On Poisson Signal Estimation under Kullback–Leibler Discrepancy and Squared Risk

Jan Hannig* and Thomas C. M. Lee

Colorado State University

July 31, 2003; revised: April 3, 2004, August 19, 2004

Abstract

Regression problems under Poisson variability arise in many different scientific areas such as, for examples, astrophysics and medical imaging. This article considers the problem of bandwidth selection for kernel smoothing of Poisson data. Its first contribution is the proposal of a new bandwidth selection method that aims to choose the bandwidth that minimizes the Kullback–Leibler (KL) distance between the estimated and the unknown true regression functions. The idea behind is to first construct an estimator of the KL distance and then chooses the minimizer of this distance estimator as the bandwidth. The consistency of this distance estimator is established. As a second contribution, this article establishes the consistency of an existing estimator that targets the L_2 risk between the true and the estimated regression functions. In a simulation study, when the targeting distance measure is the KL discrepancy, the proposed KL–based bandwidth selector outperforms a bandwidth selector that uses deviance cross–validation.

Key words and phrases: bandwidth selection; kernel smoothing; Kullback–Leibler discrepancy; L_2 risk; Poisson counts.

AMS 2000 classifiaction: Primary 62G08, secondary 62G20.

Running title: Nonparametric Poisson Regression

*Corresponding Author. Postal: Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA. Email: hannig@stat.colostate.edu; Phone: (970) 491 7460; Fax: (970) 491 7895.

1 Introduction

This article is concerned with regression function estimation under the following Poisson setting. Suppose n independent Poisson counts y_j are observed at a set of grid points x_j:

$$y_j \sim P(f_j), \quad f_j = f(x_j), \quad x_j = \frac{j}{n}, \quad j = 0, \ldots, n-1, \tag{1}$$

where $P(f_j)$ denotes a Poisson distribution with mean f_j. The goal is to estimate the unknown regression function f, which is assumed to be "smooth". Practical applications covered by this setting include the estimation of X–ray or γ–ray burst intensity maps in astrophysics (e.g., Kolaczyk 1997 and van Dyk, Connors, Kashyap & Siemiginowska 2001) and the smoothing of Poisson count data in medical imaging (e.g., Hudson & Lee 1998 and La Riviere & Pan 2000). Possible generalizations to this setting, including non–equally spaced designs, will be discussed in Section 4.

For simplicity we shall primarily focus on the following kernel–smoothed estimator for f. Let K be a kernel function. Let also h be a non–negative smoothing parameter, also known as the bandwidth, that controls the amount of smoothing. Write $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$. The kernel estimator $\hat{f}_j$ for f_j is defined as

$$\hat{f}_j = \sum_{m=0}^{n-1} K_h(x_m - x_j) y_m \Bigg/ \sum_{l=0}^{n-1} K_h(x_l - x_j), \qquad j = 0, \ldots, n-1. \tag{2}$$

Note that $\hat{f}_j$ is a function of h, but, for brevity, this dependence is suppressed from its notation. It is well known that the choice of h is much more crucial than the choice of K (e.g., see Wand & Jones 1995).

The purpose of this article is to study, both theoretically and empirically, the properties of two data–dependent methods for choosing h. The first method aims to choose the h that minimizes the following Kullback–Leibler discrepancy between $\hat{f}$ and f

$$\Delta_{\mathrm{KL}}(\hat{f}, f) = \frac{1}{n} \sum_{j=0}^{n-1} \left\{ f_j - \hat{f}_j + \hat{f}_j (\log \hat{f}_j - \log f_j) \right\}. \tag{3}$$

Derivation for $\Delta_{\mathrm{KL}}(\hat{f}, f)$ is given in Appendix A. The second method aims for minimizing the L_2 risk between $\hat{f}$ and f

$$\Delta_{\mathrm{R}}(\hat{f}, f) = \frac{1}{n} \sum_{j=0}^{n-1} (f_j - \hat{f}_j)^2.$$

Notice that both $\Delta_{\mathrm{KL}}(\hat{f}, f)$ and $\Delta_{\mathrm{R}}(\hat{f}, f)$ are unknown, therefore direct minimization of these two discrepancy measures is not possible. A common approach to overcoming this problem is first to construct an estimator for the discrepancy measure of interest, and then choose the bandwidth that minimizes such a discrepancy estimator. As mentioned in Linhart & Zucchini (1986), the rationale is that the bandwidth that minimizes the discrepancy estimator should also approximately minimize the unknown discrepancy. Other classical statistical model selection criteria that follow this rationale include Mallows' C_p and Akaike Information Criterion. This article proposes a consistent estimator for $\Delta_{\mathrm{KL}}(\hat{f}, f)$, as well as establishes the consistency of an existing estimator for $\Delta_{\mathrm{R}}(\hat{f}, f)$. It is worth mentioning that a technical challenge for estimating $\Delta_{\mathrm{KL}}(\hat{f}, f)$ occurs when f_j is close to zero; i.e., when $\log f_j$ approaches $-\infty$.

The problem of function estimation under Poisson noise has of course been studied by various authors. Earlier references include Hudson (1978) and Hudson (1985), who studies the problem from a L_2 perspective. Pawitan & O'Sullivan (1993) develop an L_2 risk based method for choosing the amount of smoothing in medical image reconstruction. In the context of generalized linear models, a computational procedure, based on cross–validating the deviance, is described in Hastie & Tibshirani (1990, Ch. 6). Xiang & Wahba (1996) propose a generalized approximate cross–validation (GACV) procedure for choosing the smoothing parameter for smoothing splines with non–Gaussian data (see also Gu & Xiang 2001). Their numerical results suggest that, in the Bernoulli noise case, GACV can be used to estimate the Kullback–Leibler discrepancy. However, no proof has been provided for supporting this observation. Further results concerning the use of smoothing splines for non–Gaussian data can be found in Gu (2002, Ch. 5). More recently, a wavelet thresholding method tailored for Poisson noise is proposed by Kolaczyk (1999*b*). Also, Kolaczyk (1999*a*) and Nowak & Kolaczyk (2000) provide Bayesian multi–scale methods for handling Poisson inverse problems.

The rest of this article is organized as follows. The main theoretical contributions of this article are presented in Section 2. In Section 3 results from numerical experiments are reported for evaluating the two bandwidth selection methods mentioned above. Generalizations and conclusions are offered in Section 4. Technical details are deferred to the appendices.

2 Theoretical Results

This section presents the main contributions of this article, namely, the proposal of a new consistent estimator for $\Delta_{\mathrm{KL}}(\hat{f}, f)$, and a theoretical study of an earlier estimator for $\Delta_{\mathrm{R}}(\hat{f}, f)$. We remark that the kernel estimator $\hat{f}_j$ can also be interpreted as a weighted average of the y_j's. It is because one could write

$$\hat{f}_j = \sum_m w_{m-j} y_m \quad \text{with} \quad w_{m-j} = \frac{K_h(x_m - x_j)}{\sum_l K_h(x_l - x_j)}. \tag{4}$$

Notice that the weights w_m's sum to unity. In what follows we will assume that f satisfies the periodic boundary condition; i.e., $f_j = f_{j+n} = f_{j-n}$ for $j = 0, \ldots, n-1$. This will allow us to have the weights w_m independent of location.

2.1 Estimating the Kullback–Leibler Discrepancy

One major difficulty behind the construction of an estimator for $\Delta_{\text{KL}}(\hat{f}, f)$ is the need for estimating $\log f_j$ when f_j is close to zero. It is because under this situation y_j will take value 0 with probability close to $1 - f_j \approx 1$ and 1 with probability close to $f_j \approx 0$. This will in turn give rise to "low count" data. The way that we handle this "low count" situation is to lump neighboring observations of y_j (i.e., $y_{j \pm k}$ for small k) together so that the sum of these y_j's is large enough to be worked with. Thus in our estimator, denoted as $\hat{\Delta}^k_{\text{KL}}(h)$, there is one integer parameter k that needs to be pre–specified. This parameter k is used to control the amount of lumping. At the end of this subsection we will discuss the issue of how to pre–specify k. The details of the construction of our estimator $\hat{\Delta}^k_{\text{KL}}(h)$, together with additional comments on k, are given in Appendix B. Here we only describe the main idea behind this construction.

When estimating $\Delta_{\text{KL}}(\hat{f}, f)$ we need to be able to estimate $\log f_j$ and $f_j \log f_j$. If Y has Poisson(λ) distribution the arguments in Appendix B show that

$$E\left\{\left(\log Y - \frac{1}{2Y}\right) I_{\{Y>0\}}\right\} \approx \log \lambda, \tag{5}$$

$$E(Y \log Y) - \frac{1}{2} \approx \lambda \log \lambda, \tag{6}$$

where I_E is the indicator function for event E. The approximation in (6) is uniformly good for all λ, which suggests estimating $\lambda \log \lambda$ with $Y \log Y - \frac{1}{2} I_{\{Y>0\}}$. This and the lumping idea described above lead directly to the definition of β^k_j below. The approximation in (5) needs bias correction for small λ. The bias corrected version of the estimator of $\log \lambda$ is then used below for the definition of α^k_j. The estimator $\hat{\Delta}^k_{\text{KL}}(h)$ is then derived from (3) by replacing of $\log f_j$ by its estimator α^k_j and $f_j \log f_j$ by β^k_j.

We now can state the exact form of our estimator $\hat{\Delta}^k_{\text{KL}}(h)$. Define

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

$$\alpha_j^k = \left\{\log\frac{y_j^k}{2k+1} + \frac{0.5}{y_j^k} - \frac{1.36177}{(y_j^k)^2} + \frac{2.15204}{(y_j^k)^3}\right\} I_{\{y_j^k>0\}} - \{\log(2k+1) + 2.10898\} I_{\{y_j^k=0\}},$$

and

$$\beta_j^k = \frac{y_j^k}{2k+1}\log\frac{y_j^k}{2k+1} - \frac{1}{2(2k+1)} I_{\{y_j^k>0\}}.$$

Our estimator admits the following expression:

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

If the target discrepancy measure is $\Delta_{\text{KL}}(\hat{f}, f)$, we propose to choose the bandwidth h as the minimizer of $\hat{\Delta}^k_{\text{KL}}(h)$.

We have established the consistency of our estimator. The results are summerized in the following theorem. The proof is given in Appendix C.

Theorem 1. *Suppose that f is Lipschitz with constant D and bounded away from 0 and ∞, and that the kernel K is compact, symmetrical, unimodal and square–integrable. Then*

$$\left|E\{\hat{\Delta}^k_{\text{KL}}(h) - \Delta_{\text{KL}}(\hat{f}, f)\}\right| \le \frac{C_1 M_1}{M_2(2k+1)^2} + \frac{C_2}{M_2 b(2k+1)} + \frac{C_3 M_1 D k(k+1)}{M_2(2k+1)n} + \frac{C_4 D k(k+1)}{nb}\{1 + 2\max(-\log M_2, \log M_1)\} + \frac{C_5 D^2 k^2 (1+k)^2}{M_1(2k+1)n^2 b}, \quad (7)$$

where $M_1 = \max f(x)$, $M_2 = \min f(x)$, b is the number of y_j's in the support of K_h, and C_1, C_2, C_3, C_4, C_5 are constants depending only on K. Furthermore

$$\text{var}\{\hat{\Delta}^k_{\text{KL}}(h) - \Delta_{\text{KL}}(\hat{f}, f)\} \le C\frac{b}{n}, \quad (8)$$

where C is a constant depending only on f.

In addition, if $k < b < n$ are simultaneously approaching infinity, $b = o\{\min(n^{1/3}, k^2)\}$ and b grows at least polynomially fast, then

$$\frac{\hat{\Delta}^k_{\text{KL}}(h) - \Delta_{\text{KL}}(\hat{f}, f)}{\Delta_{\text{KL}}(\hat{f}, f)} \to 0 \quad \textit{in probability.} \quad (9)$$

We remark that the quantity b plays a dual role to the bandwidth h. It is because $b = \lfloor Lnh \rfloor$ if L is the length of the support of K.

Now we consider the choice of k. Of course its optimal value would depend on different unknown quantities such as various properties of f. In practice these quantities may not be available, which makes pre–specifying the optimal value of k difficult. However, from our numerical experience, setting $k = 1$ is often a good and conservative choice. We have used $k = 1$ through out all our numerical experiments described in Section 3 below.

Remark 1. The established consistency of the estimator $\hat{\Delta}^k_{\mathrm{KL}}(h)$ suggests an important implication about the asymptotic behavior of our estimator. Denote $\hat{f}_{h_{\mathrm{KL},0}}$ the estimator of f calculated using the optimal bandwidth $h_{\mathrm{KL},0}$ minimizing $\Delta_{\mathrm{KL}}(\hat{f}, f)$ (not obtainable in practice) and $\hat{f}_{\hat{h}_{\mathrm{KL}}}$ the estimator of f calculated using the bandwidth $\hat{h}_{\mathrm{KL}}$ minimizing $\hat{\Delta}^k_{\mathrm{KL}}(h)$ (our estimator). Assume that

$$\frac{\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f)}{\Delta_{\mathrm{KL}}(\hat{f}, f)} \to 0$$

for both $h = \hat{h}_{\mathrm{KL}}$ and $h = h_{\mathrm{KL},0}$. This could be achieved for example by strengthening equation (9) of Theorem 1 to hold uniformly for all h. Then

$$\frac{\hat{\Delta}^k_{\mathrm{KL}}(\hat{h}_{\mathrm{KL}})}{\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)} \to 1 \quad \text{and} \quad \frac{\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f)}{\hat{\Delta}^k_{\mathrm{KL}}(h_{\mathrm{KL},0})} \to 1. \tag{10}$$

Since $\hat{f}_{h_{\mathrm{KL},0}}$ minimizes $\Delta_{\mathrm{KL}}(\hat{f}, f)$ and $\hat{f}_{\hat{h}_{\mathrm{KL}}}$ minimizes $\hat{\Delta}^k_{\mathrm{KL}}(h)$, we have

$$\frac{\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)}{\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f)} \geq 1 \quad \text{and} \quad \frac{\hat{\Delta}^k_{\mathrm{KL}}(h_{\mathrm{KL},0})}{\hat{\Delta}^k_{\mathrm{KL}}(\hat{h}_{\mathrm{KL}})} \geq 1. \tag{11}$$

From here and (10) calculate

$$\limsup_{n\to\infty} \frac{\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)}{\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f)} \leq \limsup_{n\to\infty} \frac{\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)}{\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f)} \cdot \frac{\hat{\Delta}^k_{\mathrm{KL}}(h_{\mathrm{KL},0})}{\hat{\Delta}^k_{\mathrm{KL}}(\hat{h}_{\mathrm{KL}})} = 1. \tag{12}$$

Combining (11) and (12) we have $\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)/\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f) \to 1$ concluding that $\Delta_{\mathrm{KL}}(\hat{f}_{\hat{h}_{\mathrm{KL}}}, f)$ converges to 0 at the same speed as $\Delta_{\mathrm{KL}}(\hat{f}_{h_{\mathrm{KL},0}}, f)$.

2.2 Estimating the L_2 Risk

Unbiased estimation of the L_2 risk under Poisson variability has been studied by previous authors; e.g., see Hudson (1978) and Pawitan & O'Sullivan (1993). For the current setting, the following estimator $\hat{\Delta}_{\mathrm{R}}(h)$ for $\Delta_{\mathrm{R}}(\hat{f}, f)$ can be obtained from results in Pawitan & O'Sullivan (1993):

$$\hat{\Delta}_{\mathrm{R}}(h) = \frac{1}{n} \sum_j \left\{ (y_j - \hat{f}_j)^2 + (2w_0 - 1) y_j \right\}.$$

One could choose h as the minimizer of $\hat{\Delta}_{\mathrm{R}}(h)$ if $\Delta_{\mathrm{R}}(\hat{f}, f)$ is the target discrepancy measure.

We have also studied the theoretical properties of $\hat{\Delta}_{\mathrm{R}}(h)$ in a similar fashion as for $\hat{\Delta}^k_{\mathrm{KL}}(h)$. Our results are summarized in the theorem below. In short, our contribution in this subsection is that we have established the consistency of $\hat{\Delta}_{\mathrm{R}}(h)$. Proof of the theorem is delayed to Appendix D.

Theorem 2. *Suppose that f is Lipschitz and bounded, and that K is compact, symmetrical, unimodal and square–integrable. Then*

$$E\{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)\} = 0 \tag{13}$$

and

$$\mathrm{var}\{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)\} \leq C\frac{b}{n}, \tag{14}$$

where C is a constant depending only on f.

In addition, if $b = o(n^{1/3})$

$$\frac{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)}{\Delta_{\mathrm{R}}(\hat{f}, f)} \to 0 \quad \textit{in probability.} \tag{15}$$

2.3 Computational Issues

Both of the above two bandwidth selection procedures are computationally inexpensive and straightforward to implement. It is because both $\hat{\Delta}^k_{\mathrm{KL}}(h)$ and $\hat{\Delta}_{\mathrm{R}}(h)$ can be directly computed without using any Monte Carlo type approximations. Also, since the data are assumed to be regularly spaced, fast computation of $\hat{f}_j$ can be achieved by using Fourier techniques.

3 Numerical Results

A small scale simulation study was conducted to evaluate the empirical properties of the two bandwidth selection methods discussed above. For comparative purposes, the cross–validating (CV) deviance procedure described in Hastie & Tibshirani (1990, Ch. 6) was also studied. This procedure chooses the bandwidth h that minimizes the following leave–one–out CV deviance function

$$\mathrm{CVDev}(h) = \frac{1}{n}\sum_{j=0}^{n-1}\left\{\hat{f}_{-j} - y_j + y_j(\log y_j - \log \hat{f}_{-j})\right\},$$

where $\hat{f}_{-j}$ is the estimate of f_j obtained from using all but the ith observation y_i. Notice that $\mathrm{CVDev}(h)$ is targeting the KL discrepancy.

3.1 Setup

In this study three test functions, three signal–to–noise ratios (snrs) and four sample sizes were used. The three test functions were

$$\begin{aligned}
&\text{Test Function 1:} \quad f(x) = \max\{\sin(4\pi x), \epsilon\}, \quad \epsilon = 0.000005,\\
&\text{Test Function 2:} \quad f(x) = \max\{\sin(4\pi x) + 1, \epsilon\},\\
&\text{Test Function 3:} \quad f(x) = 2\sin(4\pi x) + 3.
\end{aligned}$$

These three test functions are derived from a standard sine wave and present three different levels of difficulties. For Test Function 1 half of its domain "touches zero" (i.e., has "y–value" that are virtually zero), for Test Function 2 the valleys of the sine wave "touch zero", while for Test Function 3 the whole sine wave is shifted up so that it is sufficiently far away from zero. As indicated above a major difficulty for estimating $\Delta_{\text{KL}}(\hat{f}, f)$ is the estimation of $\log f(x)$ when $f(x) \approx 0$. Thus one may treat that Test Function 1 is a hard example, Test Function 2 is a medium example while Test Function 3 is an easy example. Plots of the test functions can be found in Figures 1 to 6.

We define signal to noise ratio (further just snr) as $\|f\|/\sqrt{\text{var}(f)} = \sqrt{\sum f_j^2 / \sum f_j}$, where $\text{var}(f)$ can be interpreted as the variance of the noise. To change the snr of a test function f, a constant c is multiplied to it so that $\sqrt{\sum (cf_j)^2 / \sum cf_j}$ reaches the pre–specified value. The three snrs used were 2, 4, and 6. The four sample sizes were $n = 200$, 400, 800 and 1600. The kernel function used was $K(x) = \frac{3}{4}(1 - x^2), x \in [0, 1]$. It is the optimal kernel of order (0, 2) derived in Gasser, Műller & Mammitzsch (1985). Throughout the whole study we set $k = 1$.

For each of the above 36 experimental settings, 250 independent data sets were simulated. For each of these simulated data sets, the bandwidths $\hat{h}_{\text{KL}}$, $\hat{h}_{\text{R}}$ and $\hat{h}_{\text{DEV}}$ that minimize respectively $\hat{\Delta}^k_{\text{KL}}(h)|_{k=1}$, $\hat{\Delta}_{\text{R}}(h)$ and $\text{CVDev}(h)$ were computed. In addition, two practically unobtainable optimal bandwidths were also computed. They were $h_{\text{KL},0}$, the bandwidth that minimizes $\Delta_{\text{KL}}(\hat{f}, f)$, and $h_{\text{R},0}$, the bandwidth that minimizes $\Delta_{\text{R}}(\hat{f}, f)$.

3.2 Results

Four numerical measures were adopted to evaluate the quality of $\hat{h}_{\text{KL}}$. Let $\hat{f}_{[h]}$ be the estimate of f computed using the bandwidth h. The four numerical measures were

$$\Delta_{\text{KL}}(\hat{f}_{[\hat{h}_{\text{KL}}]}, f), \quad \Delta_{\text{R}}(\hat{f}_{[\hat{h}_{\text{KL}}]}, f), \quad \frac{\Delta_{\text{KL}}(\hat{f}_{[\hat{h}_{\text{KL}}]}, f)}{\Delta_{\text{KL}}(\hat{f}_{[h_{\text{KL},0}]}, f)}, \quad \text{and} \quad \frac{\Delta_{\text{R}}(\hat{f}_{[\hat{h}_{\text{KL}}]}, f)}{\Delta_{\text{R}}(\hat{f}_{[h_{\text{R},0}]}, f)}.$$

The first and the third measures were used to assess the performance of $\hat{h}_{\text{KL}}$ when $\Delta_{\text{KL}}(\hat{f}, f)$ is of interest: the first assesses the quality in an absolute sense while the third assesses the quality

Test Function	Bandwidth Selection	Sample Size $n = 200$	$n = 400$	$n = 800$	$n = 1600$
1	$\hat{h}_{\text{KL}}$	0.284 (0.005)	0.150 (0.002)	0.088 (0.001)	0.051 (0.001)
	$\hat{h}_{\text{R}}$	0.927 (0.021)	0.598 (0.014)	0.404 (0.010)	0.277 (0.006)
	$\hat{h}_{\text{DEV}}$	0.333 (0.008)	0.181 (0.004)	0.104 (0.002)	0.060 (0.001)
2	$\hat{h}_{\text{KL}}$	0.081 (0.001)	0.043 (0.001)	0.025 (0.001)	0.015 (0.001)
	$\hat{h}_{\text{R}}$	0.153 (0.004)	0.090 (0.002)	0.056 (0.001)	0.034 (0.001)
	$\hat{h}_{\text{DEV}}$	0.087 (0.002)	0.047 (0.001)	0.026 (0.001)	0.015 (0.001)
3	$\hat{h}_{\text{KL}}$	0.037 (0.001)	0.020 (0.001)	0.012 (0.001)	0.007 (0.001)
	$\hat{h}_{\text{R}}$	0.037 (0.001)	0.020 (0.001)	0.012 (0.001)	0.007 (0.001)
	$\hat{h}_{\text{DEV}}$	0.037 (0.001)	0.020 (0.001)	0.012 (0.001)	0.006 (0.001)

Table 1: Averages and standard deviations (in parentheses) of $\Delta_{\text{KL}}(\hat{f}_{[h]}, f)$ for $h = \hat{h}_{\text{KL}}$ (minimizer of $\hat{\Delta}^k_{\text{KL}}(h)$), $h = \hat{h}_{\text{R}}$ (minimizer of $\hat{\Delta}_{\text{R}}(h)$) and $h = \hat{h}_{\text{DEV}}$ (minimizer of CVDev(h)).

relative to the best possible bandwidth $h_{\text{KL},0}$ that one could get only if f is known. Although $\hat{h}_{\text{KL}}$ is not targeting the L_2 risk $\hat{\Delta}_{\text{R}}(h)$, it would still be interesting and worthwhile to include the second and the fourth measures. Averages and standard deviations for these four measures, computed from the 250 repetitions for each experiment setting, are given in Tables 1 to 4. Similar values for evaluating the quality of $\hat{h}_{\text{R}}$ and $\hat{h}_{\text{DEV}}$ were also computed and are reported in the same tables.

The following empirical conclusions can be drawn from examining these tables. First, for all experimental settings, the values of $\Delta_{\text{KL}}(\hat{f}, f)$ and $\Delta_{\text{R}}(\hat{f}, f)$ decrease as n increases (see Tables 1 and 2). Secondly, for the "easy" Test Function 3, all three bandwidth selectors $\hat{h}_{\text{KL}}$, $\hat{h}_{\text{R}}$ and $\hat{h}_{\text{DEV}}$ gave very similar performances regardless of which numerical measure is being used. Thirdly, for Test Functions 1 and 2, $\hat{h}_{\text{KL}}$ seems to outperform $\hat{h}_{\text{DEV}}$ when the targeting distance measure is the KL discrepancy. Lastly, as most of the corresponding entries in Table 3 are close to 1, the proposed $\hat{h}_{\text{KL}}$ gave very good results when comparing to the best possible (but practically unobtainable) $h_{\text{KL},0}$.

To visually evaluate the quality of various estimated curves, the following was done. For Test Function 1 with snr = 4 and $n = 200$, the simulated data set that corresponds to the 125th sorted value of $\Delta_{\text{KL}}(\hat{f}_{[\hat{h}_{\text{KL}}]}, f)$ is plotted in Figures 1, together with the estimated curves computed using the corresponding $\hat{h}_{\text{KL}}$ $\hat{h}_{\text{R}}$ and $\hat{h}_{\text{DEV}}$. Similar plots were also produced for $n = 800$ and also for Test Functions 2 and 3; they are displayed in Figures 2 to 6.

Test Function	Bandwidth Selection	Sample Size			
		$n = 200$	$n = 400$	$n = 800$	$n = 1600$
1	$\hat{h}_{\text{KL}}$	1.175 (0.027)	0.754 (0.015)	0.486 (0.009)	0.321 (0.006)
	$\hat{h}_{\text{R}}$	0.814 (0.022)	0.465 (0.011)	0.270 (0.006)	0.165 (0.004)
	$\hat{h}_{\text{DEV}}$	1.164 (0.034)	0.708 (0.018)	0.444 (0.009)	0.292 (0.006)
2	$\hat{h}_{\text{KL}}$	1.066 (0.027)	0.588 (0.014)	0.345 (0.008)	0.204 (0.004)
	$\hat{h}_{\text{R}}$	0.876 (0.027)	0.428 (0.011)	0.253 (0.006)	0.147 (0.003)
	$\hat{h}_{\text{DEV}}$	0.991 (0.027)	0.524 (0.012)	0.317 (0.007)	0.194 (0.004)
3	$\hat{h}_{\text{KL}}$	0.892 (0.026)	0.499 (0.013)	0.297 (0.008)	0.164 (0.004)
	$\hat{h}_{\text{R}}$	0.889 (0.026)	0.495 (0.014)	0.284 (0.007)	0.158 (0.004)
	$\hat{h}_{\text{DEV}}$	0.905 (0.026)	0.490 (0.013)	0.290 (0.007)	0.16 (0.004)

Table 2: Similar to Table 1 but for $\Delta_{\text{R}}(f, \hat{f}_{[h]})$.

Test Function	Bandwidth Selection	Sample Size			
		$n = 200$	$n = 400$	$n = 800$	$n = 1600$
1	$\hat{h}_{\text{KL}}$	1.835 (0.031)	1.641 (0.023)	1.584 (0.021)	1.541 (0.023)
	$\hat{h}_{\text{R}}$	6.299 (0.196)	6.827 (0.198)	7.584 (0.217)	8.539 (0.223)
	$\hat{h}_{\text{DEV}}$	2.147 (0.051)	2.001 (0.044)	1.874 (0.034)	1.821 (0.036)
2	$\hat{h}_{\text{KL}}$	1.118 (0.009)	1.096 (0.008)	1.096 (0.008)	1.085 (0.008)
	$\hat{h}_{\text{R}}$	2.260 (0.067)	2.399 (0.063)	2.540 (0.066)	2.614 (0.071)
	$\hat{h}_{\text{DEV}}$	1.213 (0.017)	1.180 (0.013)	1.136 (0.010)	1.112 (0.008)
3	$\hat{h}_{\text{KL}}$	1.159 (0.018)	1.155 (0.017)	1.162 (0.017)	1.121 (0.013)
	$\hat{h}_{\text{R}}$	1.169 (0.018)	1.170 (0.019)	1.136 (0.014)	1.116 (0.009)
	$\hat{h}_{\text{DEV}}$	1.187 (0.022)	1.145 (0.017)	1.142 (0.015)	1.100 (0.010)

Table 3: Similar to Table 1 but for $\frac{\Delta_{\text{KL}}(\hat{f}_{[h]}, f)}{\Delta_{\text{KL}}(\hat{f}_{[h_{\text{KL},0}]}, f)}$.

Test Function	Bandwidth Selection	Sample Size $n = 200$	$n = 400$	$n = 800$	$n = 1600$
1	$\hat{h}_{\rm KL}$	1.759 (0.033)	1.985 (0.043)	2.149 (0.041)	2.269 (0.045)
	$\hat{h}_{\rm R}$	1.206 (0.031)	1.193 (0.029)	1.144 (0.015)	1.117 (0.012)
	$\hat{h}_{\rm DEV}$	1.746 (0.051)	1.863 (0.053)	1.954 (0.043)	2.053 (0.042)
2	$\hat{h}_{\rm KL}$	1.597 (0.041)	1.698 (0.037)	1.602 (0.032)	1.585 (0.028)
	$\hat{h}_{\rm R}$	1.249 (0.032)	1.182 (0.021)	1.137 (0.017)	1.113 (0.014)
	$\hat{h}_{\rm DEV}$	1.458 (0.036)	1.497 (0.030)	1.471 (0.027)	1.507 (0.025)
3	$\hat{h}_{\rm KL}$	1.198 (0.022)	1.197 (0.021)	1.204 (0.020)	1.157 (0.016)
	$\hat{h}_{\rm R}$	1.190 (0.021)	1.184 (0.021)	1.147 (0.017)	1.111 (0.012)
	$\hat{h}_{\rm DEV}$	1.232 (0.028)	1.170 (0.019)	1.174 (0.019)	1.120 (0.012)

Table 4: Similar to Table 1 but for $\frac{\Delta_{\rm R}(f,\hat{f}_{[h]})}{\Delta_{\rm R}(f,\hat{f}_{[h_{\rm R,0}]})}$.

4 Concluding Remarks

In this article the problem of bandwidth selection for kernel regression with Poisson data is considered. A new bandwidth selection procedure that targets the KL discrepancy is proposed and both analytically and empirically studied. In addition, an existing L_2 risk based bandwidth selection procedure is also studied. In a simulation study the proposed bandwidth selection procedure outperformed a deviance cross–validation based procedure if the KL discrepancy is the target distance measure.

Several important extensions of this work are worth considering. The first one comes when the design points are non–equally spaced. One can construct an estimator for the KL discrepancy as before, but use nearest neighbors when calculating y_j^k. This approach works well if the design points x are dense enough. For example if $\max_j(x_j - x_{j-1}) \to 0$ then the theorems of this article can be straightforwardly modified to show that the resulting estimator is consistent.

Another direct extension is to apply the above methodology to the class of linear nonparametric smoothing estimators that produce estimates $\hat{\boldsymbol{f}}$ of the form $\hat{\boldsymbol{f}} = \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{y} = (y_0, \ldots, y_{n-1})^T$ and $\boldsymbol{H}$ is known as the "hat" or the "smoother" matrix. The kernel estimator considered in this article is a member of this class. Other class members include smoothing splines and penalized regression splines.

Extension to two–dimensional regularly spaced data setting (e.g., image data) is straightforward. Another possible extension of this work we are currently investigating is to construct similar KL

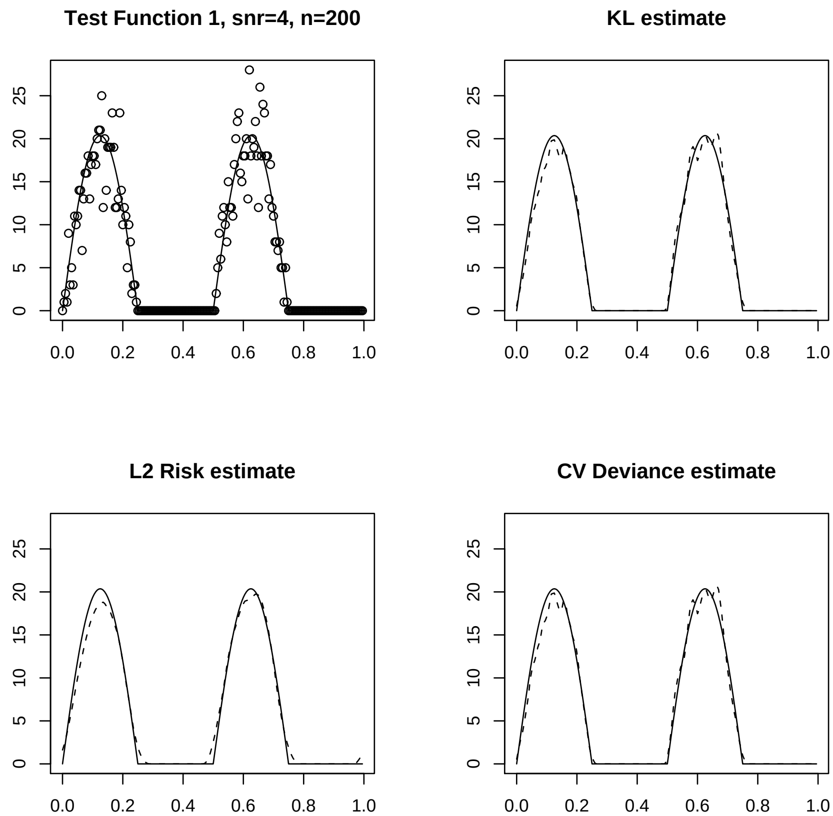

Figure 1: Visual inspection for Test Function 1 with $n = 200$. Top–left: true function (solid line) with noisy data points superimposed. Top–right: true function (solid line) with estimated function using $h = \hat{h}_{\mathrm{KL}}$ (broken line). Bottom–left: true function (solid line) with estimated function using $h = \hat{h}_{\mathrm{R}}$ (broken line). Bottom–right: true function (solid line) with estimated function using $h = \hat{h}_{\mathrm{DEV}}$ (broken line).

discrepancy estimators for the use in generalized linear and additive models.

Acknowledgement

The authors are grateful to the anonymous associate editor and the reviewers for their most constructive and helpful comments.

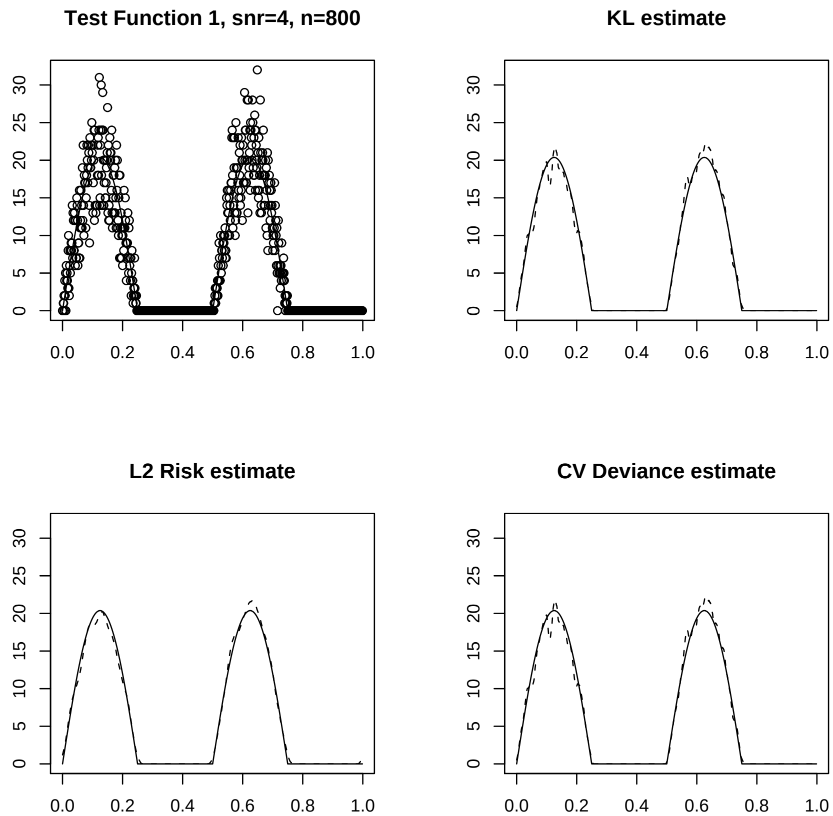

Figure 2: Similar to Figure 1 but for Test Function 1 with $n = 800$.

A Derivation of $\Delta_{\text{KL}}(\hat{f}, f)$

The KL discrepancy for measuring the distance between two discrete probability density functions (pdfs) $g_1(t)$ and $g_2(t)$ is defined as

$$d(g_1, g_2) = \sum_t g_1(t) \log \frac{g_1(t)}{g_2(t)}$$

(e.g., see Burnham & Anderson 1998). Note that $d(g_1, g_2) \neq d(g_2, g_1)$. For the current problem, in order to use $d(g_1, g_2)$ for comparing a true f and an estimate $\hat{f}$, one needs to compare them design point by design point. At design point x_j, the pdf $g_f(t)$ corresponding to f is Poisson with mean f_j. That is, $g_f(t) = e^{-f_j} f_j^t / t!$, $t = 0, 1, \ldots$. For $\hat{f}$, a natural candidate for the corresponding pdf is Poisson with mean $\hat{f}_j$. Denote this pdf as $g_{\hat{f}}(t)$, and thus $g_{\hat{f}}(t) = e^{-\hat{f}_j} \hat{f}_j^t / t!$, $t = 0, 1, \ldots$. We choose to measure the distance between f and $\hat{f}$ at x_j with

$$d(g_{\hat{f}}, g_f) = \sum_{t=0}^{\infty} g_{\hat{f}}(t) \log \frac{g_{\hat{f}}(t)}{g_f(t)} = \sum_{t=0}^{\infty} \frac{e^{-\hat{f}_j} \hat{f}_j^t}{t!} \log \frac{e^{-\hat{f}_j} \hat{f}_j^t / t!}{e^{-f_j} f_j^t / t!} = f_j - \hat{f}_j + \hat{f}_j (\log \hat{f}_j - \log f_j).$$

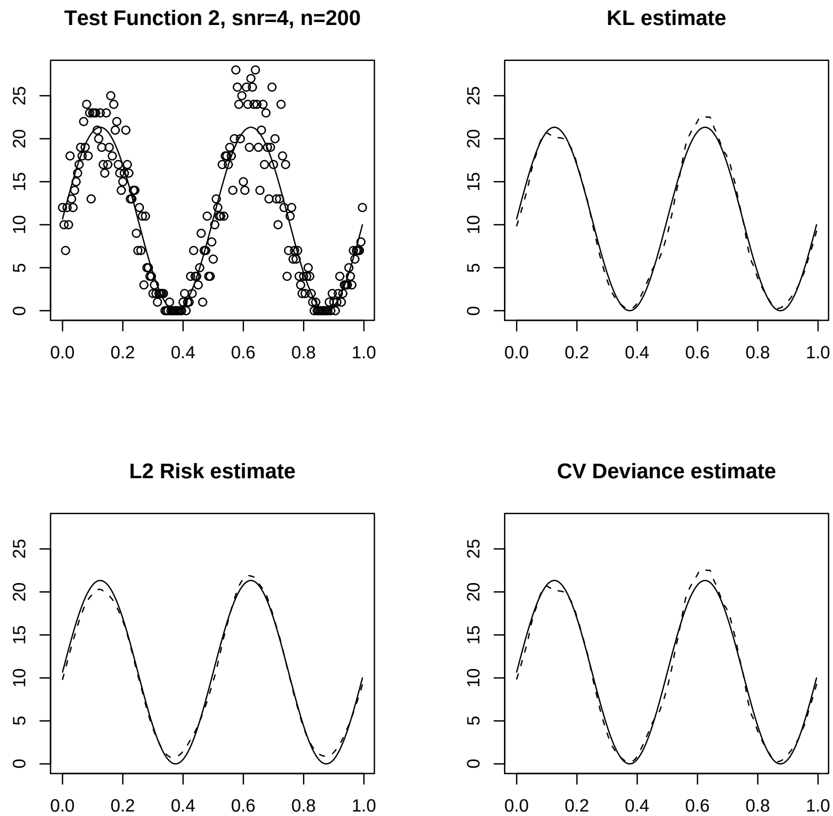

Figure 3: Similar to Figure 1 but for Test Function 2 with $n = 200$.

Upon summing over j we obtain $\Delta_{\text{KL}}(\hat{f}, f)$.

Notice that one could also use $\Delta_{\text{KL}}(f, \hat{f})$ (i.e., use $d(g_f, g_{\hat{f}})$) instead of $\Delta_{\text{KL}}(\hat{f}, f)$ (i.e., use $d(g_{\hat{f}}, g_f)$), but we choose $\Delta_{\text{KL}}(\hat{f}, f)$ for the following reason. Using the Taylor series approximation $1 - y + y \log y = (y-1)^2/2$ for $y \approx 1$, we obtain

$$\Delta_{\text{KL}}(\hat{f}, f) \approx \frac{1}{2n} \sum_{j=0}^{n-1} \frac{(f_j - \hat{f}_j)^2}{f_j}$$

and

$$\Delta_{\text{KL}}(f, \hat{f}) = \frac{1}{n} \sum_{j=0}^{n-1} \left\{ \hat{f}_j - f_j + f_j(\log f_j - \log \hat{f}_j) \right\} \approx \frac{1}{2n} \sum_{j=0}^{n-1} \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}.$$

Our belief is that $\Delta_{\text{KL}}(\hat{f}, f)$ is a better measure to use, as in the above approximation it uses a fixed quantity, the denominator term f_j, to adjust for the variance of $(f_j - \hat{f}_j)^2$ while $\Delta_{\text{KL}}(f, \hat{f})$ uses a random quantity $\hat{f}_j$. In addition, for the following reason $\Delta_{\text{KL}}(\hat{f}, f)$ is more desirable in the case when $f_j \approx 0$ for several consecutive j's. In this case it is quite possible that $\hat{f}_j = 0$ for some small values of bandwidth h, which causes $\Delta_{\text{KL}}(g_{f_j}, g_{\hat{f}_j}) = \infty$ while $\Delta_{\text{KL}}(g_{\hat{f}_j}, g_{f_j}) = f_j$, and

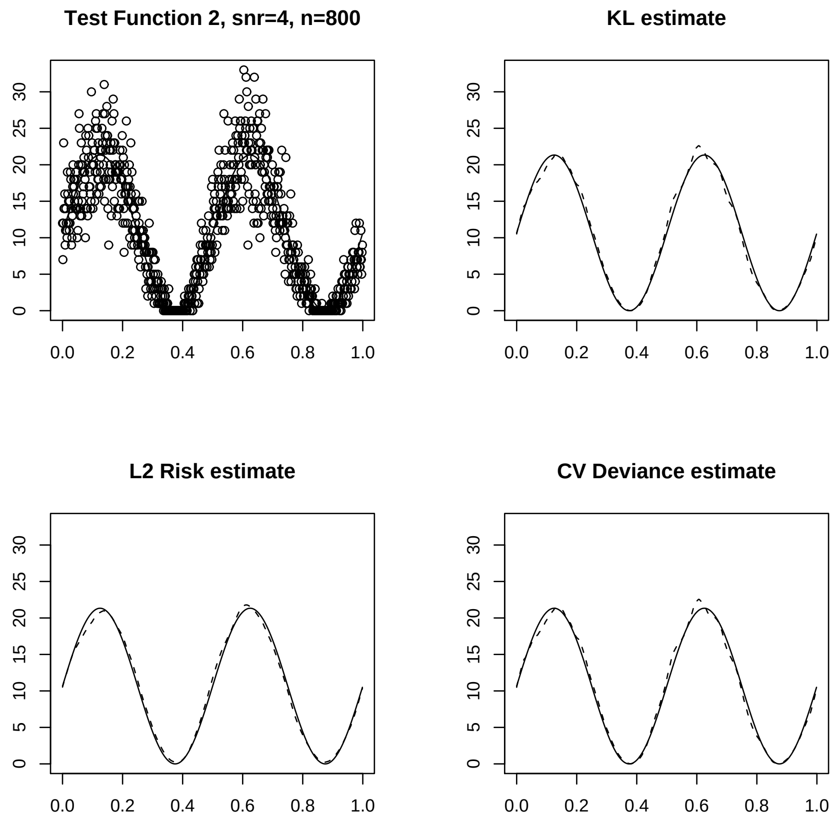

Figure 4: Similar to Figure 1 but for Test Function 2 with $n = 800$.

of course the latter is more reasonable. Thus in order to get a finite $\Delta_{\text{KL}}(g_{f_j}, g_{\hat{f}_j})$ and hence finite $\Delta_{\text{KL}}(f, \hat{f})$ the bandwidth will have to be large enough to guarantee $\hat{f}_j > 0$. This may possibly lead to oversmoothing in other parts of f. On the other hand $\Delta_{\text{KL}}(\hat{f}, f)$ does not suffer from this issue.

B Construction of $\hat{\Delta}^k_{\text{KL}}(h)$

This appendix outlines the construction of $\hat{\Delta}_{h,k}$. The goal is to find an unbiased estimator of $\Delta_{\text{KL}}(\hat{f}, f)$, which breaks down to the estimation of f_j and $\hat{f}_j \log f_j$. Estimation of f_j is straightforward as $E(y_j) = f_j$. As shown below, the estimation of $\hat{f}_j \log f_j$ can be further broken down to the estimation of $\log f_j$ and $f_j \log f_j$. However, this poses a bigger challenge as $\log f_j \approx -\infty$ whenever $f_j \approx 0$. We first work on $\log f_j$.

Here and in what follows let Y denote a Poisson(λ) random variable. Consider estimating $\log \lambda$ (i.e., $\log f_j$). The Taylor's series expansion of $\log y$ at the point λ is $\log y \approx \log \lambda + (y - \lambda)/\lambda - (y -$

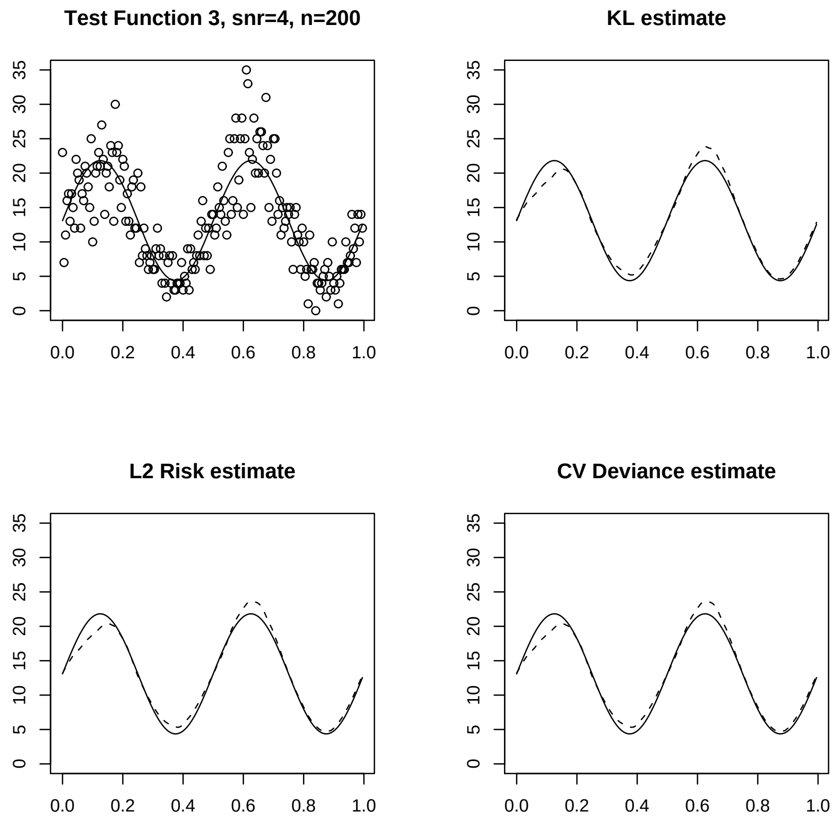

Figure 5: Similar to Figure 1 but for Test Function 3 with $n = 200$.

$\lambda)^2/(2\lambda^2)$, which leads to

$$E\{(\log Y)I_{\{Y>0\}}\} \approx \log\lambda - \frac{1}{2\lambda}. \tag{16}$$

This suggests estimating $\log\lambda$ by

$$\{\log Y - 1/(2Y)\}I_{\{Y>0\}}, \tag{17}$$

where the factor of $1/(2Y)$ is motivated by the fact that $E\{(2Y)^{-1}I_{\{Y>0\}}\} \approx 1/(2\lambda)$. The approximation in (16) works very well for large λ. However, the bias is not satisfactory for $\lambda < 10$. To correct this we suggest the following correction. Take an estimator

$$G = C_0 I_{\{Y=0\}} + \left\{\log Y - \frac{1}{2Y} + \frac{C_1}{Y^2} + \frac{C_2}{Y^3}\right\} I_{\{Y>0\}}, \tag{18}$$

and choose C_0, C_1 and C_2 to minimize $\int_1^\infty \{E(G) - \log\lambda\}^2\, d\lambda$, where $E(G)$ is considered as a function of λ. The motivation of this step is that the effect of the added terms is negligible for large values of λ. More precisely it is of the order $O(1/\lambda^2)$. At the same time the choice of C_0, C_1 and C_2 will guarantee improvement of the bias for small values of λ. We performed numerical

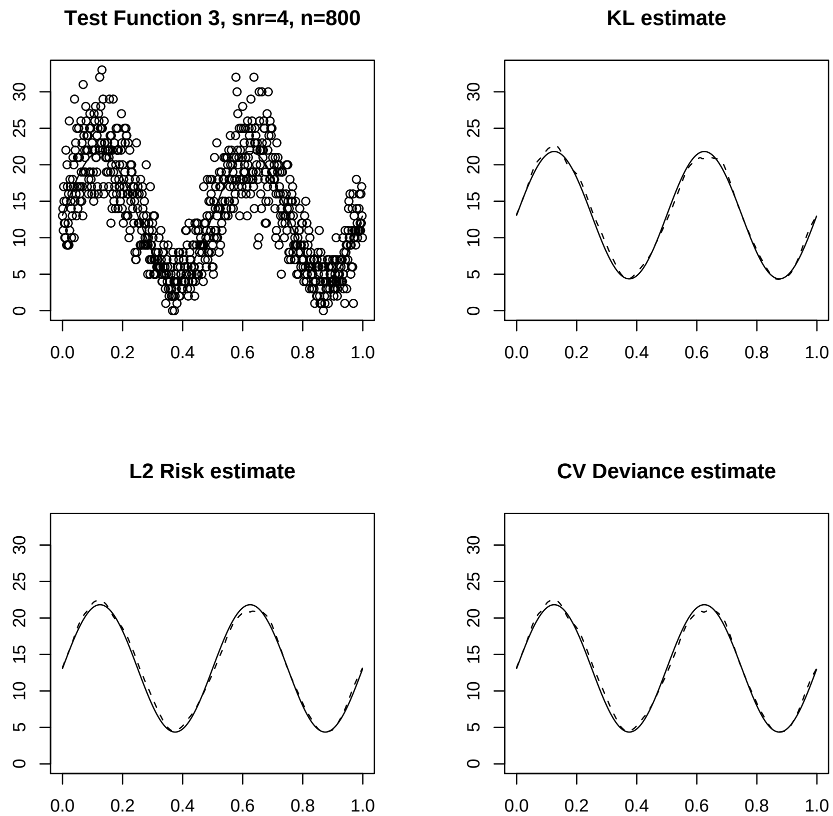

Figure 6: Similar to Figure 1 but for Test Function 3 with $n = 800$.

integration and obtain $C_0 = 2.10898$, $C_1 = 1.36177$ and $C_2 = 2.15204$. These constants improved the bias remarkably for $\lambda > 1$.

Recall that a major difficulty with estimating $\log f_j$ occurs when f_j is close to zero. To overcome this difficulty, we make use of the fact that if f is locally smooth, then $f_{j-k} \approx \ldots \approx f_{j+k}$ for small k. This implies that $y_{j-k}, \ldots, y_{j+k}$ are approximately independent and identically distributed as Poisson with mean f_j. Therefore $y_j^k = \sum_{m=-k}^{k} y_{j+m}$ has approximately Poisson distribution with mean $\lambda = (2k+1)f_j$. Now if k is large enough so that $\lambda > 1$, we have

$$E(G) \approx \log f_j + \log(2k+1). \tag{19}$$

Thus combining (18) and (19) we derive the estimator of $\log f_j$ as

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

Now we consider estimating $\lambda \log \lambda$ (or $f_j \log f_j$). The Taylor's series expansion of $y \log y$ at the point λ is $y \log y \approx \lambda \log \lambda + (y-\lambda)\ (1+\log\lambda) + (y-\lambda)^2/(2\lambda)$, which gives

$$E(Y \log Y) \approx \lambda \log \lambda + \frac{1}{2}. \tag{20}$$

Similarly as before we plug y_j^k into (20) and obtain

$$y_j^k \log y_j^k - \frac{1}{2} I_{\{y_j^k > 0\}} \approx (2k+1)\{f_j \log f_j - f_j \log(2k+1)\},$$

which leads to

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

as the estimator of $f_j \log f_j$ based on y_j^k.

To finish the derivation we decompose $\hat{f}_j \log f_j$ into two parts:

$$\hat{f}_j \log f_j = \sum_{m=-k}^{k} w_m y_{j+m} \log f_j + \left(\hat{f}_j - \sum_{m=-k}^{k} w_m y_{j+m} \right) \log f_j.$$

Since the expectation of the first part is approximately $f_j \log f_j \sum_{m=-k}^{k} w_m$, we estimate it by $\beta_j^k \sum_{m=-k}^{k} w_m$. Notice also that the first term of the second part and y_j^k are independent. Thus an approximately unbiased estimator of the second part is $\left(\hat{f}_j - \sum_{m=-k}^{k} w_m y_{j+m}\right) \alpha_j^k$. The parameter k, in a way, can be treated as a device for controlling the bias and variance of our estimator for $\hat{f}_j \log f_j$.

Finally putting the two parts together we have

$$\hat{f}_j \log f_j \approx \beta_j^k \sum_{m=-k}^{k} w_m + \left(\hat{f}_j - \sum_{m=-k}^{k} w_m y_{j+m} \right) \alpha_j^k.$$

This finishes the construction of $\hat{\Delta}_{\mathrm{KL}}^k(h)$, which is an approximately unbiased estimator of $\Delta_{\mathrm{KL}}(\hat{f}, f)$.

C Proof of Theorem 1

We first state and prove the following lemma. Let Y denote a Poisson(λ) random variable, and define residuals

$$r_1(\lambda) = E\left[\left\{ \log Y + \frac{0.5}{Y} - \frac{1.36177}{Y^2} + \frac{2.15204}{Y^3} \right\} I_{\{Y>0\}} - 2.10898 I_{\{Y=0\}} \right] - \log \lambda,$$

$$r_2(\lambda) = E\left(Y \log Y - \frac{1}{2} I_{\{Y>0\}} \right) - \lambda \log \lambda.$$

Lemma 1. *The following relations are true:*

$$E(\alpha_j^k) = \log \frac{f_j^k}{2k+1} + r_1(f_j^k), \tag{21}$$

$$E(\beta_j^k) = \frac{f_j^k}{2k+1} \log \frac{f_j^k}{2k+1} + \frac{r_2(f_j^k)}{2k+1}. \tag{22}$$

Furthermore as $\lambda \to \infty$:

$$r_1(\lambda) = O(1/\lambda^2), \tag{23}$$

$$r_2(\lambda) = O(1/\lambda). \tag{24}$$

Proofs of (21) and (22): Notice that y_j^k has a Poisson(f_j^k) distribution and direct calculation shows

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

Relation (22) follows immediately. Similarly one obtains

$$E\left(\alpha_j^k - \log \frac{f_j^k}{2k+1}\right) =$$
$$E\left[\left\{\log y_j^k + \frac{0.5}{y_j^k} - \frac{1.36177}{(y_j^k)^2} + \frac{2.15204}{(y_j^k)^3}\right\} I_{\{Y>0\}} - 2.10898 I_{\{Y=0\}}\right] - \log(f_j^k),$$

which implies (21).

Proof of (24): We first derive an upper bound for $r_2(\lambda)$. Using $\log y = \log \lambda + \log\{1 + (y-\lambda)/\lambda\}$ and $\log(1+y) \leq y - y^2/2 + y^3/3$ we get

$$\begin{aligned} E(Y \log Y) &= E(Y \log \lambda) + E\left\{Y \log\left(1 + \frac{Y-\lambda}{\lambda}\right)\right\} \\ &\leq \lambda \log \lambda + E\left[Y\left\{\frac{Y-\lambda}{\lambda} - \frac{1}{2}\left(\frac{Y-\lambda}{\lambda}\right)^2 + \frac{1}{3}\left(\frac{Y-\lambda}{\lambda}\right)^3\right\}\right] \\ &= \lambda \log \lambda + \frac{1}{2} + \frac{5}{6\lambda} + \frac{1}{3\lambda^2}, \end{aligned}$$

whence

$$r_2(\lambda) \leq \frac{5}{6\lambda} + \frac{1}{3\lambda^2} + \frac{1}{2} e^{-\lambda}.$$

Now we establish a lower bound for $r_2(\lambda)$, and we need two inequalities to proceed. The first inequality is, if $C > 0$ then $\log(1+y) \geq y - y^2/2 + y^3/3 - (1+C)y^4/4$ for $y > -D$, where $D > 0$ depends on C. The second inequality is a classical large deviation result, namely,

$P[(Y-\lambda)/\lambda \le -D] \le e^{-K\lambda}$, where K depends on D (e.g., see Grimmett & Stirzaker 2001, page 202). With these two inequalities, we proceed as

$$\begin{aligned}
E(Y\log Y) = E(Y\log\lambda) + E\left\{Y\log\left(1+\frac{Y-\lambda}{\lambda}\right) I_{\{Y>\lambda-D\lambda\}}\right\} \\
- E\left\{Y\log\left(1+\frac{Y-\lambda}{\lambda}\right) I_{\{Y\le\lambda-D\lambda\}}\right\} \\
&\ge \lambda\log\lambda + E\left\{Y\log\left(1+\frac{Y-\lambda}{\lambda}\right) I_{\{Y>\lambda-D\lambda\}}\right\} \\
&\quad + \min_{0\le x\le\lambda-D\lambda}\left\{x\log\left(1+\frac{x-\lambda}{\lambda}\right)\right\} P\left(\frac{Y-\lambda}{\lambda}\le -D\right) \\
&\ge E\left[Y\left\{\frac{Y-\lambda}{\lambda} - \frac{1}{2}\left(\frac{Y-\lambda}{\lambda}\right)^2 + \frac{1}{3}\left(\frac{Y-\lambda}{\lambda}\right)^3 - \frac{1+C}{4}\left(\frac{Y-\lambda}{\lambda}\right)^4\right\}\right] \\
&\quad + \lambda\log\lambda - \lambda e^{-K\lambda-1} \\
&= \lambda\log\lambda + \frac{1}{2} + \frac{1}{\lambda}\left(\frac{1}{12} - \frac{3C}{4}\right) + O(\frac{1}{\lambda^2})
\end{aligned}$$

and Equation (24) follows.

Proof of (23): Using similar arguments as above we conclude that

$$E(\log Y I_{\{Y>0\}}) = \log\lambda - \frac{1}{2\lambda} + O(\frac{1}{\lambda^2}).$$

Analogously we can write $x^{-1} = \lambda^{-1}\{1+(x-\lambda)/\lambda\}^{-1}$. It is again well-known that $1/(1+y) \ge 1-y$ and if $C>0$ than $1/(1+y) \le 1-y+(1+c)y^2$ for $y > -D$, where $D>0$ depends on C. From here

$$E\left(\frac{1}{Y}I_{\{Y>0\}}\right) \ge \frac{1}{\lambda}E\left(1-\frac{Y-\lambda}{\lambda}\right) I_{\{Y>0\}} = \frac{1}{\lambda} - \frac{2}{\lambda}e^{-\lambda}$$

and

$$\begin{aligned}
E(\frac{1}{Y}I_{\{Y>0\}}) &\le \frac{1}{\lambda}E(\frac{1}{1+\frac{Y-\lambda}{\lambda}}I_{\{Y>\lambda-D\lambda\}}) + \max_{1\le x\le\lambda-D\lambda}\frac{1}{x}P(1\le Y\le\lambda-D\lambda) \\
&\le \frac{1}{\lambda} + \frac{1+C}{\lambda^2} + e^{-K\lambda}.
\end{aligned}$$

Similar considerations show that

$$E(\frac{1}{Y^k}) = \frac{1}{\lambda^k} + O(\frac{1}{\lambda^{k+1}})$$

and Relation (23) follows by simple algebra. This completes proving Lemma 1 and we are now ready to give the proof for Theorem 1.

Proof of (7): To compute the bias consider $\hat{\Delta}^k_{\text{KL}}(h) - \Delta_{\text{KL}}(\hat{f}, f) = n^{-1}\sum_{j=0}^{n-1} S_j$ and decompose each summand S_j into four parts:

$$S_j = (y_j - f_j) - (\alpha_j^k - \log f_j)\sum_{|m|\geq k} w_m y_{j+m} - (\beta_j^k - f_j \log f_j)\sum_{|m|\leq k} w_m + \sum_{|m|\leq k} w_m(y_{j+m} - f_j)\log f_j. \quad (25)$$

Let us calculate $E(S_j)$ term by term:

$$E(y_j - f_j) = 0, \quad (26)$$

$$E\left\{(\alpha_j^k - \log f_j)\sum_{|m|\geq k} w_m y_{j+m}\right\} = \left\{r_1(f_j^k) - \log\frac{f_j}{f_j^k/(2k+1)}\right\}\sum_{|m|\geq k} w_m f_{j+m}, \quad (27)$$

$$E\left\{(\beta_j^k - f_j\log f_j)\sum_{|m|\leq k} w_m\right\} = r_2(f_j^k)\frac{1}{2k+1}\sum_{|m|\leq k} w_m + \left(\frac{f_j^k}{2k+1}\log\frac{f_j^k}{2k+1} - f_j\log f_j\right)\sum_{|m|\leq k} w_m, \quad (28)$$

$$E\left\{\sum_{|m|\leq k} w_m \log f_j (y_{j+m} - f_j)\right\} = \sum_{|m|\leq k} w_m (f_{j+m} - f_j)\log f_j. \quad (29)$$

Recall $M_1 = \max f$ and $M_2 = \min f$. Combining equations (26) through (29), observing the fact $\sum_{|m|\leq k} w_m \leq (2k+1)w_0$, and using inequalities

$$\left|\log\frac{y}{x}\right| \leq \frac{|x-y|}{y}, \quad \text{and} \quad |x\log x - y\log y| \leq |x-y||1+\log y| + \frac{|x-y|^2}{2y},$$

we get

$$|E(S_j)| \leq M_1 r_1(f_j^k) + w_0 r_2(f_j^k) + \frac{M_1}{M_2}\left|\frac{f_j^k}{2k+1} - f_j\right| + \left|\frac{f_j^k}{2k+1} - f_j\right|\{1 + \max(\log M_1, -\log M_2)\}(2k+1)w_0 + \frac{1}{2M_2}\left|\frac{f_j^k}{2k+1} - f_j\right|^2 (2k+1)w_0 + \max(\log M_1, -\log M_2)w_0\sum_{|m|\leq k}|f_{j+m} - f_j|.$$

Observe that $f_j^k \geq M_2(2k+1)$, and $w_0 \leq K'/b$, where K' is a constant depending only on the kernel K. Combining these observations with Lemma 1 and the fact that f is Lipschitz with constant D one obtains (7).

Proof of (8): By noting that the w_m's are zero when $|m| > b_n/2$, and that the observations are independent, we have

$$\begin{aligned}\operatorname{var}\{\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f)\} =& \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{i=0}^{n-1}\operatorname{cov}(S_i, S_j) = \frac{1}{n^2}\sum_{|i-j|\le b}\operatorname{cov}(S_i, S_j)\\ \le& \frac{1}{n^2}\sum_{|i-j|\le b}\{\operatorname{var}(S_i)\operatorname{var}(S_j)\}^{\frac{1}{2}}. \end{aligned} \tag{30}$$

Therefore we need to prove that $\operatorname{var}(S_i)$ is bounded. Using Equation (25) we get:

$$\begin{aligned}\operatorname{var}(S_j) \le 4\operatorname{var}(y_j) + 4\operatorname{var}\left\{(\alpha^k_j - \log f_j)\sum_{|m|\ge k} w_m y_{j+m}\right\}\\ + 4\left(\sum_{|m|\le k} w_m\right)^2 \operatorname{var}(\beta^k_j) + 4(\log f_j)^2\operatorname{var}\left(\sum_{|m|\le k} w_m y_{j+m}\right).\end{aligned} \tag{31}$$

Notice that the large deviations considerations mentioned before give us that

$$P(M_2 - \epsilon < y^k_j < M_1 + \epsilon) \ge 1 - e^{-ck} \text{ for } M_2 > \epsilon > 0 \text{ and some } c > 0.$$

This combined with the definition of α^k_j, β^k_j and the fact that y^k_j has a Poisson distribution immediately imply that both $\operatorname{var}(\beta^k_j)$ and $\operatorname{var}(\alpha^k_j)$ are bounded by a constant $\tilde{C}$ that depends on M_1 and M_2.

Let us now calculate each part of (31) separately:

$$\operatorname{var}(y_j) \le M_1, \tag{32}$$

$$\left(\sum_{|m|\le k} w_m\right)^2 \operatorname{var}(\beta^k_j) \le C'_3\left(\frac{k}{b}\right)^2, \tag{33}$$

$$(\log f_j)^2\operatorname{var}\left(\sum_{|m|\le k} w_m y_{j+m}\right) \le C'_4\frac{2k+1}{b^2}M_1\max(-\log M_2, \log M_1)^2. \tag{34}$$

The only part that requires a little bit more attention is:

$$\begin{aligned}\operatorname{var}\left\{(\alpha^k_j - \log f_j)\sum_{|m|\ge k} w_m y_{j+m}\right\} = E\left\{\left(\sum_{|m|\ge k} w_m y_{j+m}\right)^2\right\}\operatorname{var}(\alpha^k_j)\\ + \operatorname{var}\left(\sum_{|m|\ge k} w_m y_{j+m}\right)\left\{E(\alpha^k_j - \log f_j)\right\}^2\\ \le (M_1^2 + M_1\sum_k w_m^2)\tilde{C} + M_1(\sum_k w_m^2)\left[\log\left\{\frac{f^k_j/(2k+1)}{f_j}\right\} + r_1(f^k_j)\right]^2.\end{aligned} \tag{35}$$

Now by substituting (32) through (35) into (31) one can see that there is a universal constant C depending on the function f through M_1, M_2 and the Lipschitz constant D such that $\mathrm{var}(S_j) \leq C$. Therefore from (30) we arrive (8).

Proof of (9): We will need to use the following relations:

$$E(\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f))^2 = O\left(\frac{b_n}{n} + \frac{1}{k^4}\right), \tag{36}$$

$$\mathrm{var}\,\Delta_{\mathrm{KL}}(\hat{f}, f) = O\left(\frac{b_n}{n}\right), \tag{37}$$

$$E\{\Delta_{\mathrm{KL}}(\hat{f}, f)\} \geq \frac{C}{b_n} + o(\frac{1}{b_n}). \tag{38}$$

Recall, $b_n = o\{\min(n^{1/3}, k_n^2)\}$. Thus both $(b_n/n)^{1/2} = o(1/b_n)$, $1/k_n^2 = o(1/b_n)$ and there is r_n such that $r_n = o(1/b_n)$ and $b_n/n + 1/k^4 = o(r_n^2)$. Fix $\varepsilon > 0$ and calculate:

$$P\left(\left|\frac{\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f)}{\Delta_{\mathrm{KL}}(\hat{f}, f)}\right| > \varepsilon\right) < P\left(\left|\frac{\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f)}{r_n}\right| > \varepsilon\right) + P(\Delta_{\mathrm{KL}}(\hat{f}, f) < r_n).$$

By combining (36), (38), (37), the Markov's and Chebyshev's inequalities we get

$$P\left(\left|\frac{\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f)}{r_n}\right| > \varepsilon\right) < \frac{E(\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f))^2}{\varepsilon^2 r_n^2} \to 0,$$

$$P(\Delta_{\mathrm{KL}}(\hat{f}, f) < r_n) < P(|\Delta_{\mathrm{KL}}(\hat{f}, f) - E\Delta_{\mathrm{KL}}(\hat{f}, f)| > E\Delta_{\mathrm{KL}}(\hat{f}, f) - r_n)$$
$$< \frac{\mathrm{var}\,\Delta_{\mathrm{KL}}(\hat{f}, f)}{(E\Delta_{\mathrm{KL}}(\hat{f}, f) - r_n)^2} \to 0.$$

This proves (9). The only remaining part is to verify (36), (38) and (37).

Proof of (36): Recall

$$E(\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f))^2 = \mathrm{var}\,E(\hat{\Delta}^k_{\mathrm{KL}}(h) - \Delta_{\mathrm{KL}}(\hat{f}, f))^2 + (E\hat{\Delta}^k_{\mathrm{KL}}(h) - E\Delta_{\mathrm{KL}}(\hat{f}, f))^2.$$

Since $k_n < b_n < n$ the right-hand-side of (7) is of the order $1/k^2 + k/n$. Thus equations (7) and (8) imply (36).

Proof of (37): The proof follows along the same steps as proof of (8) and we omit the details.

Proof of (38): We need two inequalities to prove (38). Define $l(y) = 1 - y - y\log(y)$ and hence $\Delta_{\mathrm{KL}}(\hat{f}, f) = n^{-1}\sum_{j=0}^{n-1} f_j\ l(\hat{f}_j/f_j)$. Denote $S_n = \{\max|\hat{f}/f - 1| > 1/2)\}$. It follows from a large deviation argument simular to Louani (1999) $P(S_n) \leq ne^{-Kb} \to 0$. By applying the Taylor approximation $l(y) \approx \frac{1}{2}(y-1)^2$ to $l(\hat{f}_j/f_j)$ and using the assumption that f is bounded away from 0 and ∞, we obtain our first inequality:

$$C\frac{\left(\hat{f}_j - f_j\right)^2}{f_j} I_{S_n} \leq f_j\ l\left(\frac{\hat{f}_j}{f_j}\right) \quad \text{with appropriate} \quad 0 < C < \frac{1}{2}. \tag{39}$$

To get the second inequality calculate

$$\frac{E(\hat{f}_j - f_j)^2}{f_j} = 2f_j - 2E(\hat{f}_j) + \frac{E(\hat{f}_j^2) - f_j^2}{f_j}. \tag{40}$$

Notice that

$$E(\hat{f}_j^2) = \left\{E(\hat{f}_j)\right\}^2 + \sum_m w_m^2 f_{j+m}. \tag{41}$$

Thus combining (40) and (41) we get

$$\frac{1}{n}\sum_j \frac{E(\hat{f}_j - f_j)^2}{f_j} = \frac{1}{n}\sum_j \left[\frac{f_j + E(\hat{f}_j)}{f_j}\{E(\hat{f}_j) - f_j\} + \sum_m w_m^2 \frac{f_{j+m}}{f_j}\right].$$

Since the weights w_m's are zero when $|m| > b_n/2$ and the function f is Lipschitz, we have

$$\{E(\hat{f}_j) - f_j\} = \sum_{m=-n}^{2n-1} w_{m-j}\{E(y_m) - f_j\} \le D\frac{b_n}{n}. \tag{42}$$

Also notice that

$$\sum_{m=-n}^{2n-1} w_{m-j}^2 \approx \frac{L\int K^2(\omega)d\omega}{b_n}, \tag{43}$$

where L is the length of the support of K. Combining Equations (40) to (43) we can conclude that there is a constant $D_1 > 0$ depending on K such that

$$\frac{1}{n}\sum_j \frac{E(\hat{f}_j - f_j)^2}{f_j} \ge \frac{D_1}{b_n} + O\left\{\frac{b_n}{n}\right\}.$$

Equation (38) then follows from this, our first inequality (39) and the Cauchy-Schwartz inequality.

D Proof of Theorem 2

Notice that $\hat{\Delta}_{\text{R}}(h) - \Delta_{\text{R}}(\hat{f}, f) = n^{-1}\sum_{j=0}^{n-1} Z_j$ where

$$Z_j = 2y_j(f_j - \hat{f}_j) + (y_j)^2 - (f_j)^2 + (2w_0 - 1)y_j. \tag{44}$$

Using independence of y_j we get

$$E(Z_j) = 2E\{y_j w_0(f_j - y_j)\} + f_j + (2w_0 - 1)f_j = 0,$$

proving (13).

Let us now turn our attention to the variance. Notice that w_m's are zero when $|m| > b_n/2$, whence the independence of observations implies:

$$\begin{aligned}\operatorname{var}\{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)\} =& \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{i=0}^{n-1}\operatorname{cov}(Z_i, Z_j) = \frac{1}{n^2}\sum_{|i-j|\le b}\operatorname{cov}(Z_i, Z_j)\\ \le& \frac{1}{n^2}\sum_{|i-j|\le b}\{\operatorname{var}(Z_i)\operatorname{var}(Z_j)\}^{\frac{1}{2}}.\end{aligned} \tag{45}$$

Therefore we need to prove that $\operatorname{var}(Z_i)$ is bounded. Using equation (44) we get:

$$\operatorname{var}(Z_j) \le 3\operatorname{var}\{2y_j(f_j - \hat{f}_j)\} + 3\operatorname{var}\{(y_j)^2\} + 3\operatorname{var}\{(2w_0 - 1)y_j\}.$$

Since the function f is bounded and if Y has Poisson(λ) distribution then $E(Y^4) = \lambda + 7\,\lambda^2 + 6\,\lambda^3 + \lambda^4$ we conclude that there is a universal constant C depending on the function f through $\max f$, such that $\operatorname{var}(Z_j) \le C$. This and the fact that (45) has no more than $(2b+1)n$ non-zero terms implies

$$\operatorname{var}\{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)\} \le C\frac{b}{n}$$

which is (14).

Finally, we will prove (15). Notice first that arguments almost identical to those in the proof of (14) imply

$$\operatorname{var}\Delta_{\mathrm{R}}(\hat{f}, f) \le C'\frac{b}{n}. \tag{46}$$

Second, we will estimate $E\Delta_{\mathrm{R}}(\hat{f}, f)$. Substituting

$$E(\hat{f}_j^2) = \left\{E(\hat{f}_j)\right\}^2 + \sum_m w_m^2 f_{j+m} \tag{47}$$

into

$$E\{(\hat{f}_j - f_j)^2\} = (f_j)^2 - 2f_jE(\hat{f}_j) + E(\hat{f}_j^2) \tag{48}$$

we get

$$\frac{1}{n}\sum_j E(\hat{f}_j - f_j)^2 = \frac{1}{n}\sum_j\left[\{f_j - E(\hat{f}_j)\}^2 + \sum_m w_m^2 f_{j+m}\right].$$

The assumption that function f is Lipschitz assures that $|f_m - f_j| < D|m-j|/n$. Since the weights w_m's are zero when $|m| > b_n/2$, we have

$$\{E(\hat{f}_j) - f_j\}^2 = \left\{\sum_{m=-n}^{2n-1} w_{m-j}(f_m - f_j)\right\}^2 \le \left(D\frac{b_n}{n}\right)^2. \tag{49}$$

Combining Equations (43) and (48) we can conclude that there is a constant $D_2 > 0$ depending on K and M, such that

$$\frac{1}{n}\sum_j E\{(\hat{f}_j - f_j)^2\} \geq \frac{D_2}{b_n} + O\left[\left\{\frac{b_n}{n}\right\}^2\right]. \tag{50}$$

Recall, $b_n = o(n^{1/3})$. Thus $(b_n/n)^{1/2} = o(1/b_n)$ and there is r_n such that $r_n = o(1/b_n)$ and $(b_n/n)^{1/2} = o(r_n)$. Fix $\varepsilon > 0$ and calculate:

$$P\left(\left|\frac{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)}{\Delta_{\mathrm{R}}(\hat{f}, f)}\right| > \varepsilon\right) < P\left(\left|\frac{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)}{r_n}\right| > \varepsilon\right) + P(\Delta_{\mathrm{R}}(\hat{f}, f) < r_n).$$

By combining (13), (14), (46), (50), and the Chebyshev's inequality we get

$$P\left(\left|\frac{\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)}{r_n}\right| > \varepsilon\right) < \frac{\operatorname{var}\hat{\Delta}_{\mathrm{R}}(h) - \Delta_{\mathrm{R}}(\hat{f}, f)}{\varepsilon^2 r_n^2} < \frac{Cb_n/n}{\varepsilon^2 r_n^2} \to 0,$$

$$\begin{aligned} P(\Delta_{\mathrm{R}}(\hat{f}, f) < r_n) &< P(|\Delta_{\mathrm{R}}(\hat{f}, f) - E\Delta_{\mathrm{R}}(\hat{f}, f)| > E\Delta_{\mathrm{R}}(\hat{f}, f) - r_n) \\ &< \frac{\operatorname{var}\Delta_{\mathrm{R}}(\hat{f}, f)}{(E\Delta_{\mathrm{R}}(\hat{f}, f) - r_n)^2} < \frac{C' b_n/n}{(D_2/b_n - D(b_n/n)^2 - r_n)^2} \to 0. \end{aligned}$$

This proves (15).

References

Burnham, K. P. & Anderson, D. R. (1998), *Model Selection and Inference: A Practical Information-Theoretic Approach*, Springer-Verlag New York Inc.

Gasser, T., Műller, H.-G. & Mammitzsch, V. (1985), 'Kernels for nonparametric curve estimation', *Journal of the Royal Statistical Society Series B* **47**, 238–252.

Grimmett, G. R. & Stirzaker, D. R. (2001), *Probability and random processes*, third edn, The Clarendon Press Oxford University Press, New York.

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

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

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hudson, H. M. (1978), 'A natural identity for exponential families with applications in multiparameter estimation', *The Annals of Statistics* **6**, 473–484.

Hudson, H. M. (1985), 'Adaptive estimator for simultaneous estimation of Poisson means', *The Annals of Statistics* **13**, 246–261.

Hudson, H. M. & Lee, T. C. M. (1998), 'Maximum Likelihood restoration and choice of smoothing parameter in deconvolution of image data subject to Poisson noise', *Computational Statistics and Data Analysis* **26**, 393–410.

Kolaczyk, E. D. (1997), 'Nonparametric estimation of gamma–ray burst intensities using Haar wavelets', *The Astrophysical Journal* **483**, 340–349.

Kolaczyk, E. D. (1999*a*), 'Bayesian multi–scale models for poisson processes', *Journal of the American Statistical Association* **94**, 920–933.

Kolaczyk, E. D. (1999*b*), 'Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds', *Statistica Sinica* **9**, 119–135.

La Riviere, P. J. & Pan, X. (2000), 'Nonparametric regression sinogram smoothing using a roughness–penalized Poisson likelihood objective function', *IEEE Transactions on Medical Imaging* **19**, 773–786.

Linhart, H. & Zucchini, W. (1986), *Model Selection*, New York: Wiley.

Louani, D. (1999), 'Some large deviations limit theorems in conditional nonparametric statistics', *Statistics* **33**(2), 171–196.

Nowak, R. D. & Kolaczyk, E. D. (2000), 'A statistical multiscale framework for poisson inverse problems', *IEEE Transactions on Information Theory* **46**, 1811–1825.

Pawitan, Y. & O'Sullivan, F. (1993), 'Data dependent bandwidth selection for emission computed tomography reconstruction', *IEEE Transactions on Medical Imaging* **12**, 167–172.

van Dyk, D. A., Connors, A., Kashyap, V. L. & Siemiginowska, A. (2001), 'Analysis of energy spectra with low photon counts via Bayesian posterior simulation', *The Astrophysical Journal* **548**, 224–243.

Wand, M. P. & Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

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