J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 357–375

Smoothing for discrete-valued time series

Zongwu Cai

University of North Carolina, Charlotte, USA

and Qiwei Yao and Wenyang Zhang

London School of Economics and Political Science, UK

[Received March 2000. Final revision September 2000]

Summary. We deal with smoothed estimators for conditional probability functions of discrete-valued time series $\{Y_t\}$ under two different settings. When the conditional distribution of Y_t given its lagged values falls in a parametric family and depends on exogenous random variables, a smoothed maximum (partial) likelihood estimator for the unknown parameter is proposed. While there is no prior information on the distribution, various nonparametric estimation methods have been compared and the adjusted Nadaraya–Watson estimator stands out as it shares the advantages of both Nadaraya–Watson and local linear regression estimators. The asymptotic normality of the estimators proposed has been established in the manner of sparse asymptotics, which shows that the smoothed methods proposed outperform their conventional, unsmoothed, parametric counterparts under very mild conditions. Simulation results lend further support to this assertion. Finally, the new method is illustrated via a real data set concerning the relationship between the number of daily hospital admissions and the levels of pollutants in Hong Kong in 1994–1995. An *ad hoc* model selection procedure based on a local Akaike information criterion is proposed to select the significant pollutant indices.

Keywords: Adjusted Nadaraya–Watson estimator; α-mixing; Discrete-valued time series; Local Akaike information criterion; Local linear smoother; Local partial likelihood; Nonparametric estimation; Smoothed maximum likelihood estimation; Sparse asymptotics

1. Introduction

Let $\{Y_t\}$ be a strictly stationary discrete-valued time series. We apply smoothing techniques to estimate the conditional probability function of Y_t given its lagged values in both parametric and nonparametric settings. The methods proposed are applicable when Y_t is either quantitative or ordinal categorical. Such a variable can arise as a discretization of an underlying continuous variable or as an inherently discrete, but ordered, set of categories (Simonoff (1996), section 6.1). In the former case, discrete values of Y_t have real physical meanings. If Y_t indeed represents a discretization of a continuous variable with a smooth density function, the probability of $Y_t = i$ will be close to that of $Y_t = i + \Delta$ for small integer Δ. In the latter, the different values denote different categories which have a natural ordering (e.g. *very bad*, *bad*, *neutral*, *good* and *very good*). An observation falling in one particular cell provides information about the probability of falling in its neighbour's. Therefore, smoothing makes sense since we may assume that the probability function is 'continuous' (see conditions (A3) and (B2) later) in both cases. The improvement using smoothing is most evident when the distribution is *sparse* in the sense that the probability that Y_t falls in each

Address for correspondence: Qiwei Yao, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: q.yao@lse.ac.uk

1369–7412/01/63357

cell is small. To highlight this phenomenon, we develop an asymptotic approximation under the *sparse asymptotics* framework which assumes that the maximum value of the probability function converges to 0 when the sample size goes to ∞; see Simonoff (1985), Hall and Titterington (1987) and Simonoff (1996), section 6.2. We shall show that the smoothed estimators proposed have smaller asymptotic mean-squared errors than those of conventional (non-smoothed) parametric estimators when the underlying distribution is sparse; see remarks 2, 5 and 7 later.

In Section 2, we assume that the conditional distribution of Y_t given its past falls in a parametric family with the parameter depending on the value of Y_{t-1} and also an 'exogenous' variable. By assuming that the parameter is 'continuous' in Y_{t-1} (see condition (A3) and remark 1, part (c), later), we estimate the parameter by maximizing a local (i.e. smoothed) partial likelihood function. We propose a simple and intuitively appealing bootstrap method to choose the bandwidth. The asymptotic normality of the estimator is established. In Section 3, various nonparametric kernel estimation methods for the conditional probability function of Y_t given Y_{t-1} are discussed. We are in favour of the adjusted Nadaraya–Watson (ANW) estimator (Hall and Presnell, 1999; Hall *et al.*, 1999) since it enjoys the same first-order asymptotic properties as the local linear estimator and is always a proper probability function itself. The nonparametric setting in Section 3 is similar to the local polynomial estimation of continuous conditional density functions considered by Fan *et al.* (1996). However, the asymptotic theory is different since we estimate a probability function which converges to 0 itself. In Section 4, the methods proposed are illustrated through two simulated examples and a Hong Kong air pollution–disease data set, which can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

Although it appears to us that smoothing techniques have not been used in analysing discrete-valued time series data before, there is a substantial literature on their application to discrete data analysis. Aitchison and Aitken (1976), Bowman (1980), Titterington (1980) and Hall (1981) appear to be among the earliest. Lucid reviews on research in this direction can be found in Simonoff (1995) and Simonoff (1996), chapter 6. The latest developments include Dong and Simonoff (1994) on boundary-corrected kernel estimation for sparse multinomial distributions, Aerts *et al.* (1997) and Simonoff (1998) on local polynomial estimation of multinomial tables and Faddy and Jones (1998) on semiparametric smoothing for discrete probability functions.

2. Smoothed Parametric Estimation

2.1. Maximum local partial likelihood estimator

Suppose that discrete-valued time series Y_t is influenced by an exogenous variable $\mathbf{X}_t$, $\{(\mathbf{X}_t, Y_t)\}$ forms a strictly stationary process and Y_t takes non-negative integer values. We assume that the conditional probability of $Y_t = j$ given $(\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots, Y_{t-1} = i, Y_{t-2}, \ldots)$ is $p(j; \mathbf{X}_t, \boldsymbol{\beta}_i)$ which depends on $(\mathbf{X}_t, Y_{t-1} = i)$ only, where $p(\cdot; \cdot, \cdot)$ is of a given form and $\boldsymbol{\beta}_i$ is an unknown parameter vector. For example, the conditional distribution could be Poisson with mean $\mu(\mathbf{X}_t^T \boldsymbol{\beta}_i)$, where $\mu(\cdot)$ is a known link function. Given observations $\{(\mathbf{X}_t, Y_t), 1 \leqslant t \leqslant n\}$, the log-conditional-likelihood function given $\mathbf{X}_1$ and Y_1 is

$$\sum_{t=2}^{n} \log\{p(Y_t; \mathbf{X}_t, \boldsymbol{\beta}_{Y_{t-1}})\} + \sum_{t=2}^{n} \log\{f(\mathbf{X}_t; \mathbf{X}_{t-1}, \ldots, \mathbf{X}_1, Y_{t-1}, \ldots, Y_1)\},$$

where $f(X;\ Z)$ denotes the conditional probability density of X given Z. By maximizing the first sum in this expression, we obtain the maximum (partial) likelihood estimators for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots$. Such an estimator for $\boldsymbol{\beta}_i$ is in fact derived by maximizing

$$\sum_{t=2}^{n} \log\{p(Y_t;\ \mathbf{X}_t,\ \boldsymbol{\beta}_i)\}\, I(Y_{t-1} = i), \tag{2.1}$$

which depends on the pairs $(Y_{t-1},\ Y_t)$ with $Y_{t-1} = i$ only. To make more efficient use of the available data, we define a smoothed (partial) likelihood function of $\boldsymbol{\beta}_i$ by replacing the indicator function $I(Y_{t-1} = i)$ by a kernel function:

$$\sum_{t=2}^{n} \log\{p(Y_t;\ \mathbf{X}_t,\ \boldsymbol{\beta}_i)\}\, K_{n,h}(Y_{t-1} - i). \tag{2.2}$$

Maximizing this smoothed likelihood, we obtain a smoothed estimator. In this expression, $K(\cdot)$ is a kernel function, $K_{n,h}(x) = h^{-1}\, K(\delta_n x/h)$, $h > 0$, is a bandwidth which controls the amount of smoothing used in estimation and $\delta_n > 0$ reflects the sparseness of the underlying distribution; see, condition (A3) and remark 1, parts (b) and (c) in Section 2.3.

In view of the more attractive asymptotic properties of the local linear smoother relative to the local constant version (Fan *et al.*, 1998), we propose to use the local linear estimator $\hat{\boldsymbol{\beta}}_i$, where $(\hat{\boldsymbol{\beta}}_i,\ \hat{\mathbf{a}})$ is the maximizer of

$$l_n = \sum_{t=2}^{n} \log[p\{Y_t;\ \mathbf{X}_t,\ \boldsymbol{\beta}_i + \delta_n \mathbf{a}(Y_{t-1} - i)\}]\, K_{n,h}(Y_{t-1} - i). \tag{2.3}$$

Obviously, this approach can be applied to the case when the conditional probability of Y_t given its past is of the form $p(Y_t;\ \boldsymbol{\theta}_{Y_{t-1}})$ with known function p and unknown parameter $\boldsymbol{\theta}$. Further, the smoothing can also be incorporated into the quasi-likelihood approach of Wedderburn (1974) in an obvious manner. In fact the proposed smoothed (partial) likelihood function (2.2), although derived under a specified time series context, is of the form of the *local likelihood functions* explored by, among others, Tibshirani and Hastie (1987) and Fan *et al.* (1998) for independent observations.

2.2. Bandwidth selection

The bandwidth h plays an important role in smoothing estimation. Most existing bandwidth selection methods were originally designed for continuous independent data, although some of them can be adapted to handle dependence in time series. For the problem discussed in this section, there is no natural way to derive an analogue to the cross-validation method or its variations. Instead, we propose a simple bootstrap approach to choose h which is easy to implement and takes into account the dependence of the data in resampling. The bandwidth selected may be variable in the sense that different bandwidths may be used to estimate different $\boldsymbol{\beta}_i$s. The method is similar in spirit to those used by Hall *et al.* (1999) and Polonik and Yao (2000) for the estimation of (continuous) conditional distribution functions and conditional minimum volume sets.

We draw bootstrap samples conditionally on the given data $\{\mathbf{X}_t\}$ as follows. Let $\tilde{\boldsymbol{\beta}}_i$ be the parametric estimator obtained from maximizing expression (2.1) for all i. (Some initial moving average smoothing may be applied to $\{\tilde{\boldsymbol{\beta}}_i\}$ in the case that some cells contain few observations. Alternatively, $\tilde{\boldsymbol{\beta}}_i$ can be obtained nonparametrically by maximizing expression (2.2) with a small bandwidth such that the biases are small.) Draw Y_0^* from the empirical (marginal) probability function of $\{Y_t\}$. For $t = 1, \ldots, n$, draw Y_t^* from the (discrete) probability function $p(\cdot;\, \mathbf{X}_t,\, \tilde{\boldsymbol{\beta}}_{Y_{t-1}^*})$. Define $\hat{\boldsymbol{\beta}}_i^* \equiv \hat{\boldsymbol{\beta}}_i^*(h)$ in the same way as $\hat{\boldsymbol{\beta}}_i$ with $\{(\mathbf{X}_t,\, Y_t)\}$ replaced by $\{(\mathbf{X}_t,\, Y_t^*)\}$. To estimate $\boldsymbol{\beta}_i$ for a particular i, we choose the h which minimizes the conditional expectation

$$E[||\hat{\boldsymbol{\beta}}_i^*(h) - \tilde{\boldsymbol{\beta}}_i||\,|\,\{(\mathbf{X}_t,\, Y_t)\}].$$

To speed up the computation, we may use one single bandwidth for the estimation of all the $\boldsymbol{\beta}_i$, and this single bandwidth minimizes

$$M(h) = \sum_i \hat{\pi}_i\, E[||\hat{\boldsymbol{\beta}}_i^*(h) - \tilde{\boldsymbol{\beta}}_i||\,|\,\{(\mathbf{X}_t,\, Y_t)\}], \tag{2.4}$$

where $\hat{\pi}_i$ is the relative frequency estimate for the marginal probability $P(Y_t = i)$. Some initial moving average may be applied in case some cells contain no observations. For example, we may replace $\hat{\pi}_i$ by the moving average of its three nearest neighbours (including itself) with the weights $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{4}$.

2.3. Theoretical properties

We write $l(y;\, \mathbf{x},\, \boldsymbol{\beta}) = \log\{p(y;\, \mathbf{x},\, \boldsymbol{\beta})\}$, and let

$$\dot{l}(y;\, \mathbf{x},\, \boldsymbol{\beta}) = \frac{\partial l(y;\, \mathbf{x},\, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

$$\ddot{l}(y;\, \mathbf{x},\, \boldsymbol{\beta}) = \frac{\partial^2 l(y;\, \mathbf{x},\, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^{\mathrm{T}}}.$$

Define $\mu_j = \int u^j\, K(u)\, \mathrm{d}u$ and $\nu_j = \int\, u^j\, K(u)^2\, \mathrm{d}u$. For matrix $\mathbf{A} = (a_{ij})$, $||\mathbf{A}|| = (\Sigma\, a_{ij}^2)^{1/2}$. We use C to denote a finite positive constant which may be different at different places. We state some regularity conditions first.

Condition (A1). The parameter $\boldsymbol{\beta}_i$ is identifiable in the sense that $p(\cdot;\, \mathbf{X}_t,\, \mathbf{z}_1) \not\equiv p(\cdot;\, \mathbf{X}_t,\, \mathbf{z}_2)$ for any $\mathbf{z}_1 \neq \mathbf{z}_2$ and

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^{\mathrm{T}}} E\{p(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_i) | \mathbf{X}_t,\, Y_{t-1} = i\} = E\left\{ \frac{\partial^2}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^{\mathrm{T}}} p(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_i) | \mathbf{X}_t,\, Y_{t-1} = i \right\}.$$

Further, all the third partial derivatives of $p(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_i)$ with respect to $\boldsymbol{\beta}_i$ are bounded by a random variable, say $M(Y_t,\, \mathbf{X}_t)$, and $E\{M(Y_t,\, \mathbf{X}_t) | \mathbf{X}_t,\, Y_{t-1} = i\}$ is finite.

Condition (A2). $\boldsymbol{\Sigma}_i \equiv -E[E\{\ddot{l}(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_{Y_{t-1}}) | \mathbf{X}_t,\, Y_{t-1}\} | Y_{t-1} = i]$ is a positive definite matrix. Further, for some $\gamma > 2$,

$$E\{||\dot{l}(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_i)||^{\gamma} + ||\ddot{l}(Y_t;\, \mathbf{X}_t,\, \boldsymbol{\beta}_i)||^{\gamma}\} < \infty,$$

and $||\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j|| \leqslant C\delta_n |j - i|$.

Condition (A3). For $i = 0, 1, \ldots, \pi_i \equiv P(Y_t = i) = \delta_n \int_i^{i+1} g(\delta_n x)\, \mathrm{d}x$, where $g(\cdot)$ is a density function on $[0, \infty)$. Further, $\boldsymbol{\beta}_i = \mathbf{b}(\delta_n i)$ and both $g(\cdot)$ and

$$\ddot{\mathbf{b}}(x) \equiv \left(\frac{\partial}{\partial x} \right)^2 \mathbf{b}(x)$$

are continuous in a neighbourhood of $\delta_n i$.

Condition (A4). The kernel function $K(\cdot)$ is bounded, symmetric and compactly supported.

Condition (A5). The process $\{\mathbf{X}_t, Y_t\}$ is α mixing with the mixing coefficient satisfying the condition $\alpha(k) = O(k^{-\beta})$, where $\beta > 2(\gamma - 1)/(\gamma - 2)$ for γ given in condition (A2) above.

Condition (A6). As $n \to \infty$, $h \to 0$, $nh \longrightarrow \infty$ and $\delta_n \to 0$.

Remark 1.

(a) Both condition (A1) and condition (A2) are the standard conditions to ensure that the (unsmoothed) maximum likelihood estimator is consistent and asymptotically normal; see Lehmann and Casella (1998), section 6.3. We need both for the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$ as well. Together with conditions (A3)–(A6), they also ensure that the equation

$$\sum_{t=2}^{n} (1, \delta_n(Y_{t-1} - i))^{\mathrm{T}} \otimes \dot{l}\{Y_t; \mathbf{X}_t, \boldsymbol{\beta}_i + \mathbf{a}\delta_n(Y_{t-1} - i)\}\, K_{n,h}(Y_{t-1} - i) = 0 \tag{2.5}$$

has a solution which is a consistent estimator for $\boldsymbol{\beta}_i$ (see theorem 1, part (a), below), where $\otimes$ denotes the matrix Kronecker product. We call $\hat{\boldsymbol{\beta}}_i$ such a solution hereafter.

(b) Condition (A3) assumes that $\pi_i = O(\delta_n)$. Note that $\delta_n \to 0$ as $n \to 0$. This reflects the fact that the sparse asymptotics depict the performance of an estimator when the probability that Y_t falls into each cell is small. This is the case when the smoothing is most relevant.

(c) The smooth condition imposed on both $\boldsymbol{\beta}_i$ and $g(\cdot)$ in condition (A3) reflects the fact that the method proposed is designed for the cases when the conditional probability function of Y_t given Y_{t-1} is 'continuous' in Y_{t-1}, which makes smoothing estimation relevant. The assumption that $\ddot{\mathbf{b}}(\cdot)$ is continuous ensures a nice asymptotic formula for the bias of $\hat{\boldsymbol{\beta}}_i$. If we replace it by

$$||\boldsymbol{\beta}_i - \boldsymbol{\beta}_j|| \leqslant C\delta_n|i - j|,$$

$\hat{\boldsymbol{\beta}}_i$ is still asymptotically normal but with a bias of the order h.

(d) The constant δ_n is introduced to reflect the sparseness of data. It does not add any extra complication in estimating $\boldsymbol{\beta}_i$s. In fact in a practical implementation we may fix δ_n at any value; the resulting discrepancy will be absorbed in the estimation of bandwidth h automatically; see equation (2.3).

(e) The requirement in condition (A4) that $K(\cdot)$ be compactly supported is imposed for brevity of the proofs and can be removed at the cost of lengthier arguments. In particular, the Gaussian kernel is allowed.

(f) α-mixing is one of the weakest mixing conditions for weakly dependent stochastic processes. Stationary time series or Markov chains fulfilling certain (mild) conditions are α mixing with exponentially decaying coefficients; see section 2.6.1 of Fan and Yao (2002). In contrast, the assumption on the convergence rate of $\alpha(k)$ in condition (A5) is not the weakest possible and is imposed to simplify the proof.

Theorem 1. Let conditions (A1)–(A6) hold. Suppose that $x_i = i\delta_n$ is bounded away from both 0 and ∞ as $n \to \infty$, and $g(x_i) > 0$. Then the following assertions hold.

(a) Equation (2.5) admits a root $\hat{\boldsymbol{\beta}}_i$ which converges to $\boldsymbol{\beta}_i$ in probability.

(b) For any $\hat{\boldsymbol{\beta}}_i$ fulfilling (a),

$$\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = (nh)^{-1/2}\mathcal{N} + \frac{\mu_2 h^2}{2}\ddot{\mathbf{b}}(x_i) + o_P\{(nh)^{-1/2} + h^2\},$$

where $\mathcal{N}$ is a normal random vector with mean 0 and variance matrix $\nu_0\, g(x_i)^{-1}\boldsymbol{\Sigma}_i^{-1}$.

Remark 2. Under condition (A3), $P(Y_t = i) = O(\delta_n)$. It can be shown that under this condition the parametric estimator $\tilde{\boldsymbol{\beta}}_i$ derived from maximizing expression (2.1) is asymptotically normal with mean 0 and variance of the order $1/n\delta_n$ instead of $1/n$, since the expected number of observations falling in each cell is of the order $n\delta_n$ instead of n. Theorem 1 shows that the asymptotic variance of the smoothed estimator $\hat{\boldsymbol{\beta}}_i$ is of the order $1/nh$. Hence the asymptotic variance of $\hat{\boldsymbol{\beta}}_i$ converges to 0 *faster* than that of $\tilde{\boldsymbol{\beta}}_i$ for any bandwidth h for which $h/\delta_n \to \infty$. Theorem 1 indicates that the optimal bandwidth which minimizes the approximate mean-squared error (AMSE) is of the order $n^{-1/5}$. (We define the AMSE as the squared asymptotic bias plus asymptotic variance up to the first order.) With the optimal bandwidth, the AMSE of $\hat{\boldsymbol{\beta}}_i$ is of the order $n^{-4/5}$ which converges to 0 faster than the AMSE of $\tilde{\boldsymbol{\beta}}_i$ as long as $n\delta_n^5 \to 0$. This is a very mild condition. In the case that Y_t takes finite m values, $\delta_n = O(1/m)$. Therefore, the smoothed estimator $\hat{\boldsymbol{\beta}}_i$ will outperform parametric estimator $\tilde{\boldsymbol{\beta}}_i$ as long as $n = o(m^5)$.

Remark 3. In the case that the exogenous variable $\mathbf{X}_t$ is absent, theorem 1 implies that

$$\begin{aligned}\frac{p(j, \hat{\boldsymbol{\beta}}_i)}{p(j, \boldsymbol{\beta}_i)} - 1 &= p(j, \boldsymbol{\beta}_i)^{-1}\,\dot{p}(j, \boldsymbol{\beta}_i)^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\{1 + o_P(1)\}\\ &= p(j, \boldsymbol{\beta}_i)^{-1}\,\dot{p}(j, \boldsymbol{\beta}_i)^{\mathrm{T}}\left[(nh)^{-1/2}\mathcal{N} + \frac{\mu_2 h^2}{2}\ddot{\mathbf{b}}(x_i) + o_P\{(nh)^{-1/2} + h^2\}\right],\end{aligned}$$

where $p(j, \boldsymbol{\beta}_i) = P(Y_t = j | Y_{t-1} = i)$ and $\dot{p}(j, \boldsymbol{\beta}_i) = \partial p(j, \boldsymbol{\beta}_i)/\partial\boldsymbol{\beta}_i$.

The proof of theorem 1, as well as that of theorem 2 in Section 3.3, is obtainable from

`http://www.blackwellpublishers.co.uk/rss/`

3. Nonparametric estimation

3.1. Estimators

We assume that data $\{Y_t,\ 1 \leqslant t \leqslant n\}$ are available from a strictly stationary discrete-valued time series, where Y_t takes integer values $\{1, \ldots, m\}$ with $m < \infty$. Of interest is to estimate the conditional probability function

$$p_{ij} = P(Y_t = j \,|\, Y_{t-1} = i), \qquad i, j = 1, \ldots, m.$$

In fact the methods proposed below may also be extended to estimate the higher dimensional conditional probability function

$$p_{i_1, \ldots, i_k, j} = P(Y_t = j | Y_{t-1} = i_1, \ldots, Y_{t-k} = i_k)$$

for $k > 1$, which could be appealing when $\{Y_t\}$ is a kth-order Markov chain. However, such an extension is of limited practical value owing to the difficulties that are associated with the curse of dimensionality.

Note that $p_{ij} = E\{I(Y_t = j) | Y_{t-1} = i\}$. This naturally leads to the NW estimator

$$\tilde{p}_{ij} = \frac{\sum_{t=2}^{n} I(Y_t = j)\, K_{mh}(Y_{t-1} - i)}{\sum_{t=2}^{n} K_{mh}(Y_{t-1} - i)}, \tag{3.1}$$

where $K(\cdot)$ is a kernel function, $K_{mh}(\cdot) = (mh)^{-1} K(\cdot/mh)$ and $h > 0$ is a bandwidth which controls the amount of smoothness in estimation. The extreme case of $h = 0$ corresponds to the relative frequency estimate

$$\check{p}_{ij} = \frac{\sum_{t=2} I(Y_{t-1} = i,\ Y_t = j)}{\sum_{t=2} I(Y_{t-1} = i)} \qquad \left(\frac{0}{0} \equiv 0\right). \tag{3.2}$$

When $h > 0$, we use the information contained in the data (Y_{t-1}, Y_t) with $Y_t = j$ and Y_{t-1} close to i to estimate p_{ij}.

It is well known that an NW estimator exhibits boundary bias at both ends, which was addressed in Simonoff (1995) for categorical data, and it has a more complicated asymptotic bias formula (see remark 8 later). To reduce these disadvantages, the obvious correction is to use the local linear estimator, defined as $\breve{p}_{ij} = \hat{\alpha}$, because of its nice properties (Fan, 1993) such as mathematical efficiency, bias reduction and adaptation of edge effects, where $(\hat{\alpha}, \hat{\beta})$ minimizes

$$\sum_{t=2}^{n} \{I(Y_t = j) - \alpha - \beta(Y_{t-1} - i)\}^2 K_{mh}(Y_{t-1} - i). \tag{3.3}$$

However, $\breve{p}_{ij}$ is not constrained to lie between 0 and 1. In this respect, the NW method is superior, since $\tilde{p}_{ij} \in [0, 1]$ and $\Sigma_j \tilde{p}_{ij} = 1$. We propose the ANW estimator by combining the advantages from both NW and local linear estimators. The method was first introduced by Hall and Presnell (1999) for estimating conditional mean functions and was used by Hall *et al.* (1999) for the estimation of conditional distribution functions of continuous random variables.

The ANW approach is as follows. Let $w_t(i)$, for $1 \leqslant t \leqslant n$, denote weights (functions of the data $Y_1, \ldots, Y_{n-1}$, as well as i) with the property that

$$\left.\begin{aligned} w_t(i) &\geqslant 0, \\ \sum_{t=2}^{n} w_t(i) &= 1, \\ \sum_{t=2}^{n} w_t(i)(i - Y_{t-1})\, K_{mh}(Y_{t-1} - i) &= 0. \end{aligned}\right\} \tag{3.4}$$

Of course, weights $w_t(i)$ satisfying these conditions are not uniquely defined, and we specify them concisely by maximizing $\Pi_t\, w_t(i)$ subject to the constraints. As a result, $w_t(i)$ can be expressed as

$$w_t(i) = (n-1)^{-1} \{1 + \lambda\,(i - Y_{t-1})\, K_{mh}(Y_{t-1} - i)\}^{-1},$$

where λ, a function of the data and i, is uniquely defined by expressions (3.4). It is easily computed by using a Newton–Raphson scheme. Then, the ANW estimator is defined by

$$\hat{p}_{ij} = \frac{\sum_{t=2}^{n} I(Y_t = j)\, w_t(i)\, K_{mh}(Y_{t-1} - i)}{\sum_{t=2}^{n} w_t(i)\, K_{mh}(Y_{t-1} - i)}. \tag{3.5}$$

Note particularly that $0 \leqslant \hat{p}_{ij} \leqslant 1$ and $\Sigma_j \hat{p}_{ij} = 1$. We show in theorem 2 later that $\hat{p}_{ij}$ is first order equivalent to a local linear estimator which does not enjoy either of these properties.

3.2. Bandwidth selection

We may apply the generalized cross-validation (GCV) proposed by Wahba (1977) and Craven and Wahba (1979) to choose h. By ignoring the dependence on $\{I(Y_t = j)\}$ of the weight functions $\{w_i(t)\}$, it follows from equation (3.5) that

$$(\hat{p}_{Y_1,j}, \ldots, \hat{p}_{Y_{n-1},j})^{\tau} = \mathbf{H}\{I(Y_2 = j), \ldots, I(Y_n = j)\}^{\tau},$$

where $\mathbf{H} = \mathbf{H}(h)$ is the $(n-1) \times (n-1)$ hat matrix. GCV selects h which minimizes

$$\mathrm{GCV}_j(h) = \left\{1 - \frac{\mathrm{tr}(\mathbf{H})}{n}\right\}^{-2} \sum_{t=2}^{n} \{I(Y_t = j) - \hat{p}_{Y_{t-1},j}\}^2.$$

It is easy to see that

$$\mathrm{tr}(\mathbf{H}) = K_{mh}(0) \sum_{l=2}^{n} \left\{ w_l(Y_{l-1}) \Big/ \sum_{t=2}^{n} w_t(Y_{l-1})\, K_{mh}(Y_{t-1} - Y_{l-1}) \right\}.$$

It is known that GCV has a tendency to undersmooth when $\mathrm{tr}(\mathbf{H})/n$ is large, particularly for small sample sizes. To overcome this shortcoming, Hurvich *et al.* (1998) proposed the use of the corrected version of the Akaike information criterion, AICC:

$$\mathrm{AICC}_j(h) = \log\left[\sum_{t=2}^{n} \{I(Y_t = j) - \hat{p}_{Y_{t-1},j}\}^2\right] + \frac{2\{\mathrm{tr}(\mathbf{H}) + 1\}}{(n-1) - \mathrm{tr}(\mathbf{H}) - 2}.$$

It is easy to see that GCV and AICC are about the same when $\mathrm{tr}(\mathbf{H})/n$ is small, which is typically the case in the context of analysing sparse discrete data.

Alternatively, the bootstrap approach described in Section 2.2 may also be adapted as follows. We generate a Markov chain $\{Y_t^*\}$ with the transition probability $\{\check{p}_{ij}\}$ defined in expression (2.1). Let $\hat{p}_{ij}^* \equiv \hat{p}_{ij}^*(h)$ be the estimator based on data $\{Y_t^*, 1 \leqslant t \leqslant n\}$, defined in the same manner as $\hat{p}_{ij}$. We use the bandwidth h in the estimator $\hat{p}_{ij}$, which minimizes the conditional expectation

$$E\left[\sum_{j=1}^{m} |\hat{p}_{ij}^*(h) - \check{p}_{ij}| \,|\, \{Y_t\}\right].$$

3.3. Theoretical properties

We impose the following regularity conditions. Write $x_i = i/m$ for $0 \leqslant i \leqslant m$.

Condition (B1). As $n \to \infty$, $h \to 0$, $m \to \infty$, $nh/m \to \infty$, $m^2 h^3 \to \infty$ and $nh^2 \to \infty$.

Condition (B2). For $1 \leqslant i, j \leqslant m$,

$$\pi_{ij} \equiv P\{Y_{t-1} = i,\ Y_t = j\} = \frac{1}{m}\int_{x_{i-1}}^{x_i} \psi_j(x)\,\mathrm{d}x, \tag{3.6}$$

where ψ_j is a positive function defined on [0, 1] and has two continuous derivatives in a neighbourhood of x_i. Further, the second derivative of the density function

$$\psi(x) \equiv m^{-1}\sum_{j=1}^{m}\psi_j(x)$$

is bounded by a constant independent of m in the same neighbourhood.

Condition (B3). $K(\cdot)$ is a bounded, symmetric density function with a compact support.

Condition (B4). The process $\{Y_t\}$ is strictly stationary and α mixing with the mixing coefficient $\alpha(k) = O(k^{-\beta})$ as $k \to \infty$, where $\beta > 2$ is a constant.

Remark 4. Note that $p_{ij} = \pi_{ij}/\pi_i$, where $\pi_i = \int_{x_{i-1}}^{x_i} \psi(x)\,\mathrm{d}x$. It is easy to see that condition (B2) implies that $p_{ij} = O(m^{-1})$. Since we deal only with sparse distributions, we assume that the number of categories $m \to \infty$ as the sample size $n \to \infty$. To pursue good asymptotic properties we assume that ψ_j (and therefore also ψ) has two continuous derivatives. As long as ψ_j is Lipschitz continuous in a neighbourhood of x_i, our estimators are still asymptotically normal but with larger biases (of the order of h).

We present only the asymptotic normality for the ANW estimator $\hat{p}_{ij}$ in the theorem below and compare it with other methods in the discussion following. Let

$$\varphi_j(x) = \psi_j(x)/\psi(x)$$

and μ_j and ν_j be the same as in Section 2.3. We denote by $\dot{\varphi}(\cdot)$ and $\ddot{\varphi}(\cdot)$ the first and second derivatives of $\varphi(\cdot)$ respectively.

Theorem 2. Suppose that conditions (B1)–(B4) hold and $x_i = i/m$ is bounded away from both 0 and 1 as $n \to \infty$. Then, for any $1 \leqslant j \leqslant m$,

$$\frac{\hat{p}_{ij}}{p_{ij}} - 1 = \left\{\frac{m\nu_0}{nh\,\psi_j(x_i)}\right\}^{1/2} Z + h^2\frac{\mu_2\,\ddot{\varphi}_j(x_i)}{2\,\varphi_j(x_i)} + o_P\left\{\left(\frac{m}{nh}\right)^{1/2} + h^2\right\} + O_P(m^{-1}), \tag{3.7}$$

where Z denotes a standard normal random variable.

Remark 5. Since $p_{ij} \to 0$, we consider the asymptotic normality of $\hat{p}_{ij}/p_{ij}$ instead of p_{ij}. Note that the number of observations falling in each category is of the order n/m, which can be viewed as the equivalent sample size in a usual asymptotic setting where $p_{ij} > 0$ is fixed (i.e. does not converge to 0 as $n \to \infty$). This explains that the convergence rate in theorem 2 is $(nh/m)^{1/2}$ instead of the conventional $(nh)^{1/2}$. Consequently, the optimal bandwidth h which minimizes the AMSE is of the order $(m/n)^{1/5}$.

Remark 6. If conditions (B2) and (B4) hold and both n/m and m tend to ∞, it holds that $\check{p}_{ij}/p_{ij} - 1$ is asymptotically normal with mean 0 and asymptotic variance of the order m^2/n. Note that the asymptotic variance of $\hat{p}_{ij}/p_{ij}$ is of the order m/nh. Hence, with any h such that $mh \to \infty$, the asymptotic variance of the smoothed estimator $\hat{p}_{ij}$ converges to 0 *faster* than that of the unsmoothed estimator $\check{p}_{ij}$. If we use the optimal bandwidth $h_{\text{opt}} \propto (m/n)^{1/5}$, for which $mh \to \infty$ under the very mild restriction $n = o(m^6)$, this assertion on the asymptotic variance also holds for the AMSE.

Remark 7. The local linear estimator $\breve{p}_{ij}$ derived from expression (3.3) admits the same asymptotic expression as equation (3.7). For the NW estimator $\tilde{p}_{ij}$ defined in equation (3.1), the asymptotic expression still holds if we replace the second term on the right-hand side of equation (3.7) (i.e. the bias) by

$$\frac{1}{2}h^2\mu_2\left\{\frac{\ddot{\varphi}_j(x_i)/\varphi_j(x_i)+2\,\dot{\varphi}_i(x_i)\,\dot{\psi}(x_i)}{\psi(x_i)\,\varphi_j(x_i)}\right\},$$

which has one more term. Since all smoothed estimators $\hat{p}_{ij}$, $\breve{p}_{ij}$ and $\tilde{p}_{ij}$ share the same asymptotic variance and the same order (i.e. $O(h^2)$) biases, the assertions on the superiority over unsmoothed estimator $\check{p}_{ij}$ in remark 6 above are also valid for $\breve{p}_{ij}$ and $\tilde{p}_{ij}$.

Remark 8. Let i be a boundary point, i.e. $i/m = c\,h$ for some $c \in (0, 1)$. Then it can be proved that

$$\frac{\hat{p}_{ij}}{p_{ij}}-1=\left(\frac{m}{nh}\right)^{1/2}\frac{\eta_2(c)^{1/2}}{\eta_1(c)\{\varphi_j(0)\,\psi(0)\}^{1/2}}Z+h^2\frac{\eta_0(c)\,\ddot{\varphi}_j(0)}{2\,\eta_1(c)\,\varphi_j(0)}+o_P\left\{\left(\frac{m}{nh}\right)^{1/2}+h^2\right\}+O_P(m^{-1}),$$

$$\begin{aligned}\frac{\breve{p}_{ij}}{p_{ij}}-1={}&\left(\frac{m}{nh}\right)^{1/2}\frac{\left\{\int_{-c}^{1}(\mu_{c,2}-\mu_{c,1}u)^2\,K(u)^2\,\mathrm{d}u\right\}^{1/2}}{(\mu_{c,0}\mu_{c,2}-\mu_{c,1}^2)\{\varphi_j(0)\,\psi(0)\}^{1/2}}Z+h^2\frac{\ddot{\varphi}_j(0)}{2\varphi_j(0)}\frac{\mu_{c,2}^2-\mu_{c,1}\mu_{c,3}}{\mu_{c,0}\mu_{c,2}-\mu_{c,1}^2}\\&+o_P\left\{\left(\frac{m}{nh}\right)^{1/2}+h^2\right\}+O_P(m^{-1}),\end{aligned}$$

$$\frac{\tilde{p}_{ij}}{p_{ij}}-1=\left(\frac{m}{nh}\right)^{1/2}\frac{\left\{\int_{-c}^{1}K(u)^2\,\mathrm{d}u\right\}^{1/2}}{\mu_{c,0}\{\varphi_j(0)\,\psi(0)\}^{1/2}}Z+h\frac{\mu_{c,1}\,\ddot{\varphi}_j(0)}{\mu_{c,0}\,\varphi_j(0)}+o_P\left\{\left(\frac{m}{nh}\right)^{1/2}+h\right\}+O_P(m^{-1}),$$

where $\mu_{c,k}=\int_{-c}^{1}u^k\,K(u)\,\mathrm{d}u$, and

$$\eta_0(c)=\int_{-c}^{1}\frac{u^2\,K(u)}{1-\lambda_c u\,K(u)}\mathrm{d}u,$$

$$\eta_k(c)=\int_{-c}^{1}\frac{K(u)^k}{\{1-\lambda_c u\,K(u)\}^k}\mathrm{d}u \qquad (k=1,\,2).$$

In these expressions, λ_c is the root of the equation $\int_{-c}^{1}u\,K(u)/\{1-\lambda u\,K(u)\}\,\mathrm{d}u=0$. These results confirm that both the ANW estimator and the local linear estimator are also boundary adaptive in the sparse asymptotics in the sense that the asymptotic variances and biases are of the same orders as at the inner points. Therefore, both $\hat{p}_{ij}$ and $\breve{p}_{ij}$ perform better than the unsmoothed estimator $\check{p}_{ij}$ even at the boundary points; see remark 6 above.

Remark 9. In the scenario described in remark 3, we may also apply nonparametric estimation for $p_{ij}\equiv p(j,\,\beta_i)$ if Y_t takes finite values $1, \ldots, m$. Theorem 2 shows that the bias of the resulting estimator is of the order h^2, which is the same as the smoothed parametric estimator; see remark 3. The asymptotic approximation for the variance is

$$V_1(j,\,i) \equiv \mathrm{var}\left(\frac{\hat{p}_{ij}}{p_{ij}} - 1\right) \approx \frac{m}{nh}\frac{\nu_0}{\psi_j(i/m)} = \frac{1}{nh}\frac{\nu_0}{mp_{ij}\pi_i}\{1+o(1)\}.$$

The last equality follows from the relationship that $\psi_j(i/m) \sim m^2\pi_{ij} = m^2 p_{ij}\pi_i$. In contrast, the smoothed estimation yields the asymptotic variance

$$\begin{aligned} V_2(j,\,i) &\equiv \mathrm{var}\left\{\frac{p(j,\,\hat{\boldsymbol{\beta}}_i)}{p(j,\,\boldsymbol{\beta}_i)} - 1\right\} \approx \frac{\nu_0}{nh\,g(\delta_n i)\,p^2(j,\,\boldsymbol{\beta}_i)}\,\dot{p}(j,\,\boldsymbol{\beta}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\,\dot{p}(j,\,\boldsymbol{\beta}_i) \\ &= \frac{\nu_0\delta_n}{nh\pi_i\,p^2(j,\,\boldsymbol{\beta}_i)}\,\dot{p}(j,\,\boldsymbol{\beta}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\,\dot{p}(j,\,\boldsymbol{\beta}_i)\{1+o(1)\}; \end{aligned}$$

see remark 3 and theorem 1. To facilitate the comparison between two estimators, we consider the average asymptotic approximations

$$E\{V_1(Y_t,\,i)|\,Y_{t-1} = i\} \approx \frac{\nu_0}{nh\pi_i}$$

and

$$E\{V_2(Y_t,\,i)|\,Y_{t-1} = i\} \approx \frac{\delta_n\nu}{nh\pi_i}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\,E\left\{\frac{\dot{p}(Y_t,\,\boldsymbol{\beta}_i)\,\dot{p}(Y_t,\,\boldsymbol{\beta}_i)^{\mathrm{T}}}{p^2(Y_t,\,\boldsymbol{\beta}_i)}\,\middle|\,Y_{t-1} = i\right\}\right] = \frac{\nu_0}{nh\pi_i}\,\delta_n d,$$

where d is the dimensionality of $\boldsymbol{\beta}$, which is a fixed constant. This shows that the ratio of the variance for the smoothed parametric estimator to its nonparametric counterpart converges to 0 at the rate δ_n. Therefore we should use the parametric approach whenever there are grounds to. Our empirical study also confirms the superior performance of the smoothed parametric estimation; see example 2 in Section 4.

4. Numerical properties

To assess the finite sample performance of the smoothed estimators proposed, we apply them to two simulated examples and one real data set. For the simulation models, we compare the smoothed estimators with their parametric counterparts. We also compare the nonparametric and parametric smoothed estimators in example 2. Throughout this section, the Epanechnickov kernel $K(u) = 0.75(1-u^2)\,I(|u| \leqslant 1)$ is used.

4.1. *Example 1*

First we consider a Poisson time series model constructed as follows. Let $\{X_t\}$ be a sequence of independent identically distributed random variables from a uniform $[-1,\,1]$ distribution. Given $\{(X_{s+1},\,Y_s),\,s \leqslant t\}$, the conditional distribution of Y_{t+1} is Poisson with the mean $\lambda(X_{t+1},\,Y_t)$, where $\lambda(x,\,i) = \exp(\beta_{1i} + \beta_{2i}x)$, and

$$\beta_{1i} = 4 - 2.49\,\exp\{-(i-25)^2/512\}, \qquad \beta_{2i} = -1.5\,\sin(\pi i/25).$$

Obviously, $\{(X_t,\,Y_t)\}$ is a homogeneous Markov chain. Fig. 1(a) plots a sample of time series $\{Y_t\}$ of length 200. We repeat the simulation 400 times for each of the sample sizes $n =$ 100, 200, 400. For each realization, we calculate the smoothed estimator $(\hat{\beta}_{1i},\,\hat{\beta}_{2i})$ which maximizes a smoothed likelihood function defined as in equation (2.3), as well as the parametric estimator $(\tilde{\beta}_{1i},\,\tilde{\beta}_{2i})$ which is derived by maximizing the likelihood function defined as in expression (2.1). For the smoothed estimator, we set $\delta_n = 0.1$ and use the bandwidth h

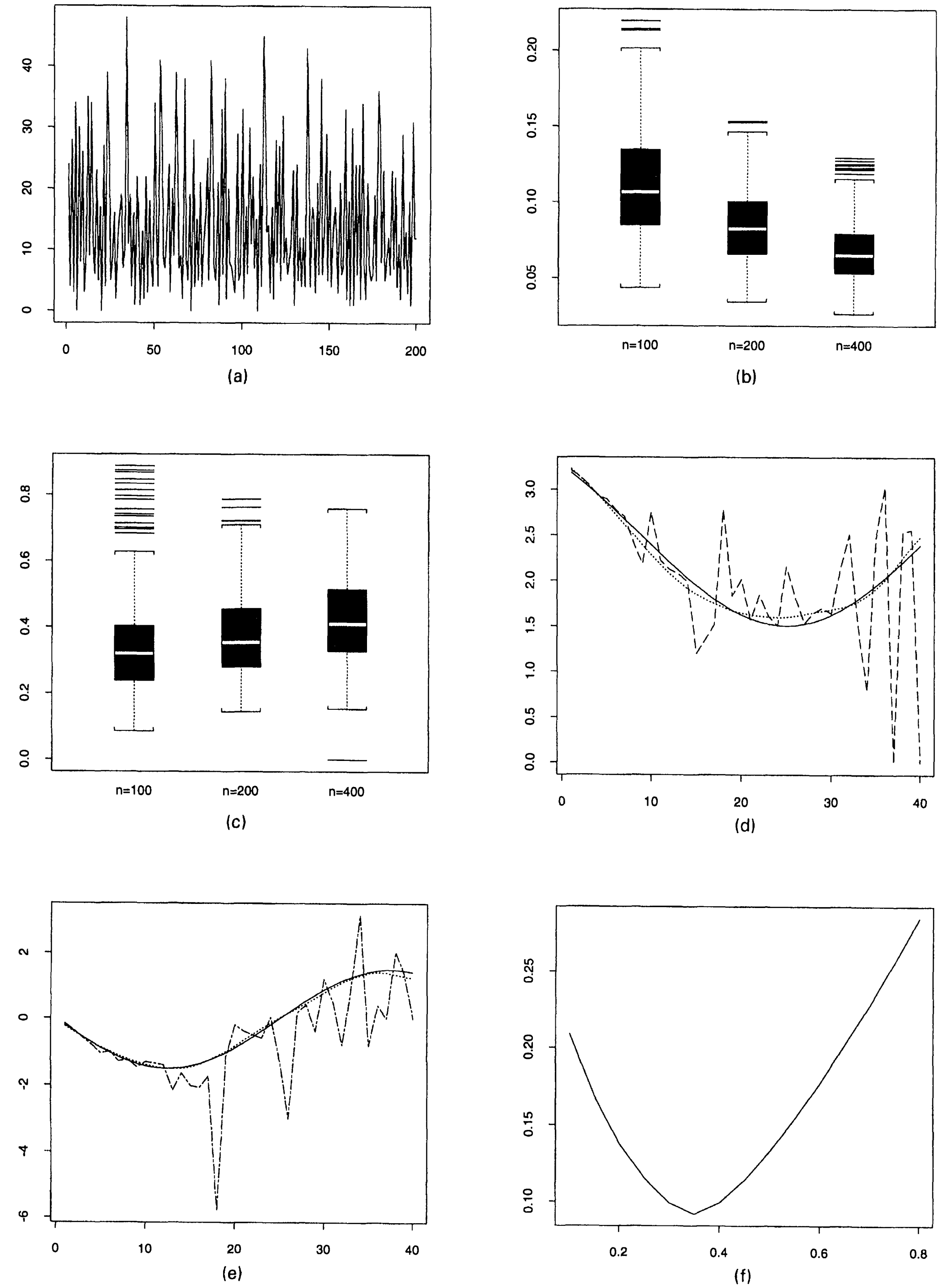


Fig. 1. Simulation results for example 1: (a) plot of a time series $\{Y_t\}$ of length 200; (b) box plot of 400 values of $\mathcal{E}$ defined in equation (4.1); (c) box plot of 400 values of $\mathcal{E}_r$ defined in equation (4.2); (d) plot of true β_{1i} (———), $\hat{\beta}_{1i}$ (·········) and $\tilde{\beta}_{1i}$ (— — —) against i; (e) plot of true β_{2i} (———), $\hat{\beta}_{2i}$ (·········) and $\tilde{\beta}_{2i}$ (· — · —) against i; (f) plot of $M(h)$ against bandwidth h

which minimizes $M(h)$ defined in equation (2.4). Fig. 1(b) presents box plots of the mean absolute deviation errors (ADEs)

$$\mathcal{E} = \sum_i \hat{\pi}_i \{|\hat{\beta}_{1i} - \beta_{1i}| + |\hat{\beta}_{2i} - \beta_{2i}|\}, \tag{4.1}$$

where $\hat{\pi}_i$ is the relative frequency estimator for the marginal probability $P(Y_t = i)$. The measure $\mathcal{E}$ decreases as the sample size increases. Fig. 1(c) depicts box plots of the mean *relative* ADEs

$$\mathcal{E}_{\mathrm{r}} = \sum_i \hat{\pi}_i \frac{|\hat{\beta}_{1i} - \beta_{1i}| + |\hat{\beta}_{2i} - \beta_{2i}|}{|\tilde{\beta}_{1i} - \beta_{1i}| + |\tilde{\beta}_{2i} - \beta_{2i}|}. \tag{4.2}$$

$\mathcal{E}_{\mathrm{r}} < 1$ implies that the smoothed estimation outperforms its parametric counterpart. Fig. 1(c) shows that the improvement of using the smoothed method is substantial, although the difference decreases as n increases. This indicates that the smoothed method is more relevant for small sample sizes, although for sample sizes as large as 400 it is still significantly better than the unsmoothed method in this example. Figs 1(d) and 1(e) plot the typical example of estimated β_{1i} and β_{2i} against i, together with their true values. A typical example is selected such that the corresponding $\mathcal{E}$ is equal to the median in the 400 replicated simulations. Fig. 1(f) plots $M(h)$ defined in equation (2.4) against bandwidth h, which indicates that the optimal bandwidth is 0.35 for this typical example.

4.2. Example 2

We consider a Markov chain time series $\{Y_t\}$ generated as follows. Given $\{Y_s, s \leqslant t\}$, the conditional distribution of Y_{t+1} is binomial$\{m, p(Y_t)\}$, where

$$\operatorname{logit}\{p(i)\} = i/m - (i/m)^2 \equiv \beta_i.$$

Figs 2(a), 2(b) and 2(c) plot segments of time series $\{Y_t\}$ of length 200 for $m = 5$, $m = 10$ and $m = 20$ respectively. For each of the sample sizes $n = 100, 200, 400$ and $m = 5, 10, 20$, we repeat the simulation 400 times. For each realization, we compute the ANW estimator $\{\hat{p}_{ij}\}$ defined in equation (3.5) with the GCV bandwidth (see Section 3.2) and the relative frequency estimator $\{\check{p}_{ij}\}$ given in equation (3.2). Fig. 2(d) presents box plots of mean ADEs

$$\mathcal{E} = \sum_i \hat{\pi}_i \sum_j |\hat{p}_{ij} - p_{ij}| \tag{4.3}$$

for $m = 5$ (the three panels on the left), $m = 10$ (the three panels in the middle) and $m = 20$ (the three panels on the right), where $\hat{\pi}_i$ is the relative frequency estimator for the marginal probability $P(Y_t = i)$. It is clear that $\mathcal{E}$ decreases as the sample size n increases. Fig. 2(e) displays the box plots of the mean *relative* ADEs

$$\mathcal{E}_{\mathrm{r}} = \sum_i \hat{\pi}_i \frac{\sum_j |\hat{p}_{ij} - p_{ij}|}{\sum_j |\check{p}_{ij} - p_{ij}|}. \tag{4.4}$$

It holds always that $\mathcal{E}_{\mathrm{r}} < 1$ in Fig. 2(e). This indicates that the nonparametric estimator always performs better than the relative frequency estimator in this example, although the difference between the two methods decreases as the sample size n increases. Note that, as m increases, $\mathcal{E}_{\mathrm{r}}$ decreases. This illustrates that, the more sparse the distribution is, the more relevant the smoothing is. Figs 2(g)–2(l) plot the typical example of estimated $\hat{p}_{ij}$ (dotted line)

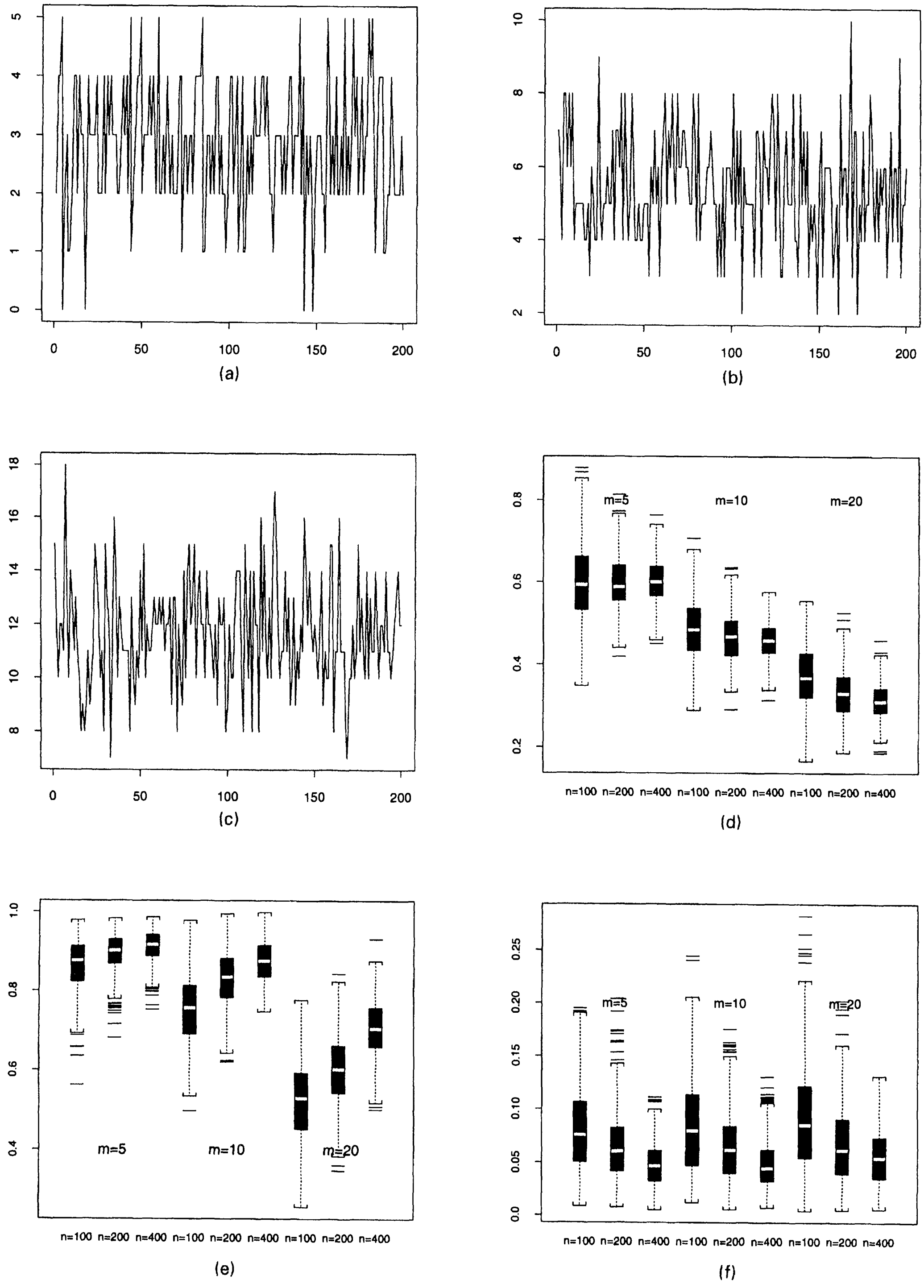
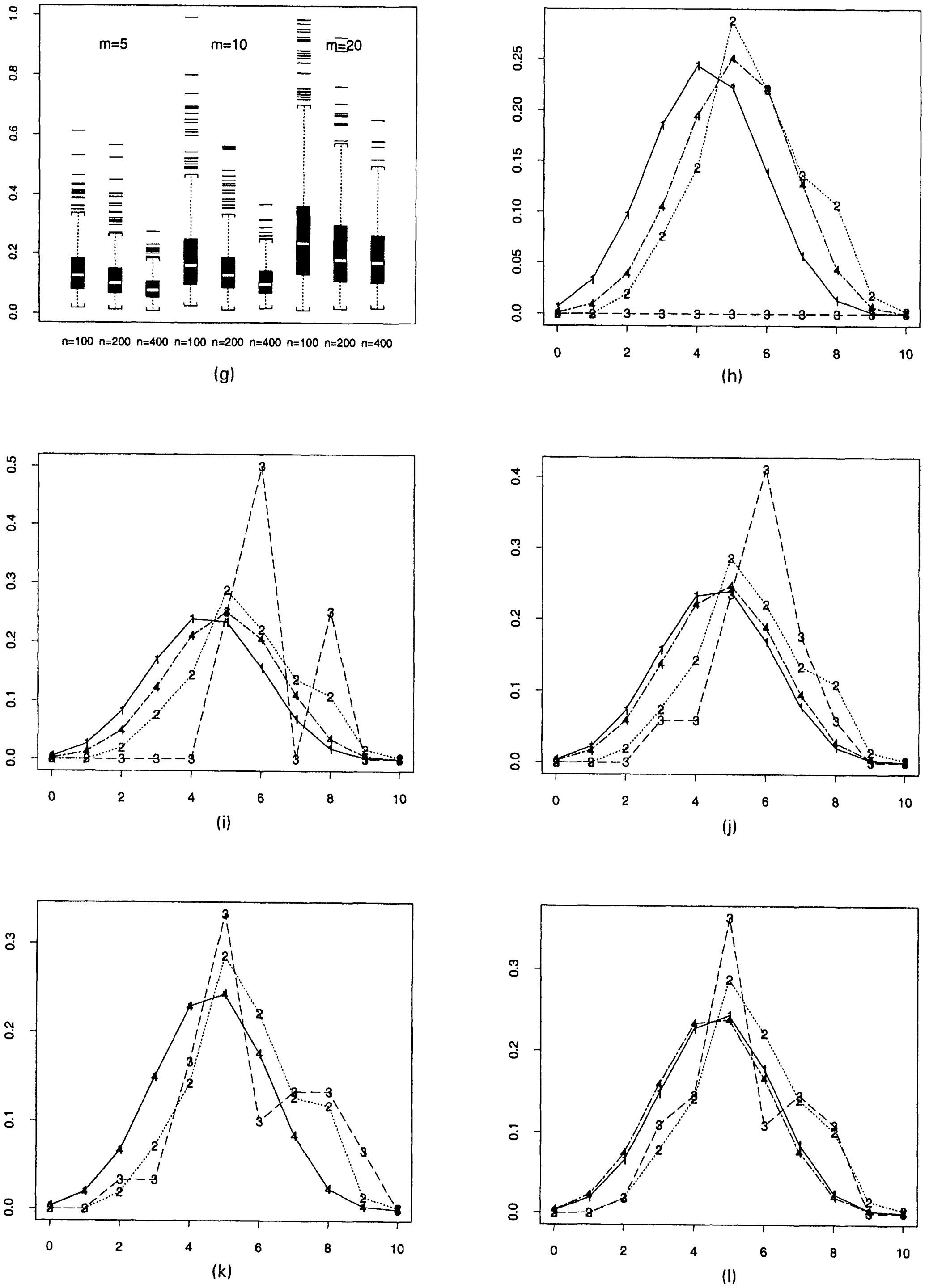


Fig. 2. Simulation results for example 2: (a)–(c) plot of a time series $\{Y_t\}$ of length 200 for $m = 5, 10, 20$; (d) box plot of 400 values of $\mathcal{E}$ defined in equation (4.3); (e) box plot of 400 values of $\mathcal{E}_r$ defined in equation (4.4); (f) box plot of 400 values of $\mathcal{E}^*$ defined in equation (4.5); (g) box plot of 400 values of $\mathcal{E}_r^*$ defined in equation (4.6);



(g)

(h)

(i)

(j)

(k)

(l)

Fig. 2. (*continued*) (h)–(l) plots of true p_{ij} (———), $\hat{p}_{ij}$ (·········), $\check{p}_{ij}$ (— — —) and p^*_{ij} (· — · —) against j for $n = 200$, $m = 10$ and $i = 1$–5 respectively and $\check{p}_{ij}$ (broken line) against j for $n = 200$ and $m = 10$, together with the true value p_{ij} (full line), for five cases: from $i = 1$ to $i = 5$. They show clearly that the ANW estimates are more accurate than the relative frequency estimates. Typical examples are selected such that the corresponding $\mathcal{E}$s are equal to the medians in the 400 replicated simulations.

To illustrate the superior performance of the smoothed parametric estimation over the purely nonparametric estimation, we apply the smoothed maximum likelihood method (see equation (2.3)) to obtain the smoothed parametric estimator $\hat{\beta}_i$, and then compare directly the derived estimator

$$p_{ij}^* = \binom{m}{j}\left\{\frac{\exp(\hat{\beta}_i)}{1+\exp(\hat{\beta}_i)}\right\}^j\left\{1-\frac{\exp(\hat{\beta}_i)}{1+\exp(\hat{\beta}_i)}\right\}^{m-j}$$

with the nonparametric estimator $\hat{p}_{ij}$ obtained above. Fig. 2(f) presents the box plots of values of

$$\mathcal{E}^* = \sum_i \hat{\pi}_i \sum_j |p_{ij}^* - p_{ij}| \tag{4.5}$$

in the 400 replications. A direct comparison with Fig. 2(d) indicates that overall p_{ij}^* is much more accurate than $\hat{p}_{ij}$. Furthermore, Fig. 2(g) displays box plots of the mean relative ADEs for p_{ij}^* over $\hat{p}_{ij}$

$$\mathcal{E}_{\mathrm{r}}^* = \sum_i \hat{\pi}_i \frac{\sum_j |p_{ij}^* - p_{ij}|}{\sum_j |\hat{p}_{ij} - p_{ij}|}, \tag{4.6}$$

which indicates a significant gain from using the parametric model, as $\mathcal{E}_{\mathrm{r}}^*$ is always smaller than 1 and in fact is smaller than 0.4 in most cases. The finding here reinforces the theoretical results in remark 9.

4.3. Example 3

Finally we apply the methods proposed to Hong Kong environmental data. The data were collected daily in Hong Kong from January 1st, 1994, to December 31st, 1995 (courtesy of Professor T. S. Lau). Of interest is to examine the relationship between the total number of daily hospital admissions (Y_t) for circulatory and respiratory problems and the levels of pollutants. Fig. 3(a) displays the number of daily hospital admissions. The covariates are taken as the levels of the pollutants sulphur dioxide X_{1t} (in micrograms per cubic metre), nitrogen dioxide X_{2t} (in micrograms per cubic metre) and dust X_{3t} (in micrograms per cubic metre). The correlation coefficient between X_{2t} and X_{3t} is 0.782, which is quite high.

Fan and Zhang (1999) used a varying-coefficient model

$$Y_t = a_1(t) + a_2(t)X_{1t} + a_3(t)X_{2t} + a_4(t)X_{3t} + \epsilon_t \tag{4.7}$$

to fit the daily data. Cai *et al.* (2000) considered a Poisson regression model for the weekly data with mean $\lambda(t, \mathbf{X}_t)$ given by

$$\log\{\lambda(t, \mathbf{X}_t)\} = a_1(t) + a_2(t)X_{1t} + a_3(t)X_{2t} + a_4(t)X_{3t}. \tag{4.8}$$

Although the models fitted are interesting, both approaches ignore the autodependence in the data; see Fig. 3(b). Cai *et al.* (2000) argued that the autocorrelation of the response variable is not strong for the weekly data. They also pointed out that X_{3t} in the above model is not significant according to a goodness-of-fit test.

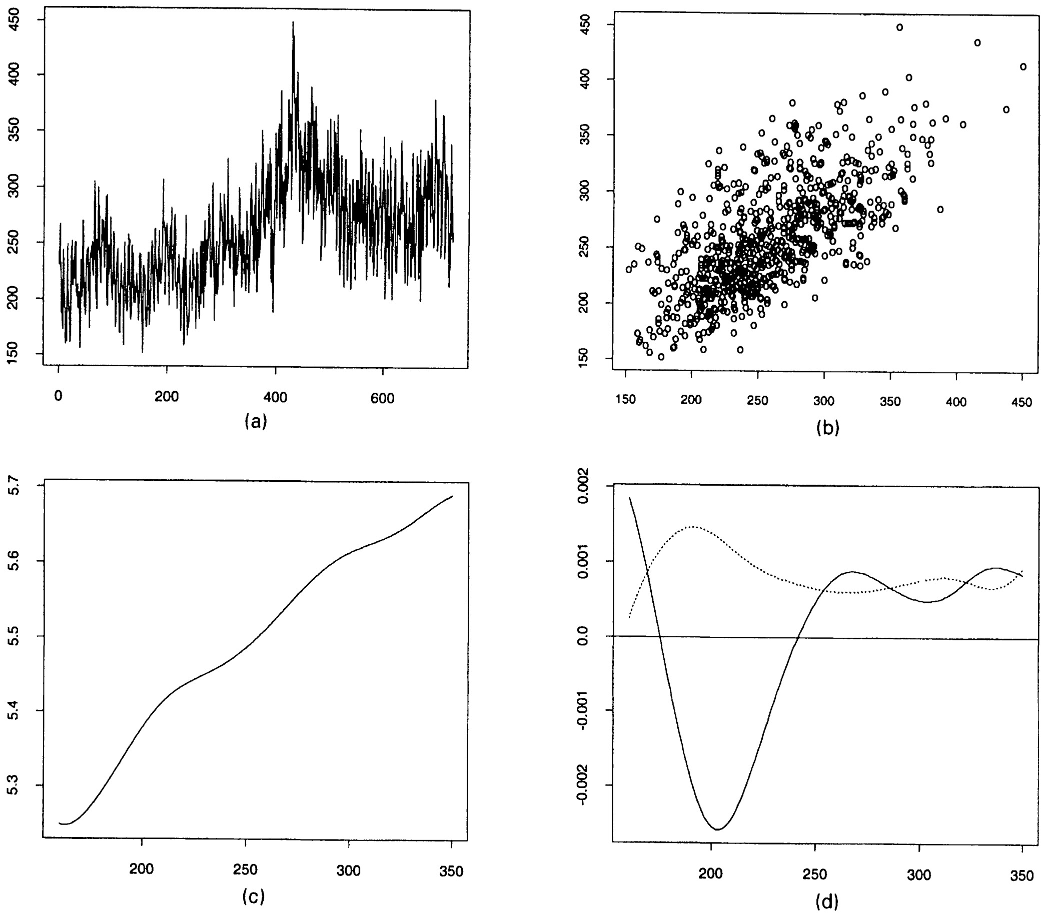


Fig. 3. (a) Time series plot of daily hospital admissions; (b) scatterplot of Y_t *versus* Y_{t-1}; (c) smoothed parametric estimation of β_1; (d) smoothed estimation of β_2 (———) and β_4 (··········)

Following the lead of Cai *et al.* (2000), we model the daily hospital admissions with Poisson distributions. However, instead of letting the parameter in the link function vary with respect to time, we assume that it is a function of the number of patients in the immediate past. By modelling data in such a way, we can incorporate the dependence in the time series in the model. We assume that the number of daily admissions Y_t follows a Poisson distribution with mean $\lambda_i(\mathbf{X}_t, \boldsymbol{\beta}_i)$, conditionally on its lagged values $Y_{t-1} = i$, Y_{t-2}, $Y_{t-3}, \ldots$, and the levels of pollutants, where $\lambda_i(\cdot, \cdot)$ is given by

$$\log\{\lambda_i(\mathbf{X}_t, \boldsymbol{\beta}_i)\} = \beta_{1i} + \beta_{2i} X_{1t} + \beta_{3i} X_{2t} + \beta_{4i} X_{3t}. \qquad (4.9)$$

This model differs from that of Cai *et al.* (2000) because the β-parameters now vary with respect to the immediately lagged value Y_{t-1} rather than time t. In this model the dependence of Y_t on Y_{t-1} has been also reflected indirectly by its association with pollutants X_{jt} for $j = 1$, 2, 3. Therefore, there is a genuine need to delete the insignificant variables in equation (4.9). For this, we propose an *ad hoc* procedure based on the *local* Akaike information criterion AIC as follows. In general, AIC is defined as

Table 1. AIC values for the eight candidate models

Model with covariate(s)	No X_js	X_1	X_2	X_3	(X_1, X_2)	(X_1, X_3)	(X_2, X_3)	(X_1, X_2, X_3)
AIC − min(AIC)	6.891	6.740	1.134	0.382	0.800	0.000	0.852	1.559

$$-2(\text{maximized log-likelihood}) + 2(\text{number of estimated parameters}).$$

See Akaike (1973). From expression (2.2), we may define the local AIC at $Y_{t-1} = i$ for this example as

$$\mathrm{AIC}_i(d) = 2\sum_{t=2}^{n} [\lambda_i(\mathbf{X}_t, \hat{\boldsymbol{\beta}}_i) - Y_t \log\{\lambda_i(\mathbf{X}_t, \hat{\boldsymbol{\beta}}_i)\}]\, K_{n,h}(Y_{t-1} - i) + 2\, d,$$

where d is the number of non-zero components of $\boldsymbol{\beta}_i$. By minimizing $\mathrm{AIC}_i(d)$ over d, we derive an 'optimum' model for the conditional distribution of Y_t given $Y_{t-1} = i$. However, we are interested in the global form of mean in this example. We simply choose the model which minimizes the average local AIC defined as

$$\mathrm{AIC}(d) = \sum_i \hat{\pi}_i \,\mathrm{AIC}_i(d), \tag{4.10}$$

where $\hat{\pi}_i$ is the relative frequency estimate for $P(Y_t = i)$. By taking δ_n to be 0.05 and using the bandwidth selector described in Section 2.2, we computed the AIC values defined in equation (4.10) for all eight possible models (with no interactions), which are reported in Table 1. This leads to the selected model

$$\log\{\lambda_i(\mathbf{X}_t, \boldsymbol{\beta})\} = \beta_{1i} + \beta_{2i} X_{1t} + \beta_{4i} X_{3t}. \tag{4.11}$$

The corresponding optimal bandwidth is 1.10. Fig. 3(c) plots the estimated intercept $\hat{\beta}_{1i}$ against i (the value of Y_{t-1}) and Fig. 3(d) plots the estimated coefficients $\hat{\beta}_{2i}$ and $\hat{\beta}_{4i}$ against i. The fitted model for λ_i is dominated by the intercept $\hat{\beta}_{1i}$ which increases monotonically as i increases. This indicates clearly that the (conditional) distribution of Y_t depends on Y_{t-1}. Further, Y_t tends to be large when Y_{t-1} is large. This reflects the autodependence observed in Fig. 3(b).

Acknowledgements

Cai's work was supported, in part, by National Science Foundation grant DMS 0072400 and funds provided by the University of North Carolina at Charlotte. Yao and Zhang's work was partially supported by Biotechnology and Biological Sciences Research Council–Engineering and Physical Sciences Research Council grant 96/MM109785. We thank Dr Jeremy Penzer, Mrs Jane Galbraith and two reviewers for helpful comments.

References

Aerts, M., Augustyns, I. and Janssen, P. (1997) Smoothing sparse multinomial data using local polynomial fitting. *J. Nonparam. Statist.*, **8**, 127–147.

Aitchison, J. and Aitken, C. G. G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and P. Csàki), pp. 267–281. Budapest: Akadémiai Kiadó.

Bowman, A. W. (1980) A note on consistency of the kernel method for the analysis of categorical data. *Biometrika*, **67**, 682–684.

Cai, Z., Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models. *J. Am. Statist. Ass.*, **95**, 888–902.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

Dong, J. and Simonoff, J. S. (1994) The construction and properties of boundary kernels for sparse multinomials. *J. Comput. Graph. Statist.*, **3**, 57–66.

Faddy, M. J. and Jones, M. C. (1998) Semiparametric smoothing for discrete data. *Biometrika*, **85**, 131–138.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, **21**, 196–216.

Fan, J., Farmen, M. and Gijbels, I. (1998) Local maximum likelihood estimation and inference. *J. R. Statist. Soc.* B, **60**, 591–608.

Fan, J. and Yao, Q. (2002) *Nonlinear Time Series: Nonparametric and Parametric Methods.* New York: Springer. To be published.

Fan, J., Yao, Q. and Tong, H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189–206.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying-coefficient models. *Ann. Statist.*, **27**, 1491–1518.

Hall, P. (1981) On nonparametric multivariate binary discrimination. *Biometrika*, **68**, 287–294.

Hall, P. and Presnell, B. (1999) Intentionally biased bootstrap methods. *J. R. Statist. Soc.* B, **61**, 143–158.

Hall, P. and Titterington, D. M. (1987) On smoothing sparse multinomial data. *Aust. J. Statist.*, **29**, 19–37.

Hall, P., Wolff, R. C. L. and Yao, Q. (1999) Methods for estimating a conditional distribution function. *J. Am. Statist. Ass.*, **94**, 154–163.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B, **60**, 271–293.

Lehmann, E. L. and Casella, G. (1998) *Theory of Point Estimation.* New York: Springer.

Polonik, W. and Yao, Q. (2000) Conditional minimum volume predictive regions for stochastic processes. *J. Am. Statist. Ass.*, **95**, 509–519.

Simonoff, J. S. (1985) An improved goodness-of-fit statistic for sparse multinomials. *J. Am. Statist. Ass.*, **80**, 671–677.

———(1995) Smoothing categorical data. *J. Statist. Planng Inf.*, **47**, 41–69.

———(1996) *Smoothing Methods in Statistics.* New York: Springer.

———(1998) Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. *Int. Statist. Rev.*, **66**, 137–156.

Tibshirani, R. and Hastie, T. (1987) Local likelihood estimation. *J. Am. Statist. Ass.*, **82**, 559–567.

Titterington, D. M. (1980) A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.

Wahba, G. (1977) A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (ed. P. R. Krisnaiah), pp. 507–523. Amsterdam: North-Holland.

Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models, and the Gaussian-Newton method. *Biometrika*, **61**, 439–447.