Bernoulli **10**(5), 2004, 919–938

Nonparametric methods of inference for finite-state, inhomogeneous Markov processes

PETER HALL[1] and EFSTATHIA BURA[2]

[1]*Centre for Mathematics and its Applications Australian National University, Canberra, ACT 0200, Australia E-mail: peter.hall@anu.edu.au*

[2]*Department of Statistics, George Washington University, Funger Hall 2201 G Street NW, Washington DC 20052 USA E-mail ebura@gwu.edu*

In some inferential problems involving Markov process data, the inhomogeneity of the process is of central interest. One example is of a binary time series of data on the presence or absence of a species at a particular site over time. Here the two states correspond to 'presence' or 'absence,' respectively, of the species, and the main topic of interest is temporal variation in the process In principle this variation can be modelled parametrically, but in the absence of information about the physical mechanism causing species numbers to fluctuate, it is usually very difficult to suggest a plausible model that explains the data, at least until a more adaptive analysis is conducted These issues argue in favour of nonparametric methods for estimating probabilities of transition, and for estimating probabilities of the process being in a given state at a given time. Such techniques, which in practice might be a prelude to parametric modelling, will be introduced and explored, under assumptions motivated by characteristics of the data set mentioned above. These assumptions will be shown to lead to consistent estimation of probabilities, and so to imply that nonparametric methodology gives accurate information about properties of the process.

Keywords bandwidth; binary time series; kernel methods; local linear methods; nonparametric regression; state probability; transition probability

1. Introduction

Some binary time series, where only the 'presence' or 'absence' of a feature is recorded, can be represented as two-state Markov processes. However, in some problems of practical interest the process is not temporally homogeneous, and important properties of the process, such as the probability that the feature is present (or absent) at a given point in time, vary with that point.

Consider, for example, the binary time series that represents the presence or absence of a species at a given epoch, and at a given site Data are usually obtained by carbon-dating fossil records Let state $i = 0$ denote absence, and $i = 1$ correspond to presence The 'state probability', $p_i(t)$, that the process is in state i at time t, is of direct scientific interest.

At a higher level, if the transition probability matrix $(p_{ij}(t_2|t_1))$, governing transitions from state i at time t_1 to state j at time t_2, were known, then the Markov process could be 'run,' starting from either of the two states, to give Monte Carlo approximations to the probability that the process was in a given state at a given time. This would be beneficial in making inference from fossil records. However, in many instances the absence of information about the process that led to temporal variations in species abundance makes it difficult to develop parametric models, without first constructing an adaptive, nonparametric estimate of transition probabilities. With these motivations, we shall suggest nonparametric methods for inference about time-inhomogeneous Markov processes from discrete data.

The Markov process cannot be simulated directly from estimated transition probabilities, since they do not satisfy the Chapman–Kolmogorov equations (see, for example, Cox and Miller 1965, p. 179). To overcome this problem, having used discrete data to estimate transition probabilities in the continuum, one might discretize the process, using a grid that is not so fine that the new process is too strongly influenced by errors in estimating transition probabilities, or so coarse that detailed information about the time-inhomogeneity of the original process is lost unnecessarily.

In Section 2 we shall suggest methods for estimating state probabilities and transition probabilities for inhomogeneous, finite-state Markov processes that vary continuously but are observed only discretely. The context in which we are developing these techniques will be illustrated, in Section 3, by a practical example.

The main features of that example are that the number of transitions over the entire period, and the number of observations made between transitions, tend to be moderately large; and the probability distribution for transitions varies smoothly with time. These properties will be used to motivate, in Section 4.1, specific models that might generate data such as those discussed in Section 3, leading, in Section 4.2, to particular regularity conditions under which theory for estimators, given in Section 2, can be developed. Theoretical results, stated in Section 4.2, show that our nonparametric methods give uniformly accurate estimators of state probabilities and transition probabilities.

Our work has connections to nonparametric methods for time series and other dependent processes, where there is an especially large literature. It includes contributions by Robinson (1983; 1997), Roussas (1990), Hart (1991), Roussas *et al.* (1992), Truong and Stone (1992), Tran (1993), Yakowitz (1993), Greenwood and Wefelmeyer (1994), Wu and Chu (1994), Masry (1996; 2001), Robinson and Hidalgo (1997), Utikal (1997), Tran *et al.* (1996), Karlsen and Tjøstheim (2001) and Masry and Mielniczuk (2001).

2. Nonparametric methods for inference

2.1. Model, and methods for first-order inference

We record data $X(T_i)$, for $1 \leqslant i \leqslant n$, being the result of observing a continuous-time, discrete-valued stochastic process $X(t)$, taking only values in the sequence $0, 1, \ldots, q-1$, at the discrete, perhaps unequally spaced times $T_1 < \ldots < T_n$ in a given compact interval $\mathcal{I}$. The time points T_i are assumed to be stochastically independent of X. The process $X(t)$

is piecewise constant, and may be supposed to be left- or right-continuous in order that its sample paths might be uniquely defined

In some instances, although not in the case of our data, the process X may be observed in the continuum. Here $n = \infty$ and, in the definitions of estimators in Section 2, series should be replaced by integrals. Once this has been done, all the estimators suggested in this section remain valid. Theoretical properties in the case of continuous data will be briefly discussed in Section 4.2

We shall suppose $X(t)$ is Markovian. That is, if i and j each take a value in $0, 1, \ldots, q-1$, and if $t_1 < t_2$ and $t_1 > s_1 > s_2 > \ldots$, then

$$p_{ij}(t_2|t_1) \equiv P\{X(t_2) = j|X(t_1) = i\}$$
$$= P\{X(t_2) = j|X(t_1) = i;\, X(s_k) = i(s_k) \text{ for } k \geqslant 1\},$$

where $0 \leqslant i(s_1), i(s_2), \quad \leqslant q-1$

Define $p_i(t) = P\{X(t) = i\}$. Local linear estimators of $p_i(t)$, for $0 \leqslant i \leqslant q-1$, which take account of the fact that $\sum_i p_i(t) = 1$, are obtained by minimizing

$$S_1(a_0, b_0, \ldots, a_{q-2}, b_{q-2}) = \sum_{i=0}^{q-2} \sum_{k=1}^{n} [I_{ki} - \{a_i + b_i(t - T_k)\}]^2\, K_k(t)$$
$$+ \sum_{k=1}^{n} \left(I_{k,q-1} - \left[1 - \sum_{i=0}^{q-2} \{a_i + b_i(t - T_k)\} \right] \right)^2 K_k(t), \qquad (2.1)$$

where $K_k(t) = K\{(t - T_k)/h\}$, $I_{ki} = I\{X(T_k) = i\}$, K denotes a kernel function and h is a bandwidth. See Fan and Gijbels (1996, pp 19ff.) for discussion of local linear regression. If $\hat{v} = (\hat{a}_0, \hat{b}_0, \ldots, \hat{a}_{q-2}, \hat{b}_{q-2})$ gives the minimum of S_1 then $\hat{p}_i(t) = \hat{a}_i$, for $0 \leqslant i \leqslant q-2$, and $\hat{p}_{q-1} = 1 - (\hat{a}_0 + \quad + \hat{a}_{q-2})$ are our estimators of $p_i(t)$. The $2(q-1)$-vector $\hat{v}$ is the solution of the $2(q-1)$ equations

$$\sum_{j=0}^{q-2} \{(1 + \delta_{ij}) A_0(t) a_j + (1 + \delta_{ij}) A_1(t) h b_j\} = B_{i0}(t) + C_{q-1\,0}(t),$$
$$\sum_{j=0}^{q-2} \{(1 + \delta_{ij}) A_1(t) a_j + (1 + \delta_{ij}) A_2(t) h b_j\} = B_{i1}(t) + C_{q-1\,1}(t), \qquad (2.2)$$

where $0 \leqslant i \leqslant q-2$, δ_{ij} denotes the Kronecker delta,

$$A_j(t) = \frac{1}{nh} \sum_{k=1}^{n} K_k(t)\{(t - T_k)/h\}^j, \qquad B_{ij}(t) = \frac{1}{nh} \sum_{k=1}^{n} K_k(t)\{(t - T_k)/h\}^j\, I_{ki}$$

and $C_{ij}(t) = A_j(t) - B_{ij}(t)$ The parameters b_j may be eliminated from equations (2.2), giving just $q-1$ equations in the $q-1$ unknowns a_j:

$$\sum_{j=0}^{q-2}(1+\delta_{ij})a_j = \frac{A_2(t)\{B_{i0}(t)+C_{q-1,0}(t)\} - A_1(t)\{B_{i1}(t)+C_{q-1,1}(t)\}}{A_2(t)A_0(t)-A_1(t)^2}. \tag{2.3}$$

In the case $q=2$, equations (2.3) reduce to their counterparts in the problem of estimating a regression with Bernoulli response variables, that is, where $I\{X(T_k) = i\} = p_i(T_k) +$ error. In particular,

$$\hat{p}_i(t) = \frac{A_2(t)B_{i0}(t) - A_1(t)B_{i1}(t)}{A_2(t)A_0(t)-A_1(t)^2}. \tag{2.4}$$

A ridge parameter, or another approach to rendering the estimator robust against instances where the denominator at (2.4) is small, may be incorporated.

In the special case $q=2$, if $\hat{p}_i$ is given by (2.4) then $\hat{p}_0+\hat{p}_1=1$. More generally, however, if we define $\hat{p}_i$ by (2.4) then it is not necessarily true that $\sum_i \hat{p}_i = 1$. When $q=2$, positivity and summation to unity are ensured by using instead the estimator $\tilde{p}_i = \min\{\max(\hat{p}_i, 0), 1\}$.

Local log-linear, or local logit, methods can be employed instead of local linear ones. In these cases, each term $a_i + b_i(t-T_k)$ in (2.1) is replaced by $\exp\{a_i + b_i(t-T_k)\}$ or $[1+\exp\{a_i+b_i(t-T_k)\}]^{-1}$, respectively. This guarantees that $\hat{p}_0, \ldots, \hat{p}_{q-2}$ are positive and (in the local logit case) that $\sum_{i\leqslant q-2}\hat{p}_i \leqslant 1$. On the other hand, advantages of the log-linear approach, versus local logit, include the fact that it uses standard, widely available software, has numerical properties that are well understood, is widely accepted as a smoothing technique, and has theory that is a little more transparent.

2.2. Second-order inference

Assume we have estimators $\hat{p}_i(t)$ of the state probabilities $p_i(t)$. We shall construct estimators $\hat{p}_{ij}(t_1, t_2)$ of the dual-state probabilities,

$$p_{ij}(t_1, t_2) = P\{X(t_1) = i, X(t_2) = j\},$$

and take

$$\hat{p}_{ij}(t_2|t_1) = \hat{p}_{ij}(t_1, t_2)/\hat{p}_i(t_1) \tag{2.5}$$

to be our estimator of $p_{ij}(t_2|t_1)$, incorporating a ridge parameter into the denominator if desired.

Analogously to (2.1), we obtain $\hat{p}_{ij}(t_1, t_2)$ by minimizing

$$
\begin{aligned}
&S_2(a_0, b_{01}, b_{02}, \ldots, a_{q-2}, b_{1,q-2}, b_{2,q-2})\\
&\quad = \sum_{j=0}^{q-2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \big[I_{ki} I_{\ell j} - \{a_j + b_{j1}(t_1 - T_k) + b_{j2}(t_2 - T_\ell)\}\big]^2 K_k(t_1)K_\ell(t_2)\\
&\qquad + \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left(I_{ki} I_{\ell,q-1} - \left[\hat{p}_i(t_1) - \sum_{j=0}^{q-2} \{a_j + b_{j1}(t_1 - T_k) + b_{j2}(t_2 - T_\ell)\} \right] \right)^2\\
&\qquad \times K_k(t_1)K_\ell(t_2) \qquad (2.6)
\end{aligned}
$$

If the minimum of S_2 occurs at $\hat{a}_0, \hat{b}_{01}, \hat{b}_{02}, \ldots, \hat{a}_{q-2}, \hat{b}_{q-2,1}, \hat{b}_{q-2,2}$ then our estimator of $p_{ij}(t_1, t_2)$ is $\hat{p}_{ij}(t_1, t_2) = \hat{a}_j$ for $0 \leqslant j \leqslant q-2$, and $\hat{p}_{i,q-1} = 1 - (\hat{a}_0 + \cdots + \hat{a}_{q-2})$ for $j = q-1$. Our method acknowledges the fact that $\sum_j p_{ij} = p_i$, and as a result our estimators satisfy $\sum_j \hat{p}_{ij}(t_1, t_2) = \hat{p}_i(t_1)$ for all t_1, t_2.

In the two-state case, equations (2.6) involve just three variables, $(a, b_1, b_2) = v^{\mathrm{T}}$ say, and have solution $A\hat{v} = \alpha$, where $\hat{v} = (\hat{a}, \hat{b}_1, \hat{b}_2)^{\mathrm{T}}$, $A = (a_{rs})$ is a 3×3 matrix, $\alpha = (\alpha_r)$ is a column vector of length 3,

$$a_{rs} = 2\,\frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} u_r(k, \ell)u_s(k, \ell)K_k(t_1)K_\ell(t_2),$$

$$\alpha_r = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \{I_{ki}(2I_{\ell 0} - 1) + \hat{p}_i(t_1)\}\, u_r(k, \ell)K_k(t_1)K_\ell(t_2),$$

$u_1(k, \ell) = 1$, $u_2(k, \ell) = t_1 - T_k$ and $u_3(k, \ell) = t_2 - T_\ell$. To ensure positivity we take $\hat{p}_{i0}(t_1, t_2) = \min\{\max(\hat{a}, 0), \hat{p}_i(t_1)\}$ and $\hat{p}_{i1}(t_1, t_2) = \hat{p}_i(t_1) - \hat{p}_{i0}(t_1, t_2)$. Then we define $\hat{p}_{ij}(t_2|t_1)$ by (2.5).

These transition probability matrix estimators will not satisfy the Chapman–Kolmogorov equations. That is, it will not be true that

$$\hat{p}_{ij}(t_2|t_1) = \sum_{k=0}^{q-2} \hat{p}_{ik}(t_3|t_1)\hat{p}_{kj}(t_2|t_3) \qquad \text{for } t_1 < t_3 < t_2$$

Therefore, the estimated transition probabilities cannot be used to generate a Markov chain in the continuum, for example by direct simulation.

Nevertheless, if we restrict attention to a grid of time points, say $k\tau$ for integers k, which is not 'too fine'; if, for adjacent points on that grid, we interpret $\hat{p}_{ij}((k+1)\tau|k\tau)$ as a $q \times q$ matrix indexed by k; and if we simulate the Markov process as an inhomogeneous Markov chain, on the grid, by multiplying the matrices together to compute probabilities of transitions across multiple grid points, then simulation is straightforward.

The reason why the grid should not be too fine is that the relative accuracy in estimation of $p_{ij}(t_1, t_2)$, for $i \neq j$, deteriorates as $t_2 - t_1$ becomes small. (Remarks following Theorem 4.2 will discuss this point.) Thus, grid width has some of the properties of a smoothing parameter. In the context of an infill asymptotic approach to theory, the grid can become finer as the quantity of data increases, but should not decrease to zero for a finite quantity of data Theoretical aspects of choice of τ will be discussed at the end of Section 4 2.

2.3. Bootstrap assessment of variability

An important feature of the problem is that it does not readily admit a 'structural model', such as those commonly imposed in more familiar nonparametric regression problems of density estimation and regression. In particular, no unchanging, smooth function is being estimated. This makes it awkward to describe, in theoretical terms, the variability of our estimators

To overcome this difficulty we suggest an empirical approach to assessing variability, as follows. Divide the data set into blocks of consecutive observations, the blocks being chosen so that the Markov chain has approximately homogeneous behaviour within each block Block lengths may be unequal Assume the data follow a time-independent Markov chain within each block. Estimate the transition probability matrix for each block, and generate new data sets (i.e. new observations in the set $\{0, 1, \ldots, q-1\}$, at the same time points as before) by running the blockwise stationary chain. Realizations may start in the same state as the original data set, or at a randomly chosen state. For each realization obtained in this way, compute the estimators and then calculate their bootstrap variance estimates. The method can be viewed as a version of the block bootstrap, tailored to the case of inhomogeneous Markov processes.

This approach can also be used to assess the robustness of our method to aberrations in the data, by introducing systematic or stochastic errors to the bootstrap sequences of 0s and 1s generated as suggested above, and observing how this affects bootstrap estimates of variability.

3. Numerical example

A fossil record of a particular species consists of a sequence of observance or non-observance of 'fossil horizons', represented mathematically by a sequence of 1s and 0s, respectively. The stratigraphic positions (e.g. depths in sediment) at which the observations are made, are converted to time points (in the case of our data), or to time intervals, by radiometric analysis.

A typical assumption in previous analyses of such data is that the observation times are uniformly distributed over the time interval between the first and last occurrences of a find. This assumption could arguably be sustained if the 'sedimentation potential', or potential for the species to leave a fossil record, were constant over time. However, that supposition is difficult to justify, and in fact little information is usually available about how sedimentation potential might fluctuate with time. This, we suggest, is just one reason why nonparametric methods are attractive for analysing data such as these.

The data presented here represent a fossil record of the *M. lacrymosum* Neogene bryozoan species of the genus *Metrarabdotos*, a sea-dwelling invertebrate, collected in the Dominican Republic and discussed by Cheetham (1986, 1987) and Jackson and Cheetham (1990). Cheetham's interpretation of these and related data played a supportive role in Stephen Jay Gould's advancement of the theory of 'punctuated equilibrium'. Cheetham, Gould (1991) argued, 'presented the most elegant and persuasive of all cases of punctuated equilibrium in his studies of the cheilostome bryozoan *Metrarabdotos*'. Gould's theory argues that, rather than being a relatively slow, continuous process, evolution consists of long, inactive periods punctuated by sudden 'shocks' or transitions.

This is an example where sedimentation potential is possibly far from uniform. Due to the stratigraphy of the Dominican Republic, there is a virtual absence of fossil-bearing rock aged between 8 and 14 million years. On the other hand, the fossil record is relatively complete between 4 and 8 million years ago, and almost absent from 3.5 million years ago to the present day.

The data are presented in Table 1. The 83 time points are in millions of years, slightly perturbed where radiometric analysis alone failed to separate epochs. (Radiocarbon dating indicated the rank order of the T_i in all cases.) Ones and zeros indicate presence and absence, respectively, of the species.

Table 1. Fossil record of *M. Lacrymosum*. The quantity T_i was determined by radiometric analysis; values of T_i were slightly perturbed where radiocarbon dating failed to separate epochs. Values 0 or 1 of $X(T_i)$ represent absence or presence, respectively, of the species at time T_i.

T_i	−8.00	−7.99	−7.91	−7.9	−7.89	−7.86	−7.85	−7.84	−7.81	−7.79	−7.76
$X(T_i)$	0	0	1	0	0	1	1	0	1	0	0
T_i	−7.75	−7.74	−7.71	−7.7	−7.69	−7.65	−7.62	−7.61	−7.59	−7.58	−7.55
$X(T_i)$	0	0	0	0	0	0	0	0	0	0	0
T_i	−7.51	−7.49	−7.45	−7.41	−7.39	−7.35	−7.30	−7.25	−7.20	−7.16	−7.14
$X(T_i)$	1	0	1	1	0	0	1	0	1	1	1
T_i	−7.05	−7.00	−6.91	−6.89	−6.80	−6.65	−6.56	−6.54	−6.52	−6.51	−6.49
$X(T_i)$	0	1	1	1	1	0	1	1	1	1	1
T_i	−6.48	−6.46	−6.45	−6.44	−6.40	−6.35	−6.26	−6.24	−6.20	−5.96	−5.94
$X(T_i)$	1	1	1	1	1	1	1	0	0	1	0
T_i	−5.90	−5.75	−5.71	−5.69	−5.66	−5.64	−5.60	−5.56	−5.55	−5.54	−5.45
$X(T_i)$	1	0	0	0	1	0	1	1	0	0	0
T_i	−5.41	−5.39	−5.35	−5.30	−5.25	−5.15	−5.10	−5.00	−4.81	−4.79	−4.55
$X(T_i)$	1	0	1	1	1	0	0	1	1	0	1
T_i	−3.85	−3.40	−3.35	−3.30	−2.50	−1.85					
$X(T_i)$	1	1	0	0	1	0					

The unbroken line in Figure 1(a) is a graph of the fitted function $\hat{p}_1$, computed using $h = 0.8$. The character of the curve changes little as h varies, and in fact the estimator is surprisingly robust against choice of h. For a wide range of bandwidths, $\hat{p}_1(t)$ rises

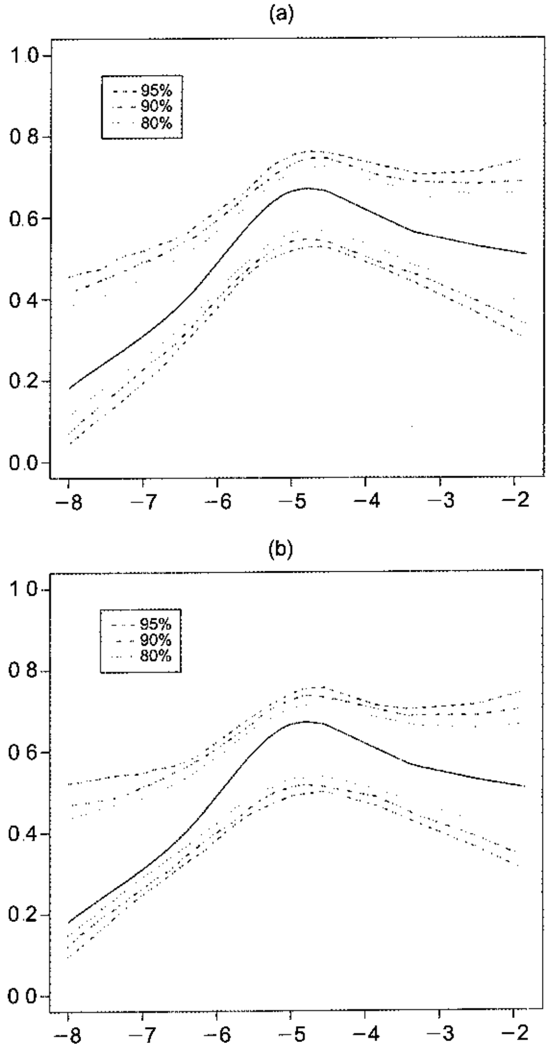

Figure 1. Graph of $\hat{p}_1(t)$, with bands. Time t, representing millions of years in the past, is shown on the horizontal axis, and $\hat{p}_1(t)$ is on the vertical axis. In each panel, the graph of $\hat{p}_1(t)$ is shown as an unbroken line. The dotted, dot-dashed and dashed lines in (a) give 80%, 90% and 95% pointwise confidence bands, respectively. The curves in (b) have the same interpretation, except that the bands were constructed after incorporating perturbation errors of 10%, independently and at random, for each state simulated at the 83 time points.

reasonably quickly from the inception of the data, approximately 8 million years ago, to a peak at approximately 5 million years ago, and then declines to a value of about 0.5, where it levels out.

Figure 2 shows graphs of estimates $\hat{p}_{ij}$ of the conditional probabilities p_{ij}, defined by

$$p_{ij}(t_2|t_1) = P\{X(t_2) = j|X(t_1) = i\},$$

for $(i, j) = (0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$, in the case where $t_2 > t_1 = -7.90$ In the calculation of $\hat{p}_{ij}(t_1, t_2)$ the bandwidth for computing $\hat{p}_i$ was kept at $h = h_1 = 0.8$, but the bandwidth for constructing the dual-state probability estimator $\hat{p}_{ij}(t_1, t_2)$ was taken to be $h = h_2 = 1.6$. Changing the bandwidths generally does not alter the character of the curve,

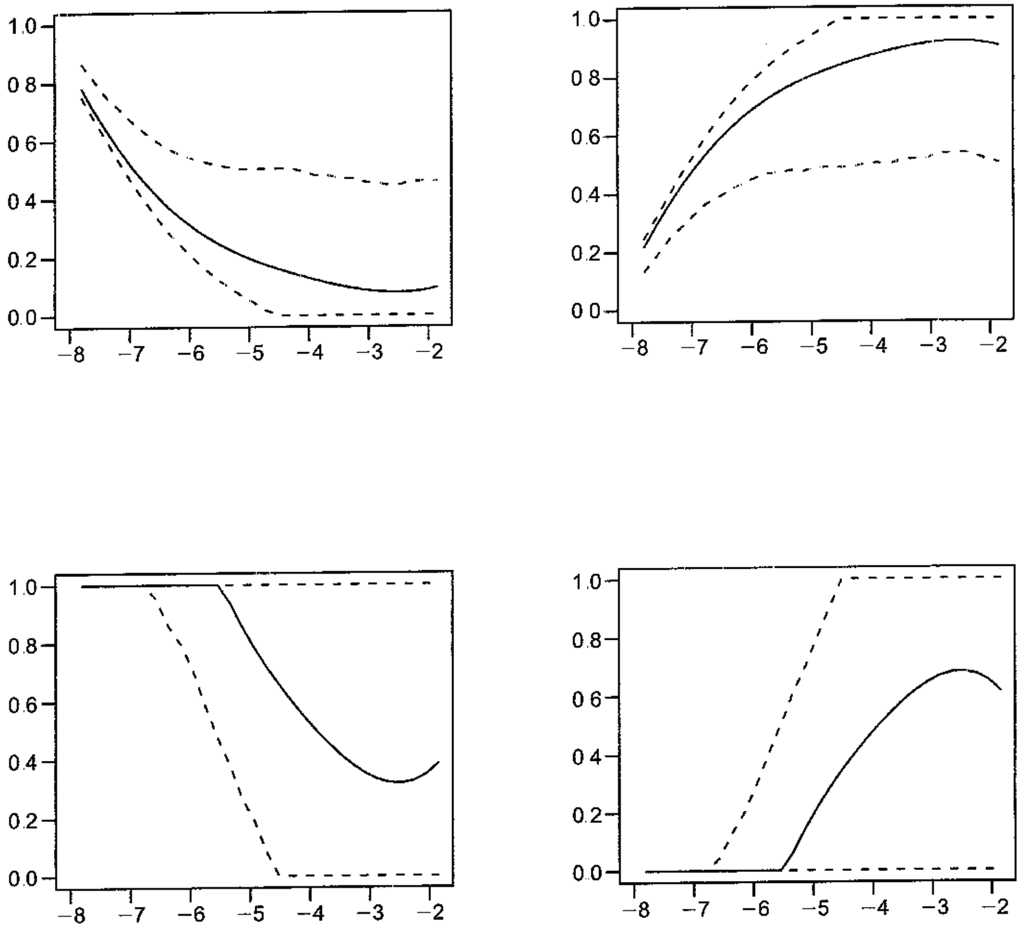

Figure 2. Graphs of $\hat{p}_{ij}(t)$. Panels across the first row represent $\hat{p}_{00}(t)$ and $\hat{p}_{01}(t)$, respectively, while panels across the second row represent $\hat{p}_{10}(t)$ and $\hat{p}_{11}(t)$, respectively. (Here, $\hat{p}_{ij}(t) = \hat{p}_{ij}(t| - 7.90)$.) The sum, across each row, of the graphed functions equals 1. Dashed lines show 70% pointwise confidence bands For each panel, time t is on the horizontal axis and $\hat{p}_{ij}(t)$ is on the vertical axis

except that decreasing the bandwidths makes the extrema more pronounced Throughout we used the standard normal kernel, although almost identical results are obtained for other choices. Potential kernels, and their relative advantages, are discussed by Simonoff (1996, p 44).

The non-differentiability of the estimates $P_{ij}(t_2|t_1)$ in Figure 2 is not scientifically reasonable In presenting our results to a different audience we would use a numerical taper, so that the curve estimator was brought smoothly to its extremum. This is of course straightforward to do However, in order to show the actual form that the 'raw' estimates take we have not used a taper in the figure.

To implement the method suggested in Section 2.3 we divided the sequence of time points into six blocks, the first five consisting of 14 points and the last of 13 points A visual inspection suggests that the chain is approximately time-homogeneous within each block Under the latter assumption we estimated one-step transition probability matrices within a block by taking numerical averages in the obvious way. The chain was then run repeatedly, starting from the observed initial state, and values of the bootstrap estimates of the p_i were computed using the method suggested in Section 2 Percentile bootstrap confidence bands were constructed in this way, at 80%, 90% and 95% levels, and are shown by broken lines in Figure 1(a)

The relatively small sample size available to us in this example means that stochastic variability is substantially greater in the bivariate problem of estimating conditional probabilities than it is in the univariate one of estimating p_1 and p_2. For this reason we did not attempt to simulate the Markov process from these data, and we show only 70% confidence bands in the plot in Figure 2. The general shape of the transition probability functions is reflected in these bands, although the actual values of the bands reveal a considerable degree of variability.

To assess the robustness of our method to errors in data collection, we subjected the simulated data to random errors as follows After 500 bootstrap samples were obtained from the blockwise Markov process, the value of each state (0 or 1) was perturbed with probability 0.10, independently of all other states. The estimate $\hat{p}_1(t)$ was then computed from each of the 500 perturbed data sets, and 80%, 90% and 95% percentile-bootstrap confidence regions were computed from the perturbed simulated data The results are shown in Figure 1(b). The field experts who provided the data to us have great confidence in the data, and argue that a 10% level of error is unduly high.

Gould's theory argues that a new species tends to arise relatively quickly, as a sharp punctuation of a pre-existing equilibrium, and then declines relatively slowly. The relatively fast increase of $\hat{p}_1(t)$, and its subsequent, gentler decrease, are consistent with, although of course do not prove, Gould's hypothesis. The confidence bands in Figure 1(a) reflect both the relatively steep increase and the subsequent slow decline The 'robustness bands' in Figure 1(b) also tend to preserve this pattern, although the fact that they reflect a higher degree of noise means that they do it less well. In particular, the bands become wider at either end, and this effect moderates the perceived, relatively steep rate of increase of the curve estimate in the early years of the species' existence. The shape of the curve estimate might be said to be under greater threat from data-recording errors than from sheer noise in the data.

4. Theoretical properties

4.1. Archetypal processes

Rather than impose regularity conditions in an abstract way, we shall describe specific processes which produce the type of data discussed in Section 3, and use these processes to motivate our formal assumptions. The latter will be given in Section 4.2, and shown in the Appendix to be satisfied by the specific processes given here.

Bearing in mind characteristics of the data set studied in Section 3, we wish to address cases where the following assumption holds:

Assumption 4.1. *(i) The number of transitions over the entire observation period, and the number of observations made between transitions, can be moderately large. (ii) The probability distribution for transitions varies smoothly with time.*

Two slightly different specific processes defined on the interval $\mathcal{I} = [0, 1]$, and which produce realizations that satisfy Assumption 4.1, will be considered next. Both are based on a $q \times q$ transition probability matrix $\Pi(t) = (\rho_{ij}(t))$, defined for $t \in \mathcal{I}$, where ρ_{ij}, for $0 \leqslant i, j \leqslant q-1$, are continuous functions with each $\rho_{ij} \geqslant 0$ and $\sum_j \rho_{ij} = 1$.

For the first Markov process, transitions having transition probability matrix $\Pi(k\delta)$ are made at respective time points $k\delta$, for $1 \leqslant k \leqslant \delta^{-1}$, where $\delta = \delta(n) \downarrow 0$ as $n \to \infty$. For the second process, probabilities for the kth transition continue to be determined by $\Pi(k\delta)$, but the time at which the kth transition occurs is now $\delta \sum_{1 \leqslant i \leqslant k} Z_i$, where $Z_1, Z_2, \ldots$ are independent, exponentially distributed random variables with a mean which may depend on the current state of the process, but nevertheless always lies in the same fixed interval $[a, b]$, where $0 < a < b < \infty$. By restricting attention to $k \leqslant \max\{j : \sum_{i \leqslant j} Z_i \leqslant \delta^{-1}\}$ we ensure that the process is studied only within $\mathcal{I}$.

For either process, the state of the chain at any given time $t \in \mathcal{I}$ is that to which it moved at the most recent transition. The initial state is chosen arbitrarily. The fact that each ρ_{ij} is a continuous function guarantees that Assumption 4.1(ii) holds.

Note that, in the case of the first process, the functions p_i and p_{ij} are not continuous. Nevertheless, they are asymptotically smooth. For example, it can be shown that for each i, p_i satisfies, for each $\epsilon_1 \in (0, 1)$,

$$\lim_{\epsilon_2 \downarrow 0} \limsup_{n \to \infty} \sup_{t_1, t_2 \in [\epsilon_1, 1] \,:\, |t_1 - t_2| \leqslant \epsilon_2} |p_i(t_1) - p_i(t_2)| = 0$$

Even for the first process, but particularly for the second, it is clear that a completely parametric methodology will fail to address the rich class of possibilities that can arise. This consideration motivated the methodology developed in Section 2.

Either of the two processes is observed within $\mathcal{I}$ at time points $T_1 < \ldots < T_n$, which, here and in Section 4.2, will be taken to be the ordered values of n independent random variables having a common density, f, supported on $\mathcal{I}$. This is a standard assumption in nonparametric regression; see, for example, Wand and Jones (1995, Section 5.3.2), Fan and Gijbels (1996, p. 57) and Simonoff (1996, Chapter 5).

Strictly speaking, we should use triangular array notation, specifying that for each n, $T_k = T_{nk}$, for $1 \leqslant k \leqslant n$, are the ordered values of n independent random variables with common density f. No assumption about the relationship among rows of the array is required, however.

By allowing n and $\delta = \delta(n)$ to vary together, in such a way that

$$n\delta \to \infty, \qquad \delta = O(n^{-\epsilon}) \quad \text{for some } \epsilon > 0, \tag{4.1}$$

we ensure that Assumption 4.1(i) holds.

4.2. Main theoretical results

Let $\eta = \eta(\delta)$ denote any sequence of constants such that, as $\delta \to 0$,

$$\eta/(\delta|\log \delta|) \to \infty \quad \text{and} \quad \eta \to 0 \tag{4.2}$$

We shall study the Markov process only on the interval $\mathcal{I}_\delta = [\eta, 1]$, not on all of $\mathcal{I}$. This allows us to 'burn' the process in for a short time period, so that our results will be valid regardless of the starting state.

For each n, we work with a potentially different Markov process on $\mathcal{I}$. In particular, we allow the state probabilities $p_i(t)$, and the dual-state probabilities $p_{ij}(t_1, t_2)$, to depend on n, reflecting the implication of Assumption 4.1(i) that the problem becomes more complex as sample size increases. We ask that whenever η satisfies (4.2), $p_i(t)$ and $p_{ij}(t_1, t_2)$ satisfy:

$$\max_{0 \leqslant i \leqslant q-1} |p_i(t_1) - p_i(t_2)| \leqslant C_1\, \delta^{-1}\{\max(|t_1 - t_2|, \delta|\log \delta|)\}^2, \tag{4.3}$$

$$\max_{0 \leqslant i,j \leqslant q-1} |p_{ij}(t_1, t_2) - p_i(t_1)p_j(t_2)| \leqslant C_2 \exp(-C_3|t_1 - t_2|/\delta), \tag{4.4}$$

for all $t_1, t_2 \in \mathcal{I}_\delta$ and all n, where the constants C_1, C_2, C_3 do not depend on t_1, t_2 or n. In order that we might derive conditional probabilities from joint probabilities without the denominator vanishing, we assume that whenever η satisfies (4.2),

$$\liminf_{\delta \to 0}\, \inf_{t \in \mathcal{I}_\delta}\, \min_{0 \leqslant i \leqslant q-1} p_i(t) > 0. \tag{4.5}$$

In the Appendix we shall show that these conditions hold for the processes introduced in Section 4.1.

Note that conditions (4.3)–(4.5) do not require strict continuity of the functions p_i and p_{ij}. Indeed, as we observed in Section 4.1, continuity is not required of these probabilities in the cases of the archetypal processes which motivate our theory.

Of the kernel K, bandwidth h and density f we make the following assumption:

Assumption 4.2. *K is a compactly supported, Hölder-continuous probability density; for some $\epsilon > 0$, $h^2\delta^{-1} + h^{-1}\delta^{1-\epsilon} \to 0$ as $n \to \infty$, and f is continuous and non-vanishing on $\mathcal{I}$.*

In particular, the conditions on h are satisfied if $h = \text{const.}\, \delta^a$, where $\frac{1}{2} < a < 1$.

Even though we effectively are working with a triangular array of Markov processes (one process for each n), and shall show that strong laws of large numbers hold for our estimators, we require no assumptions about the relationships among the processes for different values of n.

Theorem 4.1. *If* (4 1)–(4.5) *and Assumption* 4 2 *hold then the local linear estimator,* $\hat{p}_i(t)$, *of the state probability,* $p_i(t)$, *is strongly consistent, in the sense that*

$$\sup_{t\in\mathcal{I}_\delta} \max_{0\leqslant i\leqslant q-1} |\hat{p}_i(t) - p_i(t)| \to 0 \tag{4.6}$$

with probability 1.

Theorem 4.2. *If* c_δ *is any sequence of positive constants such that* $\delta/c_\delta \to 0$ *as* $n \to \infty$, *and if* (4 1)–(4 5) *and Assumption* 4 2 *hold, then the local linear estimators* $\hat{p}_{ij}(t_1, t_2)$ *are strongly consistent, in the sense that, with probability 1,*

$$\sup_{t_1\ t_2\in\mathcal{I}_\delta\,:\,|t_1-t_2|>c_\delta} \max_{0\leqslant i,j\leqslant q-1} |\hat{p}_{ij}(t_1, t_2) - p_{ij}(t_1, t_2)| \to 0. \tag{4.7}$$

Therefore, the estimated transition probabilities $\hat{p}_{ij}(t_2|t_1)$, *defined at* (2 5), *are strongly consistent with probability 1,*

$$\sup_{t_1\ t_2\in\mathcal{I}_\delta\,:\,|t_1-t_2|>c_\delta} \max_{0\leqslant i,j\leqslant q-1} |\hat{p}_{ij}(t_2|t_1) - p_{ij}(t_2|t_1)| \to 0 \tag{4.8}$$

In making Assumption 4.2, which involves h, in Theorem 4.2, we are not asking that the same bandwidth be used to compute both $\hat{p}_{ij}$ and $\hat{p}_i$ (Recall that $\hat{p}_i$ is used during construction of $\hat{p}_{ij}$; see (2.6).) It is necessary only that the bandwidth $h = h_1$ used to compute $\hat{p}_i$, and the bandwidth $h = h_2$ employed in the definition of each $K_k(t)$ in (2.6), both satisfy Assumption 4 2

For consistent inference it is essential that the bandwidth h be an order of magnitude larger than δ. This is because the amount of information about either p_i or p_{ij}, within an interval of length only a constant multiple of δ, remains bounded as $\delta \to 0$. In particular, if either of the processes discussed in Section 4 1 was observed throughout an interval of length $C\delta$ straddling t, and was not observed anywhere else, then the transition probability matrix $\Pi(t)$ could not be estimated consistently.

Since consistent estimation of $p_{ij}(t_2|t_1)$ is, in general, possible only when $t_2 - t_1$ is of larger order than δ, then if the Markov process is simulated by implementing transitions of a Markov chain on a grid with edge width τ (see Section 2 2), τ should satisfy $\delta/\tau \to 0$.

5. Technical arguments

Proof of Theorem 4.1. Note that, since the T_k are ordered values of independent random variables, and A_j is an order-invariant sum of functions of individual T_k, then $A_j(t)$ equals a sum of independent random variables. Using this property, and either standard arguments involving the Hungarian embedding (Komlós *et al.* 1976), or the technique we shall employ below to establish (5.6), it may be proved that

$$\sup_{t\in\mathcal{I}}|A_j(t)-\mathrm{E}\{A_j(t)\}|\to 0 \tag{5.1}$$

with probability 1, as $n\to\infty$, for $j=0,1,2$. (This result is valid in the triangular array setting discussed in Section 4.) The part of Assumption 4.2 pertaining to f implies that each $\mathrm{E}\{A_j(t)\}$ is uniformly bounded, and that $\mathrm{E}\{A_2(t)\}\,\mathrm{E}\{A_0(t)\}-\{\mathrm{E}A_1(t)\}^2$ is uniformly bounded above zero. Hence, by (5.1),

$$\limsup_{n\to\infty}\sup_{t\in\mathcal{I}}\max_{j=0,1,2}|A_j(t)|<\infty, \qquad \liminf_{n\to\infty}\inf_{t\in\mathcal{I}}\{A_2(t)A_0(t)-A_1(t)^2\}>0, \tag{5.2}$$

both results holding with probability 1.

Equations (2.3) may equivalently be written as

$$\sum_{j=0}^{q-2}(1+\delta_{ij})a_j=p_i(t)+r_{q-1}(t)+\frac{\Delta_i(t)+\Delta(t)}{A_2(t)A_0(t)-A_1(t)^2}, \tag{5.3}$$

where $r_i(t)=P\{X(t)\neq i\}$,

$$\Delta_i(t)=A_2(t)B_{i0}(t)-A_1(t)B_{i1}(t)-p_i(t)\{A_2(t)A_0(t)-A_1(t)^2\},$$

$$\Delta(t)=A_2(t)C_{q-1,0}(t)-A_1(t)C_{q-1,1}(t)-r_{q-1}(t)\{A_2(t)A_0(t)-A_1(t)^2\}$$

The equations

$$\sum_{j=0}^{q-2}(1+\delta_{ij})a_j=p_i(t)+r_{q-1}(t),$$

for $0\leqslant i\leqslant q-2$, have unique solution $a_i=p_i(t)$, for the same range of i. Note too that $1-C_{q-1,j}=\sum_{i\leqslant q-2}B_{ij}$, and that $1-r_{q-1}=\sum_{i\leqslant q-2}p_i$. Therefore, provided we prove that

$$\sup_{t\in\mathcal{I}_\delta}\max_{0\leqslant i\leqslant q-2}|\Delta_i(t)|\to 0 \tag{5.4}$$

with probability 1, the theorem will follow from the second part of (5.2) and from (5.3). It remains to derive (5.4).

Put $L_{jk}(t)=K_k(t)\{(t-X_k)/h\}^j$, let $\mathcal{T}$ denote the sigma-field generated by $T_1,\ldots,T_n$, and observe that $\Delta_i=A_2\,\Delta_{i0}-A_1\,\Delta_{i1}$, where

$$\Delta_{ij}(t)=\frac{1}{nh}\sum_{k=1}^{n}L_{jk}(t)\{I_{ki}-p_i(t)\}.$$

Hence, in view of the first part of (5.2), (5.4) will follow if we prove that

$$\sup_{t\in\mathcal{I}_\delta}\max_{0\leqslant i\leqslant q-2,\,j=0,1}|\mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}|\to 0, \tag{5.5}$$

$$\sup_{t\in\mathcal{I}_\delta}\max_{0\leqslant i\leqslant q-2,\,j=0,1}|\Delta_{ij}(t)-\mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}|\to 0 \tag{5.6}$$

with probability 1.

Using (4.3), and the fact that the kernel K is compactly supported; and defining

$\ell_n = (\log n)^2$, we deduce that $|\mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}|$ is dominated, uniformly in t, by a constant multiple of

$$\frac{1}{nh}\sum_{k=1}^{n} \max\bigl(|\mathcal{T}_k - t|^2/\delta,\, \delta\,\ell_n\bigr)\, K_k(t) = O\bigl(h^2\,\delta^{-1} + \delta\,\ell_n\bigr), \tag{5.7}$$

with probability 1, uniformly in $t \in \mathcal{I}_\delta$, where we have used the first part of (5.2) to derive the bound. Result (5.5) follows from (5.7), the bound on δ at (4.1), and the condition on h in Assumption 4.2.

Next we establish (5.6). Note that, for $1 \leqslant k_1 < \ldots < k_s \leqslant n$,

$$\mathrm{E}(I_{k_1 i} \ldots I_{k_s i}|\mathcal{T}) = p_i(\mathcal{T}_{k_1})p_{ii}(\mathcal{T}_{k_2}|\mathcal{T}_{k_1}) \ldots p_{ii}(\mathcal{T}_{k_s}|\mathcal{T}_{k_{s-1}}). \tag{5.8}$$

This property, (4.4) and (4.5) may be used to show that if $1 \leqslant k_1 \leqslant \ldots \leqslant k_s \leqslant n$, and if, for some $a \in [1, s-1]$, $\mathcal{T}_{k_{a+1}} - \mathcal{T}_{k_a} > B\delta\log n$, where $B > 0$ is a constant, then

$$|\mathrm{E}[\{I_{k_1 i} - p_i(\mathcal{T}_{k_1})\} \ldots \{I_{k_s i} - p_i(\mathcal{T}_{k_s})\}|\mathcal{T}]| \leqslant C_1(s)\exp\{-C_2(s)B\log n\},$$

where the constants C_1 and C_2 depend on s but not on n or B. (Here and below, C_j denotes a positive constant.) Therefore,

$$\begin{aligned} &\mathrm{E}[|\Delta_{ij}(t) - \mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}|^{2r}|\mathcal{T}] \\ &\quad\leqslant \frac{C_3(r)}{(nh)^{2r}} \sum_{k_1=1}^{n} \ldots \sum_{k_{2r}=1}^{n} w(k_1, \ldots, k_{2r})K_{k_1}(t) \ldots K_{k_{2r}}(t), \end{aligned} \tag{5.9}$$

where $w(k_1, \ldots, k_{2r}) = \exp\{-C_2(2r)B\log n\} + w_1(k_1, \ldots, k_{2r})$ and $w_1(k_1, \ldots, k_{2r})$ denotes the indicator of the event 'for each k_a there exists k_b, with $b \neq a$, such that $|\mathcal{T}_{k_a} - \mathcal{T}_{k_b}| \leqslant B\delta\log n$'.

It can be shown that, uniformly in $t \in \mathcal{I}_\delta$ and for each choice of i and j,

$$\mathrm{E}\left\{\frac{1}{(nh)^{2r}} \sum_{k_1=1}^{n} \ldots \sum_{k_{2r}=1}^{n} w_1(k_1, \ldots, k_{2r})K_{k_1}(t) \ldots K_{k_{2r}}(t)\right\} = C_4(B, r)(h^{-1}\delta\log n)^r \tag{5.10}$$

Combining (5.9) and (5.10) it can be proved that

$$\mathrm{E}[|\Delta_{ij}(t) - \mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}|^{2r}] = O\{(h^{-1}\delta\log n)^r\}, \tag{5.11}$$

uniformly in the same sense, provided B is chosen so large that $C_2(2r)B > r$.

Using (5.11), and Markov's inequality, we may prove that for each $C, \epsilon > 0$,

$$\sup_{t\in\mathcal{I}_\delta} P[|\Delta_{ij}(t) - \mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}| > (h^{-1}\delta n^{\epsilon})^{1/2}] = O(n^{-C})$$

Therefore, provided $\mathcal{J}_n \subseteq \mathcal{I}_\delta$ contains no more than $O(n^D)$ elements for some $D > 0$, we have for each $C, \epsilon > 0$,

$$P\left[\sup_{t\in\mathcal{J}_n} |\Delta_{ij}(t) - \mathrm{E}\{\Delta_{ij}(t)|\mathcal{T}\}| > (h^{-1}\delta n^{\epsilon})^{1/2}\right] = O(n^{-C}) \tag{5.12}$$

This property; an approximation to $\Delta_{ij}(t)$, for arbitrary t, by values of $\Delta_{ij}(t)$ for t on an equally spaced grid with separation n^{-D}, for any fixed $D > 0$; Assumption 4.2 on h; and use of the Hölder continuity of K noted in Assumption 4.2; imply that (5.12) continues to hold if $\mathcal{J}_n$ there is replaced by $\mathcal{I}_\delta$. Using the Borel–Cantelli lemma we see that this entails (5.6), which completes the proof. □

Proof of Theorem 4.2. It is necessary only to prove (4.7). Since, in (4.7), t_1 and t_2 are constrained to satisfy $|t_1 - t_2| > c_\delta$, and since (4.4) implies that $|p_{ij}(t_1, t_2) - p_i(t_1)p_j(t_2)| \to 0$ uniformly in such t_1 and t_2, it suffices to establish the version of (4.7) in which $p_{ij}(t_1, t_2)$ is replaced by $p_i(t_1)p_j(t_2)$. This we shall do below.

Arguing as in the proof of Theorem 4.1, we see that it suffices to derive two-state analogues of (5.5) and (5.6), which here can be reduced to:

$$\max_u \sup_{t_1, t_2 \in \mathcal{I}_\delta} \max_{0 \leqslant i,j \leqslant q-2} |\mathrm{E}\{\Delta_{ij}(t_1, t_2, u)|\mathcal{T}\}| \to 0, \tag{5.13}$$

$$\max_u \sup_{t_1, t_2 \in \mathcal{I}_\delta} \max_{0 \leqslant i,j \leqslant q-2} |\Delta_{ij}(t_1, t_2, u) - \mathrm{E}\{\Delta_{ij}(t_1, t_2, u)|\mathcal{T}\}| \to 0, \tag{5.14}$$

where

$$\Delta_{ij}(t_1, t_2, u) = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \{I_{ki} I_{\ell j} - p_i(t_1)p_j(t_2)\}\, u(k, \ell) K_k(t_1) K_\ell(t_2),$$

and there are three functions u, defined by $u(k, \ell)$ taking the values 1, $(T_k - t_1)/h$ or $(T_\ell - t_2)/h$, respectively. Note too that

$$\mathrm{E}\{\Delta_{ij}(t_1, t_2, u)|\mathcal{T}\} = \sum_{a=1}^{3} W_{ija}(t_1, t_2, u), \tag{5.15}$$

where

$$W_{ij1}(t_1, t_2, u) = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \{p_{ij}(T_k, T_\ell) - p_i(T_k)p_j(T_\ell)\}\, u(k, \ell) K_k(t_1) K_\ell(t_2),$$

$$W_{ij2}(t_1, t_2, u) = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} p_i(T_k)\{p_j(T_\ell) - p_j(t_2)\}\, u(k, \ell) K_k(t_1) K_\ell(t_2),$$

$$W_{ij3}(t_1, t_2, u) = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \{p_i(T_k) - p_i(t_1)\}\, p_j(t_2) u(k, \ell) K_k(t_1) K_\ell(t_2)$$

By (4.3),

$$|W_{ij2}(t_1, t_2, u)| \leqslant C_8\, A_0(t) \frac{1}{nh} \sum_{\ell=1}^{n} \max\big(|\mathrm{T}_\ell - \mathrm{t}_2|^2/\delta,\, \delta\, \ell_\mathrm{n}\big)\, K_\ell(\mathrm{t}_2).$$

A bound for the series on the right-hand side is given in (5.7). Using that result and the first part of (5.2) we deduce that, in the case $a = 2$,

$$\max_u \sup_{t_1, t_2 \in \mathcal{I}} \max_{0 \leqslant i,j \leqslant q-2} |W_{ija}(t_1, t_2, u)| = O\big(h^2 \delta^{-1} + \delta \ell_n\big), \tag{5 16}$$

with probability 1 A similar argument shows that (5 16) holds for $a = 3$.

Using (4 4), and the fact that K is compactly supported, we see that, for some $C_7 > 0$ and each $B > 0$, $|W_{ij1}(t_1, t_2, u)|$ is dominated by a constant multiple of

$$\frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \exp(-C_7 |T_k - T_\ell|/\delta) K_k(t_1) K_\ell(t_2) \leqslant n^{-B} A_0(t)^2 + W(t_1, t_2), \tag{5 17}$$

where

$$W(t_1, t_2) = \frac{1}{(nh)^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} I\big(|T_k - T_\ell| \leqslant B\, C_7^{-1} \delta \log n\big)\, K_k(t_1) K_\ell(t_2) \tag{5.18}$$

Although $T_1, \ldots, T_n$ are the ordered values of a sequence of independent and identically distributed random variables, $W(t_1, t_2)$ is invariant under permutations of this sequence, and so in deriving a bound for $W(t_1, t_2)$ we may treat the T_j as though they were the original independent and identically distributed quantities. Under this assumption it is straightforward, although algebraically complex, to derive bounds for $\mathrm{E}(W^r)$ for any integer $r \geqslant 1$, obtaining

$$\sup_{t_1 \, t_2 \in \mathcal{I}} \mathrm{E}\{W(t_1, t_2)^r\} \leqslant C_8(r)\big(h^{-1} \delta \log n\big)^r \tag{5.19}$$

Combining (5 17)–(5 19) we deduce that, for each integer $r \geqslant 1$,

$$\max_u \sup_{t_1 \, t_2 \in \mathcal{I}} \max_{0 \leqslant i,j \leqslant q-2} \mathrm{E}\big\{|W_{ij1}(t_1, t_2, u)|^r\big\} \leqslant C_{10}(r)\big(h^{-1} \delta \log n\big)^r$$

This result, and the argument leading to (5 6), may be used to prove that

$$\max_u \sup_{t_1 \, t_2 \in \mathcal{I}_\delta} \max_{0 \leqslant i,j \leqslant q-2} |W_{ij1}(t_1, t_2, u)| \to 0 \tag{5 21}$$

with probability 1 Properties (5 15), (5 16) for $a = 2, 3$, and (5 21) imply (5 13) The argument used to derive (5 14) is similar to that employed to obtain (5 6) □

Appendix: Motivation for assumptions in Section 4.2

Here we motivate the assumptions in Section 4.2 by showing that they are satisfied by the Markov processes introduced in Section 4 1. Indeed, it is sufficient to address the first process, since the second can be treated similarly.

Assume that the sequence of transition probability matrices $\Pi(t)$, $t \in \mathcal{I}$, introduced in Section 4 1, enjoys the following property:

Assumption A.1. *(i) For each $t \in \mathcal{I}$, there is just one eigenvalue of $\Pi(t)$ with absolute value equal to 1. (ii) For each $t \in \mathcal{I}$, all states of the stationary chain with transition probability matrix $\Pi(t)$ communicate. (iii) The components of $\Pi(t)$ have a bounded derivative, uniformly in $t \in \mathcal{I}$*

We shall prove that, if the data are generated by the first of the two Markov processes introduced in Section 4.1, and if conditions (4.1) and (4.2) hold, then Assumption A.1 implies (4.3)–(4.5).

Let k, ℓ be integers satisfying $\eta\delta^{-1} \leqslant k < \ell \leqslant \delta^{-1}$, where η is as in (4.2); and define $\Pi_r = \Pi\{(k+r)\delta\}$, for integers r. Write $(Q)_{ij}$ for the (i, j)th component of a matrix Q. In view of Assumption A.1(iii), $|(\Pi_r)_{ij} - (\Pi_0)_{ij}| \leqslant \text{const}\, |r|\, \delta$, where, here and below, 'const.' denotes a generic positive constant which does not depend on i, j, k, ℓ, n, r or δ. Hence, for $r \geqslant 1$,

$$\left|(\Pi_1 \Pi_2 \ldots \Pi_r)_{ij} - \left(\Pi_0^r\right)_{ij}\right| \leqslant \text{const.}\, r^2\delta, \tag{A.1}$$

$$\left|(\Pi_{1-r} \Pi_{2-r} \ldots \Pi_0)_{ij} - \left(\Pi_0^r\right)_{ij}\right| \leqslant \text{const.}\, r^2\delta. \tag{A.2}$$

Let $\lambda(t)$ denote the second largest absolute value of the eigenvalues of $\Pi(t)$, and put $\lambda_0 = \sup_{t\in\mathcal{I}} \lambda(t)$. Assumptions A.1(ii)–(iii) imply that $\lambda_0 < 1$. Writing π_* for the stationary distribution of Π_0, that is, for the solution of $\pi_*^{\mathrm{T}} \Pi_0 = \pi_*^{\mathrm{T}}$, we have

$$\|\pi^{\mathrm{T}} \Pi_0^r - \pi_*^{\mathrm{T}}\| \leqslant \text{const.}\, \lambda_0^r, \tag{A.3}$$

uniformly in $r \geqslant 0$, in probability distributions π and in choices $k \in [\eta\delta^{-1}, \delta^{-1} - r - 1]$ in the definition $\Pi_0 = \Pi(k\delta)$. Write π_m for the vector of state probabilities at time $(k+m)\delta$. If $k \geqslant |\log\delta|/|\log\lambda_0| \geqslant k-1$ then, by (4.2), (A.2) and (A.3),

$$\begin{aligned} \|\pi_0 - \pi_*\| &= \|\pi_{-k}^{\mathrm{T}} \Pi_{1-k} \Pi_{2-k} \ldots \Pi_0 - \pi_*^{\mathrm{T}}\| \\ &\leqslant \|\pi_{-k}^{\mathrm{T}} \Pi_0^k - \pi_*^{\mathrm{T}}\| + \text{const.}\, k^2\,\delta \leqslant \text{const}\ \delta(\log\delta)^2. \end{aligned} \tag{A.4}$$

Therefore, using (A.1), we see that for $r \geqslant 1$,

$$\|\pi_0 - \pi_r\| = \|\pi_0^{\mathrm{T}} - \pi_0^{\mathrm{T}} \Pi_1 \Pi_2 \ldots \Pi_r\| \leqslant \text{const.}\, \delta^{-1}\{\max(|t_1 - t_2|, \delta|\log\delta|)\}^2, \tag{A.5}$$

where $t_1 = k\delta$ and $t_2 = (k+r)\delta$. Using the definition of the Markov process, it is straightforward to extend the bounds at (A.5) to t_1, t_2 defined in the continuum. This implies (4.3).

To derive (4.5), note that in view of Assumption A.1(ii), for each $0 \leqslant i, j \leqslant q-1$ and each t, there is a positive probability of transiting from state i to state j in a finite number of steps, in the homogeneous chain with transition probability matrix $\Pi(t)$. Since $\Pi(t)$ is a componentwise continuous function of t then there exists a finite integer, $\nu(t) \geqslant 1$, and $0 < \sigma(t) < 1$, such that (a) $\nu_{\max} \equiv \sup_{t\in\mathcal{I}} \nu(t) < \infty$, (b) $\sigma_{\min} \equiv \inf_{t\in\mathcal{I}} \sigma(t) > 0$, and (c) for each (i, j), the probability of transiting from state i to state j in just $\nu(t)$ steps is not less than $\sigma(t)$. From these results and (A.1) we deduce that, for each (i, j), the probability of transiting from state i to state j in at most $\nu_{\max}$ steps, starting from time point $k\delta$, in the inhomogeneous chain introduced in Section 4.1, is not less $\sigma_{\min} + O(\delta)$, uniformly in $k \in [0, \delta^{-1} - \nu - 1]$. Therefore, $\inf_{t\in\mathcal{I}} \min_i p_i(t) \geqslant \sigma_{\min}^2 + O(\delta)$, from which (4.5) follows.

Next, to derive (4.4), assume without loss of generality that $t_1 < t_2$, and observe that since (4.5) holds it is sufficient to prove that for each pair (i, j),

$$|p_{ij}(t_2|t_1) - p_j(t_2)| \leqslant C_2 \exp(-C_3|t_1 - t_2|/\delta).$$

Again it is adequate to derive the result on the grid, that is, for $t_1 = k\delta$ and $t_2 = (k + r)\delta$; and for that it is enough to prove that, uniformly in probability measures $\pi_{(1)}$ and $\pi_{(2)}$, and in k and r,

$$\|(\pi_{(1)} - \pi_{(2)})^{\mathrm{T}} \Pi_1 \Pi_2 \ldots \Pi_r\| \leqslant C_4 \exp(-C_5 r). \tag{A.6}$$

If $\zeta > 0$ is sufficiently small, although fixed, and if s does not exceed the largest integer less than $\zeta|\log \delta|$, then, in view of (4.1), $s^2\delta \leqslant \text{const.}\, \lambda_0^s$. Hence, by (A.1) and (A.3),

$$\|(\pi_{(1)} - \pi_{(2)})^{\mathrm{T}} \Pi_1 \Pi_2 \ldots \Pi_s\| \leqslant \|(\pi_{(1)} - \pi_{(2)})^{\mathrm{T}} \Pi_0^s\| + \text{const.}\, s^2\delta \leqslant \text{const.}\, \lambda_0^s \tag{A.7}$$

To bound $\|(\pi_{(1)} - \pi_{(2)})^{\mathrm{T}} \Pi_1 \Pi_2 \ldots \Pi_r\|$ for general $r \geqslant s$, break r into r/s blocks of length s, and iterate the bound obtained at (A.7), renormalizing the new version of $\pi_{(1)} - \pi_{(2)}$ at each step by dividing by a constant multiple of λ_0^s. In this way we may show that if $\eta\delta^{-1} \leqslant k \leqslant k + r \leqslant \delta^{-1}$ then

$$\|(\pi_{(1)} - \pi_{(2)})^{\mathrm{T}} \Pi_1 \Pi_2 \ldots \Pi_r\| \leqslant \text{const.}\, \lambda_1^r,$$

where $\lambda_1 \in [\lambda_0, 1)$ does not depend on r. This implies (A.6).

Aknowledgements

We are grateful to two reviewers for helpful comments, and to the hospitality of the Department of Statistics at the London School of Economics, where much of this work was done.

References

Cheetham, A.H. (1986) Tempo of evolution in a Neogene bryozoan: rates of morphologic change within and across species boundaries. *Paleobiology*, **12**, 190–202.

Cheetham, A.H. (1987) Tempo of evolution in a Neogene bryozoan: are trends in single morphologic characters misleading? *Paleobiology*, **13**, 286–296.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes*. London: Methuen.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Gould, S.J. (1991) Opus 200. *Natural History*, **100**, 12–18.

Greenwood, P.E. and Wefelmeyer, W. (1994) Nonparametric estimators for Markov step processes. *Stochastic Process. Appl.*, **52**, 1–16.

Hart, J.D. (1991) Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B*, **53**, 173–187.

Jackson, J.B.C. and Cheetham, A.H. (1990) Evolutionary significance of morphospecies: A test with cheilostome Bryozoa. *Science*, **248**, 579–583.

Karlsen, H.A. and Tjøstheim, D. (2001) Nonparametric estimation in null recurrent time series. *Ann. Statist.*, **29**, 372–416.

Komlós, J., Major, P. and Tusnády, G. (1976) An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **34**, 33–58.

Masry, E. (1996) Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.*, **17**, 571–599.

Masry, E. (2001) Local linear regression estimation under long-range dependence: strong consistency and rates. *IEEE Trans. Inform. Theory*, **47**, 2863–2875.

Masry, E. and Mielniczuk, J. (2001) Local linear regression estimation for time series with long-range dependence. *Stochastic Process. Appl.*, **82**, 173–193.

Robinson, P.M. (1983) Nonparametric estimators for time series. *J. Time Ser. Anal.*, **4**, 185–207.

Robinson, P.M. (1997) Large sample inference for nonparametric regression with dependent errors. *Ann. Statist.*, **25**, 2054–2083.

Robinson, P.M. and Hidalgo, F.J. (1997) Time series regression with long-range dependence. *Ann. Statist.*, **25**, 77–104.

Roussas, G. (1990) Nonparametric regression estimation under mixing conditions. *Stochastic Process. Appl.*, **36**, 107–116.

Roussas, G., Tran, L. and Ioannides, D.A. (1992) Fixed design regression for time series: asymptotic normality. *J. Multivariate Anal.*, **40**, 262–291.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Tran, L. (1993) Nonparametric function estimation for time series by local average estimators. *Ann. Statist.*, **21**, 1040–1057.

Tran, L., Roussas, G., Yakowitz, S. and Truong Van, B. (1996) Fixed-design regression for linear time series. *Ann. Statist.*, **24**, 975–991.

Truong, Y.K. and Stone, C.J. (1992) Nonparametric function estimation involving time series. *Ann. Statist.*, **20**, 77–97.

Utikal, K.J. (1997) Nonparametric inference for Markovian interval processes. *Stochastic Process. Appl.*, **67**, 1–23.

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

Wu, J.S. and Chu, C.K. (1994) Nonparametric estimation of a regression function with dependent observations. *Stochastic Process. Appl.*, **50**, 149–160.

Yakowitz, S. (1993) Nearest neighbor regression estimation for null-recurrent Markov times series. *Stochastic Process. Appl.*, **48**, 311–318.

Received May 2003 and revised February 2004