

Proof. As in Hannig and Lee (2006), one has

$$E(\hat{f}_j^{\,2}) = \{E(\hat{f}_j)\}^2 + \sum_m w_m^2 f_{j+m}, \quad |E(\hat{f}_j) - f_j| = \mathrm{O}\left(\frac{b}{n}\right)$$

and

$$\sum_m w_m^2 \asymp \frac{L \int K^2(\omega)\, d\omega}{b}.$$

With $M_1 = \max f(x)$ and $M_2 = \min f(x)$, define the event

$$S_n = \left\{\max_{1 \le j \le n} y_j \le M_1 \log n\right\},$$

and denote with I_{S_n} and $I_{S_n'}$ the indicator functions of S_n and of its complement S_n'. On S_n, one has $\max_{1 \le j \le n} \hat{f}_j \le M_1 \log n$ and inequality (3) implies

$$f_j I(\hat{f}_j / f_j) \ge \frac{M_2}{2M_1 \log n} \frac{(\hat{f}_j - f_j)^2}{f_j}. \tag{5}$$

Using (2), (5) and the representation

$$\frac{E(\hat{f}_j - f_j)^2}{f_j} = \frac{\{E(\hat{f}_j) - f_j\}^2}{f_j} + \sum_m w_m^2 \frac{f_{j+m}}{f_j},$$

one then obtains, with a constant D depending on the kernel K,

$$E\Delta_{KL}(\hat{f}, f) \ge E\{\Delta_{KL}(\hat{f}, f) I_{S_n}\} \ge \frac{M_2}{2M_1 n \log n} \sum_j \left[\frac{E(\hat{f}_j - f_j)^2}{f_j} - \frac{E\{(\hat{f}_j - f_j)^2 I_{S_n'}\}}{f_j}\right] \ge \frac{D}{b \log n} + \mathrm{o}\left(\frac{1}{b \log n}\right), \tag{6}$$

because of (4) and because $E\{(\hat{f}_j - f_j)^2 I_{S_n'}\} = \mathrm{o}(1/b)$ after an application of the Cauchy–Schwarz inequality and of Lemma 1. Inequality (6) differs from inequality (38) in Hannig and Lee (2006) by the logarithmic term. The assumption (4) implies that $b/n = \mathrm{o}(b^{-2} \log^{-2} n)$ and $1/k^2 = \mathrm{o}(b^{-1} \log^{-1} n)$, so that $b/n + 1/k^4 = \mathrm{o}(b^{-2} \log^{-2} n)$. Hence, there exist r_n such that $b/n + 1/k^4 = \mathrm{o}(r_n^2)$ and $r_n = \mathrm{o}(b^{-1} \log^{-1} n)$ and one can follow the lines of the original proof on p. 904 in Hannig and Lee (2006). It suffices to show that $P(\{\Delta_{KL}(\hat{f}, f) < r_n\}) \to 0$. As in Hannig and Lee (2006), one obtains

$$P(\{\Delta_{KL}(\hat{f}, f) < r_n\}) \le \frac{Var \Delta_{KL}(\hat{f}, f)}{[E\Delta_{KL}(\hat{f}, f) - r_n]^2}, \tag{7}$$

and the nominator is $\mathrm{O}(b/n)$. Because of (6) and the choice of r_n, the denominator is of the order of $b^{-2} \log^{-2} n$, so that the right hand side of (7) is of the order of $(b^3 \log^2 n)/n$, which tends to zero, because of (4). This completes the proof. $\square$

Note that $b = \lfloor Lnh \rfloor$, if $\lfloor Lnh \rfloor$ is odd and $b = \lfloor Lnh \rfloor + 1$, if $\lfloor Lnh \rfloor$ is even, where L denotes the length of the support of K. Consequently, (4) can be written equivalently as $(n \log n)h = \mathrm{o}(\min\{(n \log n)^{1/3}, k^2\})$.

Acknowledgments

This work was partially supported by the Polish Ministry of Science and Higher Education under AGH local grant. The authors kindly acknowledge important and helpful comments of anonymous reviewers.

References

Antoniadis, A., Bigot, J., 2006. Poisson inverse problems. Annals of Statistics 34, 2132–2158.
Billingsley, P., 1979. Probability and Measure. Wiley, New York.
Hannig, J., Lee, T.C.M., 2006. On Poisson signal estimation under Kullback–Leibler discrepancy and squared risk. Journal of Statistical Planning and Inference 136, 882–908.
Reynaud-Bouret, P., 2003. Adaptive estimation of inhomogeneous Poisson processes via concentration inequalities. Probability Theory and Related Fields 126, 103–153.
Szkutnik, Z., 2005. B-splines and discretization in an inverse problem for Poisson processes. Journal of Multivariate Analysis 93, 198–221.