

## **USING THE SMOOTHED BOOTSTRAP FOR STATISTICAL INFERENCE FOR MARKOV CHAINS**

**Alan M. Polansky<sup>1</sup>**

<sup>1</sup> Division of Statistics, Northern Illinois University, De Kalb, IL, USA  
Email: polansky@math.niu.edu

### **ABSTRACT**

Markov chains provide a flexible model for dependent random variables with applications in such disciplines as physics, environmental science and economics. Recently the bootstrap has been used to aid in the development of statistical methods based on observed realizations from Markov chains. The bootstrap method estimates parameters of the Markov chain with an unknown transition probability matrix with those from a Markov chain with a transition probability matrix estimated using the observed realization. Unfortunately, when the length of the observed realization is not sufficiently large, the properties of the estimated transition probability matrix are often very different from those of the actual transition probability matrix. This can lead to large errors associated with the bootstrap estimates. This paper presents simple multinomial type smoothing techniques that can be applied to the estimated transition probability matrix to alleviate some of these difficulties. It is demonstrated through empirical studies that the use of the smoothed transition probability matrix can increase the reliability of the bootstrap method for Markov chains. The practical use of the method is demonstrated through an example.

### **KEYWORDS**

Dependence, Likelihood, Mean Hitting Time, Multinomial.

**2000 Mathematics Subject Classifications:** 62M05, 62G09.

## 1 INTRODUCTION

Let  $\{X_k\}_0^\infty$  be a sequence of discrete random variables, each with countable support  $S$ . The set  $S$  can contain any countable collection of distinct elements, but for simplicity this paper will assume that  $S = \{1, 2, \dots, c\}$  where  $c \in \mathbb{N} = \{1, 2, \dots\}$ . In the context of stochastic processes, the index  $k$  is generally a time index and the value of  $X_k$  is called the state of the process at time  $k$ . Let  $P$  be a  $c \times c$  matrix where  $P$  has  $(i, j)^{th}$  element  $0 \leq P_{ij} \leq 1$  and

$$\sum_{j=1}^c P_{ij} = 1,$$

for each  $i \in S$ . The  $i^{th}$  row of  $P$  is the conditional probability distribution of  $X_k$  given  $X_{k-1} = i$ . Hence  $P(X_k = j | X_{k-1} = i) = P_{ij}$  for all  $i \in S$  and  $j \in S$ . The processes studied in this paper will be assumed to have the first-order Markov property, which implies

$$P(X_k = j | X_{k-1} = i, \dots, X_0 = x_0) = P(X_k = j | X_{k-1} = i) = P_{ij},$$

for all sequences of constants  $\{x_m\}_0^{i-2}$  where each  $x_m \in S$ . The process  $\{X_k\}_0^\infty$  is known as a Markov chain with finite state space  $S$ . A review of Markov Chains can be found in Karlin and Taylor (1975). Such processes are commonly used to model dependent processes observed in such disciplines as biology, computer science, environmental science, geography, social science, physics and economics. See, for example, Clark (1965), Fuh (1993), Geary (1978), Gottschau (1992, 1994), Turchin (1986) and Yang (1979). In the applied setting, the transition probability matrix  $P$  is unknown and a finite realization  $X_0, \dots, X_n$  of the process is observed in order to estimate and test hypotheses about  $P$ . A significant amount of research has been devoted to this problem. See Basawa and Rao (1980) and Billingsley (1961a,b) for an overview of much of the relevant research.

Because the transition probability matrix completely specifies the probabilistic behavior of a Markov chain, all of the relevant parameters of interest are functions of the transition probability matrix. Hence, let  $\mathcal{P}$  be the set of  $c \times c$  transition probability matrices. Then for  $P \in \mathcal{P}$ , any parameter of interest of a Markov chain can be expressed as  $\theta = T(P)$  where  $T : \mathcal{P} \rightarrow \mathbb{R}^p$ , where  $p$  is the dimension of the parameter space of  $\theta$  which will depend on the problem of interest. The parameter  $\theta$  can be estimated using the simple plug-in estimate  $\hat{\theta} = T(\hat{P})$  where  $\hat{P} \in \mathcal{P}$  is an estimate of  $P$  computed on the observed realization  $X_0, \dots, X_n$ . Such estimates are known as bootstrap estimates. In some cases  $T(\hat{P})$  does not exist in a closed analytical form but can be approximated using simulated observed realizations from a Markov chain whose transition matrix is  $\hat{P}$ . See Athreya and Fuh (1992), Basawa, Green,

McCormick and Taylor (1990), Datta and McCormick (1992), Fuh (1993) and Kulperger and Rao (1989), for a review of the relevant methods and results.

Unfortunately, when the length of the observed realization is not sufficiently large, many of the properties of the estimated transition probability matrix can differ greatly from the actual transition probability matrix of the observed realization. For example, it might be known *a priori* for an observed Markov chain that  $P \in \mathcal{P}^{\text{com}} \subset \mathcal{P}$ , where  $\mathcal{P}^{\text{com}}$  is the set of all transition probability matrices where all states communicate. However, this property is not guaranteed to hold for the usual estimate of the transition probability matrix, particularly when the length of the observed realization is not large enough. Therefore, while  $\hat{P} \in \mathcal{P}$  it may be that  $\hat{P} \notin \mathcal{P}^{\text{com}}$ . It is clear that this can have a large effect on the accuracy of estimating certain parameters, for example the mean hitting time for certain states.

This paper presents a simple method for smoothing the estimated transition probability matrix to address some of these problems. In particular, the smooth estimated transition probability matrix will insure communication between all states and will be aperiodic. Section 2 develops the smooth estimated transition probability matrix. Section 3 provides empirical evidence that the proposed method has reasonable theoretical properties. Section 4 applies the proposed methodology to an example observed realization from a Markov chain. A general discussion of the methodology is given in Section 5.

## 2 SMOOTHED BOOTSTRAP ESTIMATES

Suppose  $X_0, \dots, X_n$  is a realization of length  $n + 1$  from a Markov chain with state space  $S$  and transition probability matrix  $P$ . For the remainder of this paper we will assume that  $P \in \mathcal{P}^z$ , the set of all transition probability matrices with non-zero entries. Such Markov chains have a single aperiodic ergodic class, and have stationary probabilities. Simple modifications can be made to the proposed methods when the transition probability matrix has a single ergodic class, but may have structural zeros. Bartlett (1951) shows that the nonparametric maximum likelihood estimator of the transition probability matrix estimates the conditional probability of moving to state  $j$ , given that the process is currently in state  $i$ , with the ratio of the number of transitions in the realization from state  $i$  to state  $j$  to the number of observed transitions in the realization from state  $i$ . Hence, let

$$\Omega_{ij} = \sum_{k=0}^{n-1} \delta(X_{k+1} = j, X_k = i),$$

and

$$\Omega_i = \sum_{k=0}^{n-1} \delta(X_k = i),$$

where  $\delta$  is the indicator function. Hence, the nonparametric maximum likelihood estimator of  $P$  is given by  $\hat{P}$  whose  $(i, j)^{th}$  element is

$$\hat{P}_{ij} = \begin{cases} \Omega_{ij}/\Omega_i & \text{if } \Omega_i > 0; \\ 0 & \text{if } \Omega_i = 0. \end{cases} \quad (2.1)$$

The assignment of the elements in row  $i$  of  $\hat{P}$  is somewhat arbitrary when  $\Omega_i = 0$ , and conventions have been suggested other than what is used in Equation (2.1). For example, another common convention is to set  $\hat{P}_{ij} = \delta(i = j)$  for  $i = 1, \dots, c$  when  $\Omega_i = 0$ . The estimate given in Equation (2.1) will be used for the remainder of the paper. The properties of the estimate given in Equation (2.1) are very closely related to the multinomial distribution, which provides the basis for studying the theoretical properties of the estimator. In particular,  $\hat{P}$  is consistent and has an asymptotic  $c^2$ -dimensional multivariate normal distribution, where each dimension corresponds to one of the  $c^2$  parameters in  $P$ . For additional properties, see Bartlett (1951), Billingsley (1961a,b) and Basawa and Rao (1980).

Because each row of  $P$  is required to sum to one, the total number of parameters estimated in the transition probability matrix is  $c(c - 1)$ . A sample size of at least  $c^2$  is required before each possible transition could be observed even once. Less likely transitions may require moderate to large sample sizes before they are observed. When a particular transition does not occur within an observed realization from the Markov chain, the corresponding estimated transition probability is zero. This can cause major problems when one wishes to estimate some properties of the Markov chain from the estimated transition probability matrix.

For example, consider a three state Markov chain with transition probability matrix in  $\mathcal{P}$  given by

$$P = \begin{bmatrix} 0.500 & 0.475 & 0.025 \\ 0.475 & 0.500 & 0.025 \\ 0.450 & 0.450 & 0.100 \end{bmatrix}. \quad (2.2)$$

Suppose a realization of length  $n = 30$  is observed from the chain. One such simulated realization is given in Table 2.1. From the observed realization it is clear that the maximum

Table 1: A simulated realization of length  $n = 30$  from a Markov chain with transition probability matrix given in Equation (??). The initial state is  $x_0 = 1$ . Observations  $X_1, \dots, X_{30}$  are given in the table. The remainder of the observed realization is read left to right across each row.

1	2	2	1	2	1	2	2	2	2
1	1	2	2	2	2	2	2	2	1
1	2	1	2	2	2	2	1	1	2

likelihood estimator of  $P$  is

$$\hat{P} = \begin{bmatrix} 0.300 & 0.700 & 0.000 \\ 0.316 & 0.684 & 0.000 \\ 0.000 & 0.000 & 0.000 \end{bmatrix}.$$

Define the first hitting time for state  $k \in S$  given  $X_0 = x_0 \in S$  as

$$G_{x_0,k} = \begin{cases} \inf\{n \geq 1 : X_n = k\} \\ \infty \text{ if no such } n \text{ exists,} \end{cases}$$

and the mean hitting time for state  $k \in S$  as  $\theta_{x_0,k} = E[G_{x_0,k}] = T_{x_0,k}(P)$ . The bootstrap estimate of  $\theta_{x_0,k}$  is  $\hat{\theta}_{x_0,k} = T_{x_0,k}(\hat{P})$ . See Athreya and Fuh (1992) for the asymptotic properties of these estimators. Note that since the third column of  $\hat{P}$  is the zero vector, State 3 is not accessible from States 1 or 2 in a Markov chain with transition probability matrix given by  $\hat{P}$ . Hence, conditional on the initial state  $x_0 = 1$ ,  $\hat{\theta}_{1,3} = \infty$ . To compute  $\hat{\theta}_{1,1}$  and  $\hat{\theta}_{1,2}$  we note from Theorem 12.2 of Gut (2005) that

$$\hat{\theta}_{x_0,k} = \sum_{n=1}^{\infty} P(G_{x_0,k}^* \geq n | X_0, \dots, X_n),$$

where  $G_{x_0,k}^*$  represents the hitting time for state  $k$  of a Markov chain with transition probability matrix  $\hat{P}$ . To compute  $P(G_{x_0,k}^* \geq n | X_0, \dots, X_n)$ , Athreya and Fuh (1992) replace the  $k^{\text{th}}$  row of the estimated transition probability matrix with that of an absorbing state. That is, the  $k^{\text{th}}$  row of the estimated transition probability matrix is replaced by  $(0, 0, \dots, 1, \dots, 0)$  where the one is in the  $k^{\text{th}}$  position. Let  $\hat{Q}$  denote this new matrix. It follows that  $P(G_{x_0,k}^* \leq n) = \hat{Q}_{x_0,k}^{(n)}$ , the  $(x_0, k)$  element of the  $n^{\text{th}}$  power of  $\hat{Q}$ . Therefore,

$$\hat{\theta}_{x_0,k} = \sum_{n=0}^{\infty} (1 - \hat{Q}_{x_0,k}^{(n)}),$$

where  $\hat{Q}_{x_0,k}^{(0)} = P(G_{x_0,k}^* \leq 0) = 0$ , which in practice will need to be approximated by a finite truncation of the infinite series. This is applied to our  $3 \times 3$  example.

An alternative method for computing these estimates is based on simulation. To compute this estimate, a simulated realization of a Markov chain with state space  $S = \{1, 2, 3\}$  and probability transition matrix  $\hat{P}$  is generated using the starting point  $X_0 = x_0$ , until the state  $k$  is observed for the first time. This process is repeated independently  $b$  times. Denote the observed length of the  $m^{\text{th}}$  simulated realization as  $G_{x_0,k}^*(m)$  for  $m = 1, \dots, b$ . Then the bootstrap estimate  $\hat{\theta}_{x_0,k} = T_{x_0,k}(\hat{P})$  can be approximated with

$$\hat{\theta}_{x_0,k} \simeq b^{-1} \sum_{m=1}^b G_{x_0,k}^*(m).$$

For this example the simulation method was used with  $b = 10000$  to obtain the bootstrap estimates  $\hat{\theta}_{1,1} = 3.21$  and  $\hat{\theta}_{1,2} = 1.43$ . The actual mean hitting times, which are computed in the same way as the estimates, only the actual transition probability matrix is used, are  $\theta_{1,1} = 2.05$ ,  $\theta_{1,2} = 2.83$  and  $\theta_{1,3} = 40.00$ . A smoothed estimate of  $P$  may increase the reliability of the estimate for  $\theta_{1,3}$ . This will be demonstrated later in this section. Note that even if state  $k$  communicates with the initial state for the Markov chain generated from the transition probability matrix  $\hat{P}$ , the bootstrap may overestimate the mean of  $G_{x_0,k}$  if the number of possible paths from the initial state to state  $k$  is much smaller in  $\hat{P}$  than in  $P$ .

A smoothed bootstrap estimate of a Markov chain parameter  $\theta = T(P)$  is  $\tilde{\theta} = T(\tilde{P})$  where  $\tilde{P}$  is a smoothed estimate of  $P$ . The typical method for smoothing an estimate of a probability distribution consists of computing a mixture distribution of an original estimate, which may not have some desired properties, with a distribution that does possess these properties. For the case of transition probability matrices we will make the assumption that  $P \in \mathcal{P}^z$ . We will construct the smoothed estimate of  $P$  to have this property as well. If the support of the conditional distribution corresponding to each row of  $P$  is nominal in structure, a discrete uniform distribution on  $S$  is used as the mixing distribution to compute the smoothed estimate. If the support of the conditional probability distribution consists of ordered categories then more sophisticated mixing distributions could be used as well.

Assuming that  $S$  has nominal structure, let  $\tilde{P}$  denote the smoothed estimate of the transition probability matrix  $P$ . Then the  $(i, j)^{\text{th}}$  element of  $\tilde{P}$  is denoted as  $\tilde{P}_{ij}$  and is defined in terms of the maximum likelihood estimate  $\hat{P}$  as

$$\tilde{P}_{ij} = \begin{cases} h_i c^{-1} + (1 - h_i) \hat{P}_{ij} & \text{if } \Omega_i > 0. \\ c^{-1} & \text{if } \Omega_i = 0. \end{cases} \quad (2.3)$$

where  $h_i \in [0, 1]$  is an adjustable mixing or smoothing parameter, for  $i = 1, \dots, c$ . Note that in this case larger values of  $h_i$  correspond to more smoothing of  $\hat{P}$ , and that if  $h_i \in (0, 1)$  then  $\tilde{P} \in \mathcal{P}^c$ . Such estimates are not new. See, for example, Aitchison and Aitken (1976), Fienberg and Holland (1973), Good (1965), Stone (1974), and Titterton (1980). Section 2.2 of Santner and Duffy (1989) provides an overview of several of these methods.

Standard methods can be implemented to choose an optimal smoothing rate, or more specifically an optimal smoothing parameter, for the estimation of the probability transition matrix. Unfortunately, these optimal smoothing parameters usually are not optimal for implementation in the bootstrap estimate. For an example of the potential difficulty in smoothing the bootstrap in the discrete case, see Guerra, Polansky and Schucany (1997). In this paper we will consider several simple smoothing parameters that attempt to preserve the asymptotic normality of the estimated transition probability matrix. Empirical studies later in the paper will provide evidence about the conditions under which the proposed methods are most reliable.

It is well known from Billingsley (1961a) that

$$n^{1/2}[\text{vec}(\hat{P}) - \text{vec}(P)] \xrightarrow{w} \mathcal{N}(\mathbf{0}, \Sigma),$$

which is a  $c^2$ -dimensional multivariate normal distribution where  $\mathbf{0}$  is a  $c^2 \times 1$  vector with each element equal to 0. The covariance matrix  $\Sigma$  is a  $c^2 \times c^2$  matrix composed of  $c \times c$  sub-matrices. Denote the  $(i, j)$ <sup>th</sup> sub-matrix as  $\Sigma_{ij}$  for  $i = 1, \dots, c$  and  $j = 1, \dots, c$ . Then  $\Sigma_{ij}$  has  $(k, l)$ <sup>th</sup> element  $\Sigma_{ij(kl)} = \delta_{ik}P_{ij}(\delta_{jl} - P_{il})$ , for  $k = 1, \dots, c$  and  $l = 1, \dots, c$ . Suppose the smoothing parameter has the form  $h_i = n^{-r}$  for  $i = 1, \dots, c$  and some  $r > 0$ . It is easy to see that  $\tilde{P}_{ij} = \hat{P}_{ij} + O(n^{-r})$  as  $n \rightarrow \infty$ . It follows that

$$n^{1/2}[\text{vec}(\tilde{P}) - \text{vec}(P)] = n^{1/2}[\text{vec}(\hat{P}) - \text{vec}(P)] + O(n^{-r+1/2}),$$

as  $n \rightarrow \infty$ . Therefore if  $r > \frac{1}{2}$  then it follows from Slutsky's Theorem that

$$n^{1/2}[\text{vec}(\tilde{P}) - \text{vec}(P)] \xrightarrow{w} \mathcal{N}(\mathbf{0}, \Sigma), \tag{2.4}$$

as  $n \rightarrow \infty$ . When  $h_1 = \dots = h_c = n^{-r}$ , the smoothed estimate  $\tilde{P}$  defined in Equation (??) globally smooths the entire transition probability matrix regardless of the number of transitions observed from each state. We will denote this smoothing parameter by  $h^g$ , and the corresponding smoothed transition probability matrix as  $\tilde{P}^g$ . A more flexible approach allows for different smoothing parameters for each row of the estimated transition probability

matrix. Define

$$h_i^l = \begin{cases} \Omega_i^{-r} & \text{if } \Omega_i > 0, \\ 1 & \text{if } \Omega_i = 0, \end{cases}$$

for  $i = 1, \dots, c$ , with the corresponding smoothed transition probability matrix being denoted as  $\tilde{P}^l$ . Theorem 1 of Athreya and Fuh (1992) implies that  $\Omega_i/n \rightarrow \pi_i$  with probability 1 as  $n \rightarrow \infty$ , where  $\pi_i$  is the stationary probability of state  $i$ , for  $i = 1, \dots, c$ . It follows that  $\Omega_i^{-r} = O(n^{-r})$  as  $n \rightarrow \infty$ , or that  $r > \frac{1}{2}$  will provide a smoothed estimator with the property given in Equation (??).

More adaptivity to the observed frequencies can be realized by using one of the many smoothing parameters based on the  $\chi^2$  statistic for testing uniformity. In particular we consider the proposal of Good (1965). In our case this proposal is equivalent to using the local smoothing parameter

$$h_i^c = \begin{cases} \min\{Z_i^{-1}, 1\} & \text{if } \Omega_i > 0, \\ 1 & \text{if } \Omega_i = 0, \end{cases}$$

where

$$Z_i = \frac{c}{(c-1)\Omega_i} \sum_{j=1}^c (\Omega_{ij} - \Omega_i/c)^2,$$

for  $i \in S$ , with the corresponding smoothed transition probability matrix being denoted as  $\tilde{P}^c$ . The asymptotic behavior of  $\tilde{P}^c$  depends on whether the  $i^{\text{th}}$  row of  $P$  follows a discrete uniform distribution. If the  $i^{\text{th}}$  row of  $P$  does not follow a discrete uniform distribution then following the development of Equation (9.6-2) of Bishop, Fienberg and Holland (1975),

$$\begin{aligned} nZ_i^{-1} &= \frac{n}{\Omega_i} \frac{(c-1)\Omega_i^2}{c} \left[ \sum_{j=1}^c \left( \Omega_{ij} - \frac{\Omega_i}{c} \right)^2 \right]^{-1} \\ &= \frac{n}{\Omega_i} \frac{c-1}{c} \left[ \sum_{j=1}^c \left( \frac{\Omega_{ij}}{\Omega_i} - P_{ij} \right)^2 + 2 \sum_{j=1}^c \left( \frac{\Omega_{ij}}{\Omega_i} - P_{ij} \right) \left( P_{ij} - \frac{1}{c} \right) + \right. \\ &\quad \left. \sum_{j=1}^c \left( P_{ij} - \frac{1}{c} \right)^2 \right]^{-1}. \end{aligned}$$

Now noting that  $\Omega_{ij}/\Omega_i \xrightarrow{P} P_{ij}$  as  $n \rightarrow \infty$  by Equation (5.4) of Billingsley (1961b), it follows that

$$\sum_{j=1}^c \left( \frac{\Omega_{ij}}{\Omega_i} - P_{ij} \right)^2 \xrightarrow{P} 0,$$

and

$$2 \sum_{j=1}^c \left( \frac{\Omega_{ij}}{\Omega_i} - P_{ij} \right) \left( P_{ij} - \frac{1}{c} \right) \xrightarrow{p} 0,$$

as  $n \rightarrow \infty$ . Since  $n/\Omega_i \rightarrow \pi_i^{-1}$  by Theorem 1 of Athreya and Fuh (1992), it follows that

$$nZ_i^{-1} \xrightarrow{p} \frac{c-1}{c\pi_i} \left[ \sum_{j=1}^c \left( \frac{1}{c} - P_{ij} \right)^2 \right]^{-1},$$

as  $n \rightarrow \infty$ , or that  $Z_i^{-1} = O_p(n^{-1})$  as  $n \rightarrow \infty$ . It follows that the asymptotic normality property given in Equation (??) holds. In the case where the  $i^{\text{th}}$  row of  $P$  follows a discrete uniform distribution the asymptotic normality result will not hold as there will be a non-zero probability that increases to 1 as  $n \rightarrow \infty$  that the difference between row  $i$  of  $\tilde{P}$  and  $P$  is exactly 0. Convergence in this case may be faster than the rate  $n^{-1/2}$ .

Returning to the example realization of Table 1, the smoothed estimate of  $P$  using the global smoothing parameter with  $r = 1$  yields  $h^g = n^{-1} = 0.033$  and a smoothed transition probability matrix equal to

$$\tilde{P}^g = \begin{bmatrix} 0.301 & 0.688 & 0.011 \\ 0.316 & 0.673 & 0.011 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}.$$

The local smoothing parameters are  $h_1^l = \Omega_1^{-1} = 0.100$ ,  $h_2^l = \Omega_2^{-1} = 0.053$ , and  $h_3^l = 1.000$ , since  $\Omega_3 = 0$ . The resulting smoothed transition probability matrix is

$$\tilde{P}^l = \begin{bmatrix} 0.303 & 0.663 & 0.033 \\ 0.317 & 0.666 & 0.017 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}.$$

The  $\chi^2$  smoothing parameters are  $h_1^c = 0.270$ ,  $h_2^c = 0.150$  and  $h_3^c = 1.000$ , since  $\Omega_3 = 0$ . The resulting smoothed transition probability matrix is

$$\tilde{P}^c = \begin{bmatrix} 0.309 & 0.601 & 0.090 \\ 0.318 & 0.631 & 0.050 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}.$$

The resulting smoothed bootstrap estimates based on each of these smooth estimates are given in Table 2. One can observe from the table that the smoothed bootstrap provides much

Table 2: Bootstrap estimates of  $\theta_{1,k}$  for  $k = 1, 2, 3$  based on the maximum likelihood estimator  $\hat{P}$  and the three smoothed estimates  $\tilde{P}^g$ ,  $\tilde{P}^l$  and  $\tilde{P}^c$  for the example realization given in Table ??.

$k$	$\theta_{1,k}$	$\hat{\theta}_{1,k} = T_{1,k}(\hat{P})$	$\tilde{\theta}_{1,k}^g = T_{1,k}(\tilde{P}^g)$	$\tilde{\theta}_{1,k}^l = T_{1,k}(\tilde{P}^l)$	$\tilde{\theta}_{1,k}^c = T_{1,k}(\tilde{P}^c)$
1	2.05	3.21	3.21	3.19	3.15
2	2.10	1.43	1.47	1.55	1.76
3	40.01	$\infty$	89.51	43.70	15.37

more reliable estimates for  $\theta_{1,3}$  and also slightly increases the reliability of the estimates of  $\theta_{1,1}$  and  $\theta_{1,2}$ . It is also apparent that the smoothing parameter may have a great influence on the reliability of the smoothed bootstrap methodology. The empirical study in the next section will provide evidence as to the conditions under which each smoothing parameter can be used reliably.

### 3 AN EMPIRICAL STUDY

To investigate the performance of the proposed smoothed bootstrap methods, a small empirical study was conducted. Simulated realizations of length  $n = 5, 10, 25$  and  $100$  were generated from a four state Markov chain with state space  $S = \{1, 2, 3, 4\}$  and transition probability matrix of the form

$$P_\gamma = \begin{bmatrix} \frac{1}{3} - \frac{\lambda}{3} & \frac{1}{3} - \frac{\lambda}{3} & \frac{1}{3} - \frac{\lambda}{3} & \lambda \\ \frac{1}{2} - \lambda & \frac{1}{2} - \lambda & \lambda & \lambda \\ 1 - 3\lambda & \lambda & \lambda & \lambda \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix},$$

for  $\lambda = 0.05$  and  $0.10$ . A total of 1000 simulated realizations were generated for each combination of  $n$  and  $\lambda$ . The starting point of each realization was taken to be  $X_0 = 1$ . For each simulated realization the mean hitting time until State 4 is visited was estimated using the estimates  $\hat{\theta}_{1,4}$ ,  $\tilde{\theta}_{1,4}^g$ ,  $\tilde{\theta}_{1,4}^l$ , and  $\tilde{\theta}_{1,4}^c$  described in the previous section. The average and standard deviation of this estimate over the 1000 simulated realizations was computed to estimate the bias and standard error of the estimate. An estimate of the root mean square error of the estimates was also computed. For the case of the estimate  $\hat{\theta}_{1,4}$ , the average and

standard deviation was computed only for the finite estimates. The number of times that  $\hat{\theta}_{1,4}$  was infinite was also computed.

The results of the simulation are given in Tables ?? and ?. The results are split up depending on whether  $\hat{\theta}_{1,4}$  is finite or not. It is reasonable to expect that all of the estimates of the mean hitting time will tend to underestimate  $\theta_{1,4}$  when  $\hat{\theta}_{1,4} < \infty$  and overestimate  $\theta_{1,4}$  when  $\hat{\theta}_{1,4} = \infty$ . This is the case for the smaller sample sizes. One can observe from Tables ?? and ?? that when  $\hat{\theta}_{1,4} < \infty$  that all three smoothed estimators have reasonable behavior when compared to  $\hat{\theta}_{1,4}$ . When  $\hat{\theta}_{1,4} = \infty$ , the global smoothed estimator  $\tilde{\theta}_{1,4}^g$  and the local smoothed estimator  $\tilde{\theta}_{1,4}^l$  tend to overestimate  $\theta_{1,4}$ , particularly with the larger sample sizes. In these cases the  $\chi^2$  smoothed estimator  $\tilde{\theta}_{1,4}^c$  tends to underestimate  $\theta_{1,4}$ . In the case where the results are combined, the local smoothed estimator tends to have the lowest bias of the three smoothed estimators, but at the cost of large variability. The  $\chi^2$  smoothed estimator, which has larger bias than the local smoothed estimator, has much lower variability, so that its estimated root mean square error is often lower than that of the local smoothed estimator. With respect to bias and mean squared error the global smoothed estimator is clearly deficient when compared to the other smooth estimators, except in the case of the smallest sample size ( $n = 5$ ). Note that all three smoothed bootstrap estimators compare well with  $\hat{\theta}_{1,4}$  when the sample size is 100 and  $\lambda = 0.10$ .

## 4 EXAMPLE

As an example we consider the multivariate binary time series data studied by Gottschau (1994). The data consist of observations from milk samples from each of a cow's four teats. The presence or absence of bacteria in each of the milk samples is recorded on time intervals of approximately three months. Gottschau (1994) considers a Markov chain model for this data as follows. Let  $FR_i$ ,  $FL_i$ ,  $RR_i$  and  $RL_i$  be binary random variables that correspond to the presence or absence of bacteria at time  $i = 0, \dots, n$  in the cow's Front Right, Front Left, Rear Right and Rear Left teats, respectively. Each variable will be assigned the value of 1 if bacteria is present and 0 if bacteria is not present. Consider the observed stochastic process  $\{Y_i\}_{i=1}^n$ , where the state of the process is defined as

$$Y_i = LR_i + 2RR_i + 4LF_i + 8RF_i, \quad (4.1)$$

for  $i = 1, \dots, n$ . As a basic model, Gottschau (1994) considers this process to follow a homogeneous Markov chain with state space  $S = \{0, 1, \dots, 15\}$ . Gottschau (1994) considers a

Table 3: Estimated mean and mean squared error for the estimate of the mean first hitting time for State 4 when  $x_0 = 1$  and  $n = 5$  and  $n = 10$ .

Estimator	Finite $\hat{\theta}_{1,4}$			Infinite $\hat{\theta}_{1,4}$			Combined	
	Number	Average	$\sqrt{\text{MSE}}$	Number	Average	$\sqrt{\text{MSE}}$	Average	$\sqrt{\text{MSE}}$
$n = 5, \lambda = 0.05, \theta_{1,4} = 19.750$								
$\hat{\theta}_k$	179	2.80	16.97	821				
$\hat{\theta}_k^g$	179	3.07	16.71	821	15.43	5.66	13.22	8.73
$\hat{\theta}_k^l$	179	3.82	15.95	821	6.03	13.75	5.63	14.17
$\hat{\theta}_k^c$	179	4.03	15.73	821	4.66	15.13	4.54	15.24
$n = 5, \lambda = 0.10, \theta_{1,4} = 9.935$								
$\hat{\theta}_k$	333	2.66	7.34	667				
$\hat{\theta}_k^g$	333	2.95	7.04	667	15.42	6.87	11.27	6.92
$\hat{\theta}_k^l$	333	3.89	6.10	667	5.95	4.12	5.26	4.87
$\hat{\theta}_k^c$	333	4.01	5.93	667	4.61	5.42	4.41	5.60
$n = 10, \lambda = 0.05, \theta_{1,4} = 19.750$								
$\hat{\theta}_k$	347	7.14	12.84	653				
$\hat{\theta}_k^g$	347	6.84	13.39	653	35.80	18.01	25.63	16.56
$\hat{\theta}_k^l$	347	5.44	14.37	653	10.98	9.03	9.05	11.18
$\hat{\theta}_k^c$	347	4.45	15.31	653	5.65	14.19	2.24	14.59
$n = 10, \lambda = 0.10, \theta_{1,4} = 9.935$								
$\hat{\theta}_k$	565	6.35	4.46	435				
$\hat{\theta}_k^g$	565	5.84	4.60	435	33.32	25.93	17.79	17.45
$\hat{\theta}_k^l$	565	5.05	5.08	435	10.55	2.77	7.45	4.23
$\hat{\theta}_k^c$	565	4.34	5.63	435	5.36	4.80	4.78	5.29

Table 4: Estimated mean and mean squared error for the estimate of the mean first hitting time for State 4 when  $x_0 = 1$  and  $n = 25$  and  $n = 100$ .

Estimator	Finite $\hat{\theta}_{1,4}$			Infinite $\hat{\theta}_{1,4}$			Combined	
	Number	Average	$\sqrt{\text{MSE}}$	Number	Average	$\sqrt{\text{MSE}}$	Average	$\sqrt{\text{MSE}}$
$n = 25, \lambda = 0.05, \theta_{1,4} = 19.750$								
$\hat{\theta}_k$	707	16.67	7.77	293				
$\hat{\theta}_k^g$	707	14.42	7.65	293	99.07	79.60	39.22	43.58
$\hat{\theta}_k^l$	707	11.04	9.32	293	30.58	11.06	16.77	9.87
$\hat{\theta}_k^c$	707	6.04	13.77	293	9.37	10.62	7.02	12.93
$n = 25, \lambda = 0.10, \theta_{1,4} = 9.935$								
$\hat{\theta}_k$	909	11.94	7.27	91				
$\hat{\theta}_k^g$	909	10.69	5.56	91	97.27	88.54	18.57	27.24
$\hat{\theta}_k^l$	909	8.71	3.71	91	29.99	20.45	10.65	7.11
$\hat{\theta}_k^c$	909	5.22	4.84	91	9.06	2.33	5.57	4.67
$n = 100, \lambda = 0.05, \theta_{1,4} = 19.750$								
$\hat{\theta}_k$	997	25.04	17.13	3				
$\hat{\theta}_k^g$	997	23.28	13.71	3	399.59	379.84	24.41	24.91
$\hat{\theta}_k^l$	997	20.47	9.51	3	131.29	111.55	20.80	11.29
$\hat{\theta}_k^c$	997	12.09	8.50	3	29.38	9.69	12.14	8.51
$n = 100, \lambda = 0.10, \theta_{1,4} = 9.935$								
$\hat{\theta}_k$	1000	11.22	5.07					
$\hat{\theta}_k^g$	1000	10.97	4.68					
$\hat{\theta}_k^l$	1000	10.45	3.98					
$\hat{\theta}_k^c$	1000	7.09	3.48					

Table 5: Data from 13 observations of a single cow from Figure 1 of Gottschau (1994).

$i$	$LR_i$	$RR_i$	$LF_i$	$RF_i$	$Y_i$
0	1	0	0	0	1
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	1	8
5	0	0	0	1	8
6	0	0	0	1	8
7	0	0	0	0	0
8	0	0	0	1	8
9	0	1	0	0	2
10	0	1	0	1	10
11	1	0	0	0	1
12	0	0	0	0	0

large sample of cattle which makes the estimation of the parameters of the transition probability matrix for this process easy to estimate without smoothing. However, if one wishes to estimate parameters of this model for a single cow, or for a small cohort of cattle, the large number of parameters in the transition probability matrix will make this difficult.

As an example, consider the data from a single cow as given in Figure 1 of Gottschau (1994). These data, along with the corresponding values of the stochastic process given in Equation (??) are given in Table ???. Suppose we assume that all of the transition probabilities for this process are non-zero. Using the Markov model for this data, we would like to estimate the mean number of transitions from each of the non-zero states (infection) to the zero state (no infection). Gottschau (1994) includes the maximum likelihood estimator of the transition probability matrix for this process based on a study of 1233 cattle. We will treat this probability transition matrix as the true matrix in order to evaluate the quality of our estimates. See Table 2 of Gottschau (1994) for this estimated transition probability matrix.

Table 6: Estimates of the number of transitions from State  $k$  to State 0 based on the maximum likelihood estimator and the locally smoothed estimator of the transition probability matrix. The population value is computed using the transition probability matrix from Table 2 of Gottschau (1994).

$k$	$\theta_{k,0}$	$\hat{\theta}_{k,0}$	$\tilde{\theta}_{k,0}^g$	$\tilde{\theta}_{k,0}^l$	$\tilde{\theta}_{k,0}^c$
1	3.79	1.00	1.34	5.42	5.87
2	3.91	3.00	3.75	9.83	10.68
3	4.32	$\infty$	5.55	9.80	10.66
4	3.84	$\infty$	5.57	9.83	10.68
5	4.62	$\infty$	5.56	9.83	10.69
6	4.78	$\infty$	5.57	9.87	10.66
7	5.13	$\infty$	5.56	9.84	10.65
8	3.81	3.56	4.12	8.09	10.20
9	4.52	$\infty$	5.56	9.86	10.69
10	4.76	$\infty$	2.59	9.83	10.69
11	4.95	$\infty$	5.55	9.87	10.70
12	4.45	$\infty$	5.55	9.86	10.70
13	4.65	$\infty$	5.54	9.80	10.68
14	4.75	$\infty$	5.56	9.81	10.67
15	5.00	$\infty$	5.58	9.83	10.61

Estimates of  $\theta_{k,0}$  for  $k = 1, \dots, 15$  were computed using the methodology of the previous section. The parameter  $\theta_{k,0}$  was computed using the transition probability matrix from Table 2 of Gottschau (1994). Table ?? summarizes these estimates. One can observe from Table ?? that the usual estimator  $\hat{\theta}_{k,0}$  is infinite in all but three of the cases considered. The smoothed estimates alleviate this problem. The most reliable smoothed estimate in this example is the global estimator  $\tilde{\theta}_{k,0}^g$ . This is consistent with the results of the empirical study of the previous section which suggested that the global estimator is most reliable when the sample size was very small compared to the number of parameters in the transition probability matrix.

## 5 DISCUSSION

This paper presents a simple method for computing a smoothed estimate of a transition probability matrix based on an observed realization from a Markov chain. It was demonstrated that the use of the smoothed estimate can improve the performance of bootstrap estimates of certain parameters of the Markov chain. An empirical study suggested under what conditions the methods were reliable. An example was presented where the standard estimate of the mean hitting time of some states of the Markov chain gave an infinite answer, while the smoothed estimates gave finite reasonable answers.

Alternative smoothing methods with more refined smoothing parameters may increase the reliability of these estimates. However, it is likely that each different situation would warrant different optimal methods. A key idea of the methods presented in this paper are their simplicity, though further study may be required to insure that the proposed methods perform reasonably in a variety of situations. One possible solution may involve the use of the calibrated bootstrap of Loh (1987) as a method of smoothing parameter selection.

The mean hitting time for a state was used as an example parameter for purposes of exposition. The methods proposed here can be easily extended to include any other parameter of a Markov chain by simply changing the functional  $T$  in the appropriate calculations.

## ACKNOWLEDGMENTS

The author would like to thank Professor S. Ejaz Ahmed for the kind invitation to contribute to this issue, and the referee whose comments greatly improved the presentation of the paper.

## REFERENCES

1. Aitchison, J., Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.
2. Athreya, K. B., Fuh, C. D. (1992). Bootstrapping Markov chains. In: *Exploring the Limits of Bootstrap*. Eds: R. LePage and L. Billard. New York: John Wiley and Sons. pp. 49–64.
3. Bartlett, M. S. (1951). The frequency goodness of fit test for probability chains. *Proceedings of the Cambridge Philosophical Society* **47**, 89–110.

4. Basawa, I. V., Green, T. A., McCormick, W. P., Taylor, R. L. (1990). Asymptotic bootstrap validity for finite Markov chains. *Communications in Statistics – Theory and Methods*, **19**, 1493–1510.
5. Basawa, I. V., Rao, B. L. S. P. (1980). *Statistical Inference for Stochastic Processes*. New York: Academic Press.
6. Billingsley, P. (1961a). Statistical methods in Markov chains. *Annals of Mathematical Statistics*, **32**, 12–40.
7. Billingsley, P. (1961b). *Statistical Inference for Markov Processes*. Chicago, IL: The University of Chicago Press.
8. Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.
9. Clark, W. A. V. (1965). Markov chain analysis in Geography: An application to the movement of rental housing areas. *Annals of the Association of American Geographers*, **55**, 351–359.
10. Datta, S., McCormick, W. P. (1992). Bootstrap for a finite state Markov chain based on I.I.D. resampling. *Exploring the Limits of Bootstrap*. R. LePage and L. Billard, Editors. New York: John Wiley and Sons. pp. 77–97.
11. Fienberg, S. E., Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**, 683–691.
12. Fuh, C. D. (1993). Statistical inquiry for Markov chains by bootstrap method. *Statistica Sinica*, **3**, 53–66.
13. Geary, K. (1978). Indicators of educational progress - a Markov chain approach applied to Swaziland. *Journal of Modern African Studies*, **16**, 141–151.
14. Good, I. J. (1965). *The Estimation of Probabilities*. Cambridge MA.: M.I.T. Press.
15. Gottschau, A. (1992). Exchangeability in multivariate Markov chain models. *Biometrika*, **48**, 751–763.
16. Gottschau, A. (1994). Markov chain models for multivariate binary panel data. *Scandinavian Journal of Statistics*, **21**, 57–71.

17. Guerra, R., Polansky, A. M., Schucany, W. R. (1997). Smoothed bootstrap confidence intervals with discrete data. *Computational Statistics and Data Analysis*, **26**, 163–176.
18. Gut, A. (2005). *Probability: A Graduate Course*. Springer, New York.
19. Karlin, S., Taylor, H. M. (1975). *A First Course in Stochastic Processes*. San Diego, CA.: Academic Press.
20. Kulperger, R. J., Rao B. L. S. P. (1989). Bootstrapping a finite state Markov chain. *Sankhya, Series A, Indian Journal of Statistics*, **51**, 178–191.
21. Loh, W.-Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, **82**, 155–162.
22. Santner, T. J., Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer.
23. Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–515.
24. Titterton, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.
25. Turchin, P. B. (1986). Modelling the effect of host patch size on Mexican bean beetle emigration. *Ecology*, **67**, 24–132.
26. Yang, M. C. K. (1979). Some results on size and utilization distribution of home range. *Statistical Distributions in Ecological Work, Volume 4*. J. K. Ord, G. P. Patil and C. Taillie, editors. Fairland, Maryland: International Co-operative Publishing House. 429–449.