Statistica Sinica **19** (2009), 177-196

ESTIMATING THE POPULATION SIZE FOR A MULTIPLE LIST PROBLEM WITH AN OPEN POPULATION

Huazhen Lin, Paul S. F. Yip and Feng Chen

Sichuan University, University of Hong Kong and University of New South Wales

Abstract: A semiparametric method using a local polynomial is proposed to estimate the population size at a specific time from multiple lists of an open population. The asymptotic distribution for the proposed estimators is derived. Simulation studies show that the proposed procedure works much better than existing methods. In addition, we provide a simple and efficient method to deal with the variable selection problem in a log-linear model when the number of the lists is large. The method is applied to estimate the number of drug-abusers in Hong Kong over the period 1977-1997.

Key words and phrases: Drug abusers, local polynomial, multiple-list, open population.

1. Introduction

This work is motivated by estimating the number of drug users in Hong Kong for the period 1977-1997. A Central Registry of Drug Abuse (CRDA) was established in the Narcotics Division of the Hong Kong Government to monitor the number of drug-abusers in Hong Kong. Reports on known or suspected drug-abusers are compiled by different agencies and submitted to the CRDA on a standard record sheet on a semi-annual basis. Among the agencies, there are four major lists: Police Department, Correctional Services Department, Social Welfare Department, and Hospitals.

Table 1 gives the four-list presence-absence data for each half-year from 1977 to 1997, where each pattern of 1's and 0's represents being recorded or not being recorded in a particular list. For example, "1111" denotes a case listed in all the lists, 1101, a case listed by the Police, the Correctional Services Department, and the Hospitals, but not by the Social Welfare Department, etc. There are $2^4 - 1 = 15$ entries for each half-year, and 42 contingency tables for the twenty-one years. At present, the Narcotic Bureau simply adds up the distinct individuals among all the lists to form an estimate of the number of drug abusers in Hong Kong. Certainly, there are individuals who were not "captured" by any of the lists. Our interest is to estimate the number of drug abusers for

Table 1. The observed numbers of drug users in Hong Kong for each half-year recorded during 1977-1997. The Four lists are Police Department, Correctional Services Department, Social Warefare Department, and Hospitals. A — first half-year; B — second half-year. Full table available as an online supplementary document.

year	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
77A	1575	3325	556	2199	184	213	53	1264	127	115	31	1096	151	120	29
77B	1199	3398	527	2053	169	290	77	1154	121	178	50	824	126	116	37
78A	1415	3444	678	1945	155	254	64	1044	111	125	46	760	164	84	44
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
97B	1145	3314	371	2300	94	488	76	1498	57	301	30	908	47	246	26

the 42 periods, or the number of missing drug abusers for each time period. The pattern of overlapping among the lists provides useful information on the number of "uncaptured" individuals (IWGDMF (1995a,b)).

There are two complications with the drug-user data. First, at different time points, the lists relate to different but overlapping populations; second, the information of each individual being captured among the lists over the period was not available. The first feature implies that it is an open population problem. The second feature implies that our estimators must be based on the marginal distributions.

In the literature, many methods have been proposed to handle the multiple-list for a closed population; for example, the Poisson log-linear model (Fienberg (1972), Cormack (1989) and IWGDMF (1995a,b)), the multinomial model (Cormack and Jupp (1991)) and the sample coverage method (Chao and Lee (1992)). There also exist methods to deal with the open population problem; for example, Huggins and Yip (2003), Huggins, Yang, Chao and Yip (2003), Yang and Huggins (2003) and Yang, Huggins and Clark (2003). However, estimation of an open population size for multiple-list experiments has yet to be developed.

One obvious method to handle the open population with multiple-list problem is to regard each unit as a closed population and to estimate the population size in each unit based on the observed individuals in that unit, using the Poisson log-linear model or the multinomial model. We refer to this as the simple imputation (SI) method. Since the SI estimate of the population size for each period is based solely on the corresponding contingency table for that period, the results may suffer from considerable variability and thus be unreliable. For example, the dotted line in Figure 6.3 reports the SI estimates of the number of drug abusers in Hong Kong. These estimates are seen to be very unstable, with variations during half-years as large as 40,000, as against baseline figures of 20,000. To assess the trend in the numbers, stable estimates of the population sizes are needed.

One purpose of this paper is to provide a stable estimate of the population sizes at different time points. Our idea is based on the following observation. The population is subject to change, but the characteristics of the population usually vary slowly over time, so observations from adjacent units carry useful information on the current population size. Thus it is possible to improve on the SI estimator by including data from adjacent units. In this paper, we achieve this by combining the local polynomial technique (Fan and Gijbels (1996)) and the log-linear model. Compared with the SI estimator, our estimator has a much smaller variance and a slightly larger bias, hence a considerably smaller mean squared error.

A problem with the log-linear model is that the number of variables is $K = 2^k - 1$, where k is the number of lists, and hence the model may be overly complex when the number of lists is moderately large. Another purpose of this paper is to address the issue of variable selection. The classical approach to model selection, such as the likelihood ratio test, may be sensitive to null parameters and thus the choice of the prespecified model. Another method for variable selection is the subset variable selection method; it compares all possible models using some information criterion such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), and selects the best one. This method can become impractical in our situation because the computational effort required for exhaustively searching over $2^k - 1$ models can be expensive when k is only moderately large. In this paper, extending the penalized likelihood estimator with the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001, 2002), and later extended by Fan, Lin and Zhou (2006) to the nonparametric setting, we propose a model selection method that requires the computations of only one model and a tuning parameter. The computational effort of selecting the tuning parameter does not increase with the number of lists. Another advantage of the SCAD penalized method over existing methods is that the true regression coefficients that are zero are automatically estimated as zero, with the remaining coefficients being estimated as if the correct submodel were known in advance. Hence, the SCAD method is not sensitive to the choice of the prespecified model.

This paper is organized as follows. Section 2 describes the model and the estimation method. The proposed estimator is shown to be asymptotically normal. In Section 2 we also discuss the estimation of variance and bandwidth selection. Based on the SCAD penalty, Section 3 provides a model selection method for the log-linear model. Section 4 presents a simulation study in which we compare the proposed method with the SI method, and investigate the robustness of the proposed method. We apply our method by applying it to the Hong Kong drug-abuser data in Section 5. A discussion is given in Section 6.

2. Model and Estimation Methods

2.1. Notation and model

The time axis $[0, \tau]$ is divided into subintervals of equal length, labeled $j = 1, \ldots, n$. Suppose we have k lists. For each time unit j, a natural way of recording the information obtained by matching the lists is in the form of an incomplete 2^k contingency table, with one margin per list. Each cell of the table corresponds to a subset of lists and contains the count of individuals that are recorded on exactly that subset of lists. The cell count corresponding to individuals on none of the lists is missing, and this is the parameter of interest. In this way, we obtain a set of contingency tables and each contingency table corresponds to a time point. Let n_{cj} be the observed cell count for cell c at time j, for $c = 1, \ldots, K$, $j = 1, \ldots, n$, where $K = 2^k - 1$, and n_{0j} be the unobserved count. Let p_c denote the probability of an individual being in the cell c and p_0 the probability of being in the empty cell. Let $\Lambda_j = E(\sum_{c=0}^{K} n_{cj})$. The main problem is to estimate Λ_j for $j = 1, \ldots, n$.

We suppose that the cell counts n_{cj}, $j = 1, \ldots, n$, $c = 0, \ldots, K$, are independently Poisson distributed with means $\Lambda_j p_c$. Following the literature (Cormack (1989) and Fienberg (1972)), we use the log-linear model to reparameterize the cell probabilities. For example, the saturated model for a 3-list experiment is given by

$$\begin{aligned}
&\log p_{111} = \theta_0,\\
&\log p_{011} = \theta_0 + \theta_1, \ \ \log p_{101} = \theta_0 + \theta_2, \ \ \log p_{110} = \theta_0 + \theta_3,\\
&\log p_{001} = \theta_0 + \theta_1 + \theta_2 + \theta_{12}, \ \ \log p_{010} = \theta_0 + \theta_1 + \theta_3 + \theta_{13},\\
&\log p_{100} = \theta_0 + \theta_2 + \theta_3 + \theta_{23},\\
&\log p_{000} = \theta_0 + \theta_1 + \theta_2 + \theta_{12} + \theta_3 + \theta_{13} + \theta_{23} + \theta_{123},
\end{aligned} \tag{2.1}$$

where p_{111} denotes the probability of an individual being in cell "111", and so on. The models for general k-list experiments are given similarly. The parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k, \ldots, \theta_{12\cdots k})^T$ represent the baseline effect, main effects, and interactions effects among the lists. For identifiability considerations, the highest order interaction $\theta_{12\cdots k}$ among the lists is always set to 0.

2.2. Estimation

For ease of presentation, we temporarily assume that the parameters Λ_j are reparameterized by $\Lambda_j(\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is a set of unknown parameters. Our model assumptions imply that $n_j = \sum_{c=1}^{K} n_{cj}$ is Poisson distributed with mean $(1 - p_0(\boldsymbol{\theta}))\Lambda_j(\boldsymbol{\beta})$ and that, given n_j, the frequency counts $n_{1j}, \ldots, n_{Kj}$ follow a multinomial distribution based on n_j trials with cell probabilities $q_c(\boldsymbol{\theta})$, $c = 1, \ldots, K$, where $q_c(\boldsymbol{\theta}) = p_c(\boldsymbol{\theta})/(1 - p_0(\boldsymbol{\theta}))$. The joint distribution of n_{cj}, $c = 1, \ldots K$, $j = 1, \ldots, n$ is then

$$\prod_{j=1}^{n} \frac{[\Lambda_j(\boldsymbol{\beta})(1 - p_0(\boldsymbol{\theta}))]^{n_j} e^{-\Lambda_j(\boldsymbol{\beta})(1-p_0(\boldsymbol{\theta}))}}{n_j!} \times \prod_{j=1}^{n} \frac{n_j! \prod_{c=1}^{K} q_c(\boldsymbol{\theta})^{n_{cj}}}{\prod_{c=1}^{K} n_{cj}!}, \tag{2.2}$$

and the log-likelihood of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ based on the observed cell counts n_{cj}, $c = 1, \ldots, K$, $j = 1, \ldots, n$, can be decomposed as $L(\boldsymbol{\theta}, \boldsymbol{\beta}) = L_1(\boldsymbol{\theta}) + L_2(\boldsymbol{\beta}, p_0(\boldsymbol{\theta}))$, where

$$L_1(\boldsymbol{\theta}) = \sum_{j=1}^{n} \sum_{c=1}^{K} n_{cj} \log(q_c(\boldsymbol{\theta})) + k_1, \tag{2.3}$$

$$L_2(\boldsymbol{\beta}, p_0(\boldsymbol{\theta})) = \sum_{j=1}^{n} \left\{ n_j \log \Big[(1-p_0(\boldsymbol{\theta}))\Lambda_j(\boldsymbol{\beta}) \Big] - [1-p_0(\boldsymbol{\theta})]\Lambda_j(\boldsymbol{\beta}) \right\} + k_2, \tag{2.4}$$

and k_1, k_2 are independent of the parameters. Thus $L(\boldsymbol{\theta}, \boldsymbol{\beta})$ is a sum of the conditional log-likelihood $L_1(\boldsymbol{\theta})$ of $n_{1j}, \ldots, n_{Kj}$ given n_j and the marginal likelihood $L_2(\boldsymbol{\beta}, p_0(\boldsymbol{\theta}))$ of n_j. Statistical inference on $\boldsymbol{\theta}$ can be based on the conditional likelihood $L_1(\boldsymbol{\theta})$ or the unconditional likelihood $L(\boldsymbol{\theta}, \boldsymbol{\beta})$. The comparison of inference based on conditional and unconditional likelihoods is a topic which has been discussed extensively (Fienberg (1972), Sanathanan (1972), Sandland and Cormack (1984), Cormack and Jupp (1991) and Yip (1991)). Under suitable regularity conditions, the estimate, and inference on $\boldsymbol{\theta}$, based on the conditional and unconditional likelihood are asymptotically equivalent (Sanathanan (1972)). In this paper, we estimate $\boldsymbol{\theta}$ by maximizing $L_1(\boldsymbol{\theta})$ and denote the estimator by $\hat{\boldsymbol{\theta}}$. The likelihood function $L_1(\boldsymbol{\theta})$ does not include any information on $\boldsymbol{\beta}$, so statistical inference on $\boldsymbol{\beta}$ is based on $L_2(\boldsymbol{\beta}, p_0(\boldsymbol{\theta}))$. Maximization of $L_2(\boldsymbol{\beta}, p_0(\hat{\boldsymbol{\theta}}))$ with respect to $\boldsymbol{\beta}$ leads to a ML estimator of $\boldsymbol{\beta}$. A computationally appealing feature of this two stage method is that, instead of solving a semiparametric optimization problem, we only need to do separately parametric and nonparametric optimizations, which can significantly reduce the computational complexity.

Consider the estimate of Λ_j. Since the subpopulation sizes are unobservable, it may be difficult to correctly specify the parametric form for $\Lambda_j(\boldsymbol{\beta})$. One obvious way is to take $\boldsymbol{\beta} = (\Lambda_1, \ldots, \Lambda_n)^T$ and treat Λ_j, $j = 1, \ldots, n$, as n independent parameters. This fully nonparametric method leads to the SI estimates $\hat{\Lambda}_j = n_j/(1 - p_0(\hat{\boldsymbol{\theta}}))$, $j = 1, \ldots, n$. As mentioned in the introduction, the SI estimates suffer from unacceptably large variation.

To improve the accuracy of the estimates of the Λ_j, we assume the Λ_j can be embedded into a smooth function Λ via $\Lambda_j = \Lambda(t_j)$ with $t_j = j\delta_n$ and $\delta_n = \tau/n$. Let $\lambda(t)$ be the population-normalized intensity of the drug-user process at time $t \in [0, \tau]$, and η be the size of the underlying population, so that we can write $\Lambda_j = \eta \int_{t_j-\delta_n}^{t_j} \lambda(s)ds$ and take $\Lambda(t)$ to be $\eta \int_{t-\delta_n}^{t} \lambda(s)ds$.

Since Λ is differentiable, for any fixed t_0 and each t close to t_0, a Taylor expansion gives,

$$\Lambda(t) \approx \Lambda(t_0) + \Lambda'(t_0)(t - t_0) = \beta_1 + \beta_2(t - t_0). \tag{2.5}$$

Inserting $\Lambda_j = \beta_1 + \beta_2(t_j - t_0)$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ into (2.4) and introducing a kernel function K with a bandwidth h, we obtain the local log-likelihood for $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ as

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{n} \Big\{ n_j \log\Big((1 - p_0(\hat{\boldsymbol{\theta}}))[\beta_1 + \beta_2(t_j - t_0)] \Big) \\ -(1 - p_0(\hat{\boldsymbol{\theta}}))[\beta_1 + \beta_2(t_j - t_0)] \Big\} K_h(t_j - t_0), \tag{2.6}$$

where $K_h(u) = K(u/h)/h$. The kernel and the bandwidth are introduced to ensure that essentially only those data near t_0 are used to estimate $\Lambda(t_0)$. Note that β_1 and β_2 are dependent on t_0, and so is $\ell(\boldsymbol{\beta})$. Maximizing $\ell(\boldsymbol{\beta})$ gives the estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$, and $\Lambda(t_0)$ is estimated by $\hat{\Lambda}(t_0) = \hat{\beta}_1$. The whole curve $\hat{\Lambda}(\cdot)$ is obtained by running the above local linear procedures with t_0 varying in $[0, \tau]$. The Λ_j are estimated by $\hat{\Lambda}_j = \hat{\Lambda}(t_j)$, $j = 1, \ldots, n$. We refer to these as local linear ML estimates.

2.3. Asymptotic properties

In this subsection, we state the results concerning the asymptotic normality of the proposed estimators. The proof of the results can be found in the appendix. The asymptotics are developed as the underlying population size η goes to infinity. The notation "$\xrightarrow{d}$" means "converges in distribution to". Before considering asymptotic properties of $\hat{\boldsymbol{\theta}}$, we first note that with the restriction $\sum p_c(\boldsymbol{\theta}) = 1$, the number of free θ's in (2.1) is $K - 1$. Remove any component, say θ_0, from $\boldsymbol{\theta}$ and, by a slight misuse of notation, we use $\boldsymbol{\theta}$ to denote the free parameters remaining, with parameter space the $(K-1)$-dimensional Euclidean space $\mathbb{R}^{K-1}$.

Theorem 1. *Let $\hat{\boldsymbol{\theta}}$ be the conditional MLE of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^0$ be the true value of $\boldsymbol{\theta}$, and $\mathbf{I}(\boldsymbol{\theta}) = (i_{rs}(\boldsymbol{\theta}))_{(K-1)\times(K-1)}$ be the matrix defined by*

$$i_{rs}(\boldsymbol{\theta}) = \sum_{c=1}^{K} \frac{1}{q_c(\boldsymbol{\theta})} \frac{\partial q_c(\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial q_c(\boldsymbol{\theta})}{\partial \theta_s}, \quad r, s = 1, \ldots, K-1.$$

with θ_j, $j = 1, \ldots, K-1$, being the jth component of $\boldsymbol{\theta}$. Then $\mathbf{I}(\boldsymbol{\theta}^0)$ is nonsingular, and if $\int_0^\tau \lambda(t)dt > 0$,

$$\eta^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \Sigma) \quad \text{as} \quad \eta \to \infty, \tag{2.7}$$

where $\Sigma = [\int_0^\tau \lambda(t)dt]^{-1/2}[\mathbf{I}(\boldsymbol{\theta}^0)]^{-1}$.

Theorem 2. *For any $t_0 \in (0, \tau)$, let $\hat{\Lambda}(t_0)$ be the local linear ML estimator of $\Lambda(t)$. Assume that $\lambda(\cdot)$ is positive and twice continuously differentiable at t_0 and that the kernel $K(\cdot)$ is a symmetric probability density with bounded support. Assume $h \to 0$, $n \to \infty$, $nh \to \infty$ and $\eta/n \to C \in (0, \infty)$ as $\eta \to \infty$. Then, as $\eta \to \infty$,*

$$(nh)^{\frac{1}{2}}\left(\hat{\Lambda}(t_0) - \Lambda(t_0) - \frac{1}{2}h^2 C\tau\lambda''(t_0)u_2\right) \xrightarrow{d} \mathbf{N}\left(0, \frac{v_0 C\tau^2\lambda(t_0)}{1 - p_0(\boldsymbol{\theta}^0)}\right), \tag{2.8}$$

where $u_2 = \int_{-\infty}^{\infty} x^2 K(x)dx$ and $v_0 = \int_{-\infty}^{\infty} K^2(x)dx$.

Remark 1. The condition $\int_0^\tau \lambda(t)dt > 0$ assumed by Theorem 1 guarantees that when $\eta \to \infty$, the number of observed individuals will also tend to ∞ (in probability). This condition is clearly fulfilled when $\lambda(\cdot)$ is positive and continuous at any point $t_0 \in (0, \tau)$.

Remark 2. The condition $\eta/n \to C \in (0, \infty)$ means that when the underlying population size grows to infinity, the number of observation intervals grows to infinity at the same rate. One implication of this assumption is that when η (or n) is large, $\Lambda(t) = \eta \int_{t-\delta_n}^{t} \lambda(s)ds \approx C\tau \int_{t-\delta_n}^{t} \lambda(s)ds/\delta_n \approx C\tau\lambda(t)$, so that the $\hat{\Lambda}_j$ are virtually the estimates of the intensity $\lambda(\cdot)$ of the drug user process at the end points of the observation intervals (up to a constant $C\tau$).

Remark 3. The $\hat{\boldsymbol{\theta}}$ in (2.6) is an estimate of the parameter $\boldsymbol{\theta}$. Therefore, the variance of $\hat{\Lambda}(t_0)$ should in general contain an extra term reflecting the uncertainty introduced by replacing $\boldsymbol{\theta}$ by an estimator. However, from Theorem 2 and its proof, the variance of $\hat{\Lambda}(t_0)$ performs as well as if $\boldsymbol{\theta}$ were known. This occurs due to the fact that the rate of convergence of $\hat{\boldsymbol{\theta}}$ is faster than that of $\hat{\Lambda}(t_0)$, so that the uncertainty from $\hat{\boldsymbol{\theta}}$ can be ignored.

2.4. Variance estimation and the bandwidth selection

It can be shown that the covariance matrix of $(\hat{\beta}_1(t_0) - \beta_1(t_0), h(\hat{\beta}_2(t_0)) - \beta_2(t_0))^T$ can be estimated by

$$(nh)^{-1}(\hat{A}_n)^{-1}\widehat{\Sigma}_n(\hat{A}_n)^{-1}, \tag{2.9}$$

where,

$$\hat{A}_n = \frac{1}{n}\sum_{j=1}^{n} \frac{n_j x_{t_j} x_{t_j}^T}{\Big(\hat{\beta}_1(t_0) + \hat{\beta}_2(t_0)(t_j - t_0)\Big)^2} K_h(t_j - t_0),$$

$$\widehat{\Sigma}_n = \frac{h}{n}\sum_{j=1}^{n} \left(\frac{n_j}{\hat{\beta}_1(t_0) + \hat{\beta}_2(t_0)(t_j - t_0)} - (1 - p_0(\hat{\boldsymbol{\theta}}))\right)^2 x_{t_j} x_{t_j}^T K_h^2(t_j - t_0),$$

in which $x_t = (1, (t - t_0)/h)^T$. The variance of $\hat{\Lambda}(t_0)$ is estimated by entry $(1, 1)$ of matrix (2.9).

Our simulation shows that when p_0 is larger, the variance formula (2.9) tends to underestimate the true variance. This might be due to the fact that when $1 - p_0$ is close to 0, the number of listed individuals will be too small for $p_0(\hat{\boldsymbol{\theta}})$ to estimate p_0 reliably, so that the variance of $\hat{\boldsymbol{\theta}}$ will make a practically nonnegligible contribution to the variance of $\hat{\Lambda}(t_0)$. For this reason, a parametric bootstrap procedure is recommended to estimate the variance of $\hat{\Lambda}_j$, $j = 1, \ldots, n$. The procedure is as follows. Get estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\Lambda}_j$, $j = 1, \ldots, n$ using the proposed method; generate a set of B i.i.d. bootstrap samples where each bootstrap sample $\{n_{cj}^*;\ c = 1, \ldots K,\ j = 1, \ldots, n\}$ is generated according to the independent distributions $n_{cj}^* \sim \text{Pois}(\hat{\Lambda}_j p_c(\hat{\boldsymbol{\theta}}))$; for each bootstrap sample, calculate the estimates of the Λ_j, $j = 1, \ldots, n$. So for each j, we have B bootstrap samples for $\hat{\Lambda}_j$, $\{\hat{\Lambda}_j^{*(1)}, \ldots, \hat{\Lambda}_j^{*(B)}\}$, and we take their variance $(B/(B-1))[\sum_{i=1}^{B}(\hat{\Lambda}_j^{*(B)})^2/B - (\sum_{i=1}^{B}\hat{\Lambda}_j^{*(B)}/B)^2]$ as an estimate of the variance of $\hat{\Lambda}_j$. Our numerical study shows this procedure works reasonably well when $B \geq 100$.

One issue in the use of the local linear method described in this paper is the selection of the bandwidth h which trades off variance and bias. Although a suitable bandwidth for $\hat{\Lambda}(t)$ might be selected subjectively by visually examining the fitted curves with different bandwidths, an automatic bandwidth selection procedure based on the data is useful to provide an indication of a suitable bandwidth range. Here, since the estimate of Λ_j is of interest, we take the bandwidth that minimizes the Cumulated Mean Squared Error (CMSE) given by

$$CMSE = \sum_{j=1}^{n}\left\{Bias^2[\hat{\Lambda}_j] + Var[\hat{\Lambda}_j]\right\} w_j.$$

The estimation of the variance of $\hat{\Lambda}_j$ can be obtained from (2.9), or by the bootstrap method as explained in the preceding paragraph. However, the biases of nonparametric estimates are generally difficult to estimate, since they involve higher order derivatives (see Theorem 2). For this reason, we use the empirical bias method proposed by Ruppert (1997), which we find to work well in our simulations and in our example.

3. Variable Selection

The number of variables in the log-linear model is $2^k - 1$, where k is the number of lists. Hence, the variable selection is an important problem when k is only moderately large. Fan and Li (2001, 2002) proposed a family of new variable selection methods based on a nonconcave penalized likelihood. Their methods are different from traditional ones in that they delete insignificant variables by estimating their coefficients as 0, and simultaneously select significant variables and estimate regression coefficients. LASSO, proposed by Tibshirani (1996), is a member of this family with an L_1 penalty. From their simulations, Fan and Li (2001) showed that the penalized likelihood estimator with smoothly clipped absolute deviation (SCAD) penalty outperforms the best subset variable selection in terms of computational cost and stability. In addition, they showed that SCAD improves the LASSO in terms of estimation biases. Furthermore, they demonstrated that with a proper choice of tuning parameter and penalty functions, for example, SCAD, the penalized likelihood estimator possesses an oracle property. Motivated by the work of Fan and Li (2001, 2002), we select variables and estimate coefficients simultaneously by maximizing the penalized log likelihood function

$$Q(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}) - N \sum_{j=1}^{K} p_{\varrho}(|\theta_j|), \tag{3.1}$$

where $N = \sum_{j=1}^{n} n_j$, $p_{\varrho}(\cdot)$ is a penalty function, ϱ is a tuning parameter, and θ_j is the jth component of $\boldsymbol{\theta}$. The coefficients of redundant variables are estimated as 0 automatically, with probability tending to 1, and the non-zero components are estimated as well as in the case where the correct submodel is known; hence, the objectives of variable selection and coefficients estimation are simultaneously achieved by maximizing (3.1). Using local quadratic approximations (Fan and Li (2001)) and the Newton-Raphson algorithm, we can maximize (3.1) by iterating the following equation untill convergence,

$$\begin{aligned}\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - &\left\{ \frac{\partial^2 L_1(\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - N\Sigma_{\varrho}(\boldsymbol{\theta}^{(k)}) \right\}^{-1} \\ &\times \left\{ \frac{\partial L_1(\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}} - N\Sigma_{\varrho}(\boldsymbol{\theta}^{(k)})\boldsymbol{\theta}^{(k)} \right\},\end{aligned} \tag{3.2}$$

where $\Sigma_\varrho(\boldsymbol{\theta}) = \mathrm{diag}\{p'_\varrho(|\theta_1|)/|\theta_1|, \ldots, p'_\varrho(|\theta_K|)/|\theta_K|\}$. The estimator from the full model can be used as starting value $\boldsymbol{\theta}^{(0)}$ of the iteration.

A good penalty function $p_\varrho(\cdot)$ should result in an estimator with the following three properties: unbiasedness for large coefficients to attenuate overall bias, sparsity (many small coefficients are estimated as zero) to reduce model complexity, and continuity to avoid unnecessary variation in model prediction. Necessary conditions for unbiasedness, sparsity and continuity have been derived by Antoniadis and Fan (2001) and Fan and Li (2001). A simple penalty function that satisfies the three requirements is the SCAD penalty,

$$p'_\varrho(\theta) = \varrho\left\{I(\theta \le \varrho) + \frac{(a\varrho - \theta)_+}{(a-1)\varrho} I(\theta > \varrho)\right\} \quad \text{for some } a > 2 \text{ and } \theta > 0, \qquad (3.3)$$

that involves two unknown tuning parameters ϱ and a. In practice, we can search for the best pair (ϱ, a) over a two-dimensional grid using some suitable criterion, such as AIC or BIC or cross-validation. Using Bayesian risk analysis tools with a normal prior distribution for θ, Fan and Li (2001) found that $a \approx 3.7$ is an appropriate choice in a wide variety of situations. This value will be used in our numerical example.

Let the true value of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ be $\boldsymbol{\theta}^0 = (\theta_{10}, \ldots, \theta_{K0})^T = (\boldsymbol{\theta}_{10}^T, \boldsymbol{\theta}_{20}^T)^T$. Without loss of generality, assume that $\boldsymbol{\theta}_{20} = 0$. Following Fan and Li (2001), we can prove that if SCAD penalty is taken and $\varrho \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of $Q(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_p(\eta^{-1/2})$. Hence by choosing a proper ϱ, there exists a $\sqrt{\eta}$ consistent penalized maximum likelihood estimator. Furthermore, again using the method of Fan and Li (2001), we can prove that this estimator must possess (1) the sparsity property, and (2) the oracle property, i.e., the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$ is the same as the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$ when $\boldsymbol{\theta}_2 = 0$. In other words, when the true parameters have some zero components, they are estimated as 0 with probability tending to 1, and the non-zero components are estimated as well, asymptotically, as in the case where the correct submodel is known.

4. Simulation

In this section we present simulations to investigate the performance of our estimator, which include testing its robustness to possible dependence among the populations. The performance of the estimator $\hat{\Lambda}_j$ is assessed via the cumulated square errors (CSE), CSE $= (1/n)\sum_{j=1}^n (\hat{\Lambda}_j - \Lambda_j)^2$, or the root mean square error (RMSE) at t_j RMSE $= \sqrt{E(\hat{\Lambda}_j - \Lambda_j)^2}$, $j = 1, \ldots, n$. In both simulations, we use the Gaussian kernel. We considered different scenarios with the simulation replicated 5,000 times in each one.

Table 2. Simulation details for three lists: lists 1 and 2 are dependent and each is independent of list 3.

	$P(S_1)$	$P(S_2\|S_1)$	$P(S_2\|\overline{S}_1)$	$P(S_3)$	p_0
Scenario 0	3/4	6/7	1/7	3/4	0.054
Scenario 1	1/3	2/3	1/3	3/4	0.111
Scenario 2	1/4	2/3	1/3	1/2	0.250
Scenario 3	1/4	4/5	1/5	1/5	0.480
Scenario 4	1/7	3/7	1/10	1/8	0.675

Simulation 1. The purpose here was to compare the performances of the proposed two-stage method and the SI method. The number of periods was $n = 100$, and the number of lists was 3. Lists 1 and 2 were assumed to be dependent but independent of list 3. The dependence structures considered are itemized in Table 2, where S_i denotes the event of "being captured by list i", $\overline{S}_i$ denotes the event of "not being captured by list i", and p_0 denotes the probability of not being captured by any list, changes from 0.111 to 0.675, corresponding Scenario 1 to 4, respectively. The function Λ was set to $\Lambda(t) = 8,000 + 1.5(t - 50)^2 + 10t$. The observed data n_{cj} were independently generated according to Poisson distributions with means $\Lambda_j p_c$, $j = 1, \ldots, n$, $c = 1, \ldots,$ $K = 2^3 - 1$. Scenario 0 is for testing (2.9).

Table 3 gives the results using the proposed method with bandwidth $h = 4$ and using SI methods, including the bias, the empirical standard deviation (SD) and RMSE of the estimates. Comparisons were made for time periods $j = 10, 25, 50, 75, 90$, which correspond to the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of the observed time. The results show that the proposed method greatly reduces the variance of the SI estimators for all cases, and reduces the bias of the SI estimators for $p_0 = 0.48$ and $p_0 = 0.675$, so that the RMSE for the proposed method was about 1/10 of that of SI results, on average.

Figure 6.1 shows the estimates (dashed) and 95% confidence limits (dotted-linear) for Λ, with $p_0 = 0.675$, from a typical sample with a bandwidth $h = 4$ chosen by minimizing the estimated *CMSE*, denoted by *ECMSE*. The typical sample was selected in such a way that its *CSE*-value is the median of the 5,000 *CSE*-values. In order to choose the bandwidth, we first gave a series of values of h. For each pre-selected bandwidth h, we computed

$$ECMSE = \sum_{j=1}^{n} \left\{ \widehat{bias}^2\{\hat{\Lambda}(t_j, h)\} + \widehat{Var}\{\hat{\Lambda}(t_j, h)\} \right\},$$

Table 3. The average, the SD and the MSE over the 5,000 replications at time point $j = 10, 25, 50, 75, 90$.

			Time				
p_0	Method		10	25	50	75	90
	true		10500	9187.5	8500	9687.5	11300
0.111	proposed	bias	31.73	37.25	37.51	38.24	32.96
		SD	29.40	26.34	25.05	26.95	30.18
		RMSE	43.26	45.62	45.11	46.78	44.69
	SI	bias	1.49	5.69	6.57	7.02	1.73
		SD	156.34	149.70	142.60	149.25	161.20
		RMSE	156.35	149.81	142.75	149.41	161.21
0.25	proposed	bias	32.11	37.63	37.07	37.19	33.23
		SD	39.76	33.95	32.62	35.62	40.32
		RMSE	51.11	50.68	49.38	51.50	52.25
	SI	bias	8.10	11.88	10.76	11.27	5.64
		SD	264.75	252.37	246.44	262.95	286.99
		RMSE	264.87	252.65	246.67	263.19	287.05
0.48	proposed	bias	31.69	36.69	36.66	37.35	32.79
		SD	80.65	70.87	65.56	74.53	85.50
		RMSE	86.65	79.81	75.12	83.36	91.57
	SI	bias	76.49	56.22	63.32	64.02	64.52
		SD	736.36	693.27	680.09	712.31	759.88
		RMSE	740.32	695.55	683.03	715.19	762.61
0.675	proposed	bias	31.02	37.36	36.42	36.06	32.16
		SD	81.46	71.32	66.34	74.23	86.15
		RMSE	87.17	80.51	75.68	82.52	91.96
	SI	bias	71.71	48.96	49.90	68.31	63.43
		SD	740.54	695.37	662.97	702.02	762.06
		RMSE	744.00	697.09	664.84	705.34	764.69

where $\widehat{bias}\{\hat\Lambda(t_j, h)\} = \nu_1 h^2 + \nu_2 h^3$ and ν_1, ν_2 were estimated by fitting the polynomial regression

$$E[\hat\Lambda(t_j, b)] = \nu_0 + \nu_1 b^2 + \nu_2 b^3$$

to the data $\{(b, \hat\Lambda(t_j, b)),\ b = h \pm r,\ r = 0, 0.1, 0.2, 0.3, 0.4\}$. The variance estimate $\widehat{Var}\{\hat\Lambda(t_j, h)\}$ was obtained by the bootstrap method described in Section 2. The plot of *ECMSE* vs h is shown in Figure 6.2. From Figure 6.1, we see that the proposed procedure with the bandwidth choice method produces reasonable estimates of the true population size function (solid line). Figure 6.1 also displays the SI estimates (dotted line), and shows that the SI estimator has large variance and leads to unreasonable variability of the results.

We also examined the accuracy of (2.9) when p_0 was small. The standard deviations, denoted by SD in Table 4, of 500 estimated $\hat\Lambda_j$, based on 500 simu-

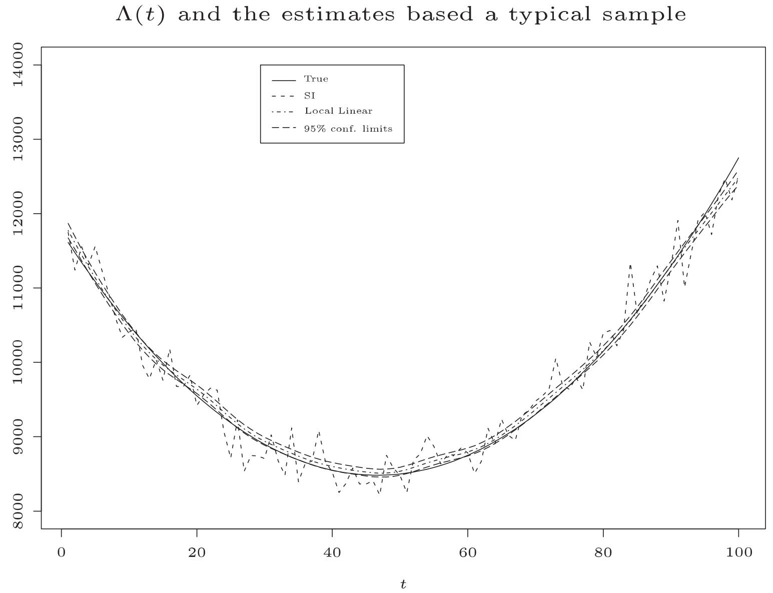

Figure 6.1. The estimated function $\widehat{\Lambda}_j$ using the proposed methods (dashed), and their 95% confidence interval (dotted-linear) using the bandwidth $h = 4$ chosen by minimizing $ECMSE$, as well as the true value (solid), and SI (dotted) estimators for a typical sample with $p_0 = 0.675$.

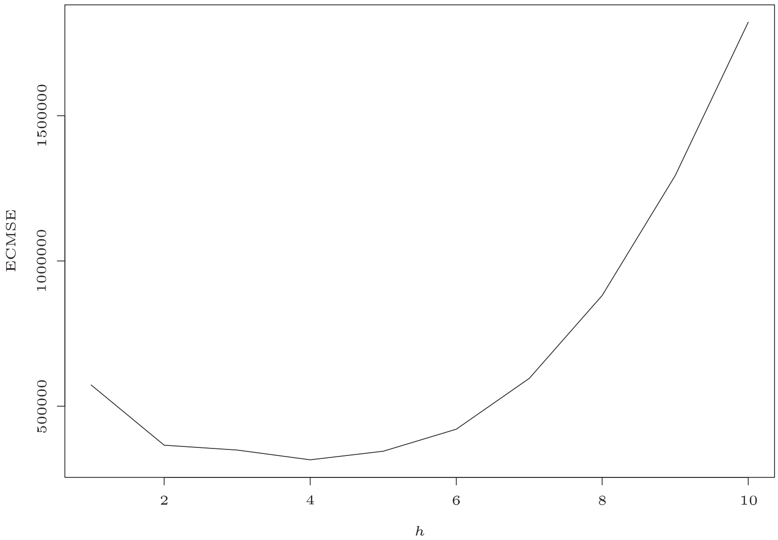

Figure 6.2. The $ECMSE$ against h for a typical sample from Ssimulation 2.

Table 4. True and estimated standard errors when using bandwidth = 5 for Scenario 0 and 1 at time points $j = 10, 25, 50, 75, 90$.

	$p_0 = 0.054$		$p_0 = 0.111$	
time	SD	$SD_{ave}(SD_{std})$	SD	$SD_{ave}(SD_{std})$
10	24.921	25.251(4.716)	27.510	25.824(5.126)
25	24.494	23.095(4.533)	25.948	23.753(4.675)
50	23.633	22.082(4.451)	25.042	22.904(4.322)
75	24.822	23.408(4.544)	27.027	24.512(4.844)
90	28.983	26.260(5.388)	30.397	26.464(5.294)

lations, can be regarded as the true standard errors. The average and the standard deviation of 500 estimated standard errors, denoted by SE_{ave} and SE_{std}, summarize the overall performance. Table 4 presents the results at the points $j = 10, 25, 50, 75, 90$. It suggests that our standard error formula is basically consistent with the true standard deviation when p_0 is small.

Simulation 2. This simulation was used to investigate the robustness of the proposed method to the possible dependence among population sizes at different periods. We applied our estimation procedure to a situation where the underlying target population N_t is modeled by an $AR(1)$ process

$$N_t|N_{t-1},\ldots,N_1 \overset{d}{=} \text{Bin}(N_{t-1},\gamma) + \text{sgn}(\Lambda_t - \gamma\Lambda_{t-1})\text{Pois}(|\Lambda_t - \gamma\Lambda_{t-1}|), t \geq 2$$
$$N_1 \overset{d}{=} \text{Pois}(\Lambda_1). \qquad (4.1)$$

It can be seen that $E[N_t] = \Lambda_t$ and that if γ is small so that $\Lambda_t - \gamma\Lambda_{t-1} > 0$ for all $t = 2, 3, \ldots$, then $Var[N_t] = \Lambda_t$. We chose $\Lambda_t = 8,000 + 1.5(t - 500)^2 + 10t$ as before, and $\gamma = 0.5$. The results are shown in Table 5. A comparison with Table 3 shows that our estimator behaves almost as well as in the case of independent N_t's.

Simulations were also carried out under a stronger dependence structure within a similar scheme as in Table 5, except that $\gamma = 0.9$. Similar results were obtained, suggesting that our method is robust against possible dependence among the populations to some extent.

5. An Example: Number of Drug Users in Hong Kong

In this section we re-examine the drug-user data in Table 1. The number of drug-users was known to have changed over the period from 1977 to 1997. The observed numbers of drug-users fluctuated around 10,000.

The proposed variable selection method in Section 3 was applied to the data set with the tuning parameter ϱ chosen by minimizing

$$AIC = -2L_1(\hat{\boldsymbol{\theta}}) + 2p \quad \text{and} \quad BIC = -2L_1(\hat{\boldsymbol{\theta}}) + \log(N)p,$$

Table 5. The average, the SD and the MSE over the 5,000 replications at time point $j = 10, 25, 50, 75, 90$ for the data with the dependent populations.

			Time				
p_0	Method		10	25	50	75	90
	true		10500	9187.5	8500	9687.5	11300
0.111	proposed	bias	21.96	24.45	24.46	23.54	21.45
		SD	49.48	44.36	43.37	45.68	50.54
		RMSE	54.13	50.66	49.80	51.39	54.90
	SI	bias	1.12	3.74	4.71	2.82	0.66
		SD	157.73	146.45	141.57	151.57	163.53
		RMSE	157.73	146.50	141.65	151.59	163.53
0.250	proposed	bias	23.52	24.76	23.67	24.43	23.34
		SD	55.66	50.55	47.36	52.21	57.08
		RMSE	60.42	56.29	52.95	57.64	61.67
	SI	bias	11.32	15.85	11.17	12.99	9.93
		SD	272.06	256.08	247.41	260.22	284.32
		RMSE	272.30	256.57	247.66	260.54	284.49
0.480	proposed	bias	25.75	25.29	25.18	25.26	24.50
		SD	91.76	81.56	75.39	84.12	95.24
		RMSE	95.31	85.39	79.48	87.84	98.34
	SI	bias	65.79	69.07	74.56	68.18	69.74
		SD	733.67	699.45	674.26	709.06	747.32
		RMSE	736.61	702.85	678.37	712.33	750.57
0.675	proposed	bias	22.83	26.46	25.78	26.40	26.22
		SD	151.53	134.20	123.76	141.39	164.01
		RMSE	153.24	136.78	126.42	143.83	166.09
	SI	bias	120.87	154.43	169.47	142.00	200.17
		SD	1412.90	1314.61	1270.48	1364.86	1452.73
		RMSE	1418.06	1323.65	1281.73	1372.22	1466.46

where p is the number of non-zero components in $\hat{\boldsymbol{\theta}}$. The AICs' and BICs' results suggest $\varrho = 0.15$, which find $\hat{\theta}_0 = -8.4918$, $\hat{\theta}_1 = 0.9638$, $\hat{\theta}_3 = 0.7679$, $\hat{\theta}_4 = 1.9445$, $\hat{\theta}_{12} = 1.9736$, $\hat{\theta}_{34} = 0.7323$, and the others estimated as zero. These results suggest that the four lists can be divided into two nearly independent pairs: Police Department and Correctional Services Department, and the Social Welfare Department and Hospitals. While the two within-pair interactions are rather strong, there basically is no interaction between the two pairs. Another interesting implication of $\theta_2 = 0$ and $\theta_{12} > 0$ is that the drug users recorded by the Correctional Services Department might have all been recorded by the Police Department.

We used the local linear technique to estimate the number of drug users in Hong Kong for all the 42 half-year periods during 1977-1997. Figure 6.3 plots the estimated numbers of the drug users (solid line) and their pointwise approximate

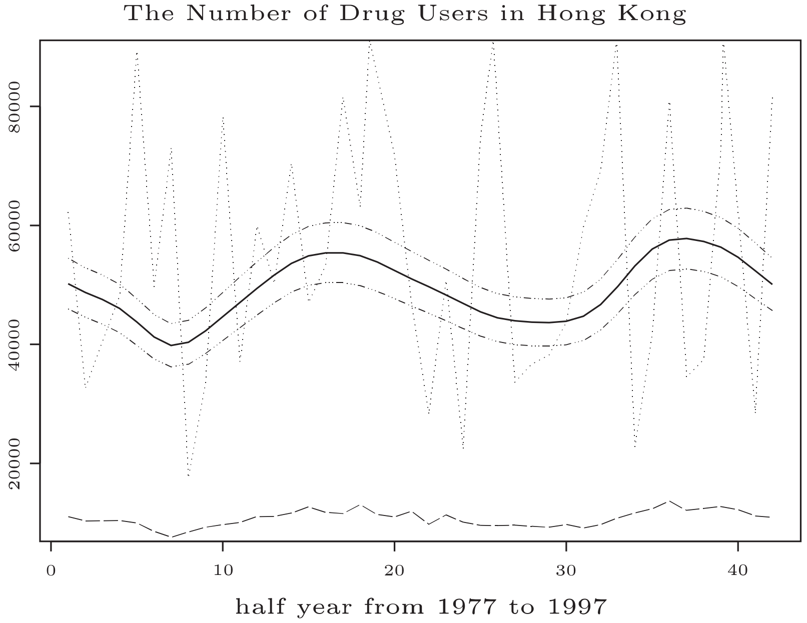

Figure 6.3. The estimated numbers (solid) of drug users, with 95% confidence limits (dotted-linear), for the bandwidth $h = 1.5$ using the proposed method and the SI (dotted) method in Hong Kong during the period from 1977 to 1997, as well as the observed numbers (dashed).

95% confidence limits (dotted-linear line), with the bandwidth $h = 1.5$ chosen by minimizing ECMSE. We also plot the SI (dotted line) estimates and the number of the observed drug users (dashed line) in Figure 6.3. It can been seen that, while the local linear ML estimates are fairly stable, the SI estimates show extraordinarily large variation. It is also clear that a simple summation of the observed numbers from different lists greatly underestimates the total number of drug users by a factor of around 4.6, on average.

Our estimates and the observed numbers of drug-users achieve peaks and troughs roughly at the same times and both curves show similar increasing then decreasing trend over time, which suggests that the proposed methods and the bandwidth selection are reasonable. These results also shows a seasonal pattern in the number of drug-users. The peaks are roughly in the years 1975, 1985 and 1995, and the waves seem to recur after periods of about 10 years. An explanation of this would require additional covariate information. Apparently, peaks and troughs are in line with the economic cycle of Hong Kong over the period. Low unemployment rate is linked to the peaks, whereas troughs were found in the times of high unemployment rate. This suggests that in good economic times, people can afford to pay for drugs (especially so-called party drugs, for example, Ketamine). The times with high unemployment rate would reduce affordabilty among the drug users in the community.

6. Discussion

The proposed two-stage procedure to estimate the population size in multi-contingency tables overcomes the closed-population assumption and makes use of the information on the population sizes without specifying the form of Λ_j, $j = 1, \ldots, n$. It is shown that the method outperforms the SI method. The method is especially suitable for the case where the population size changes smoothly with time.

The popular log-linear model for the cell probabilities $p_c, c = 1, \ldots, 2^k - 1$ gets too complex when k, the number of lists, is large, and makes it difficult to interpret the fitted model coefficients. We provide an efficient and easy-to-implement method to cope with the model selection problem.

When additional covariate information is available for each individual, one can formulate log-linear models for capture probabilities that incorporate this information. Further research is needed to accommodate heterogeneity among individuals for an open population.

Acknowledgement

Lin's research was supported in part by the Fund of National Natural Science (Grant 10771148) of China. The authors thank the Editors, an associate editor, and a referee for their valuable comments that have led to a significant improvement of this paper.

Appendix. Proofs

Proof of Theorem 1. First note that, given $N = \sum_{j=1}^{n} n_j$, the (conditional) likelihood of the parameter $\boldsymbol{\theta}$ based on n_{cj}, $c = 1, \ldots, K$; $j = 1, \ldots, n$, is

$$\exp L_1(\boldsymbol{\theta}) = \frac{N!}{\prod_{c=1}^{K} n_{c\cdot}!} \prod_{c=1}^{K} (q_c(\boldsymbol{\theta}))^{n_{c\cdot}},$$

where $n_{c\cdot} = \sum_{j=1}^{n} n_{cj}$. We now apply proposition (iv) in Subsection **5e.2** of Rao (1973). Since the parameterizations (2.1) are smooth enough for $q_c(\boldsymbol{\theta})$, $c = 1, \ldots, K$, to have continuous first-order partial derivatives, we only need to verify the non-singularity of $\mathbf{I}(\boldsymbol{\theta}^0)$ and the strong identifiability condition, ie., that for any $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\inf_{|\boldsymbol{\theta}-\boldsymbol{\theta}^0|>\delta} \sum_{c=1}^{K} q_c(\boldsymbol{\theta}^0) \log \frac{q_c(\boldsymbol{\theta}^0)}{q_c(\boldsymbol{\theta})} \geq \epsilon. \qquad \text{(A.1)}$$

Denote the sum on the left-hand side of (A.1) by $f(\boldsymbol{\theta})$. Introduce the $(K-1)$-dimensional vectors $\mathbf{b}_c$, $c = 0, \ldots, K$, so that (2.1) can be written as $\log p_c = \theta_0 + \mathbf{b}_c^T\boldsymbol{\theta}$. By the model specification, we have $\theta_0 = -\log(\sum_{c=0}^K e^{\mathbf{b}_c^T\boldsymbol{\theta}})$. Straightforward calculations show that

$$\frac{\partial f(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{c=1}^K (q_c(\boldsymbol{\theta}) - q_c(\boldsymbol{\theta}^0))\mathbf{b}_c,$$

$$\frac{\partial^2 f(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \sum_{c=1}^K q_c(\boldsymbol{\theta})\mathbf{b}_c\mathbf{b}_c^T - \sum_{c=1}^K q_c(\boldsymbol{\theta})\mathbf{b}_c \sum_{c=1}^K q_c(\boldsymbol{\theta})\mathbf{b}_c^T.$$

Since for any $\boldsymbol{\theta}$ in the parameter space, the $q_c(\boldsymbol{\theta})$ are all positive and the dimension of $\mathbf{b}_c$ is smaller than K, we can prove $[\partial^2 f(\boldsymbol{\theta})]/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$ is positive definite. So $f(\boldsymbol{\theta})$ has vanishing partial derivatives at $\boldsymbol{\theta}^0$ and is strictly convex over the parameter space. This shows that $\boldsymbol{\theta}^0$ is the unique minimizer of $f(\boldsymbol{\theta})$. By the continuity of $f(\boldsymbol{\theta})$, we have (A.1). The non-singularity of $\mathbf{I}(\boldsymbol{\theta}^0)$ follows immediately if we write $\mathbf{I}(\boldsymbol{\theta}) = [\partial^2 f(\boldsymbol{\theta})]/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$. By proposition (iv) in Subsection **5e.2** of Rao (1973) and the notes at the end of that subsection, we have, as $N \to \infty$ along a deterministic sequence,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}^0)^{-1}). \tag{A.2}$$

This distributional convergence still holds if $N \to \infty$ in probability, but in our case this is clearly true when $\eta \to \infty$, since N has a Poisson distribution with mean $\eta\int_0^\tau \lambda(s)ds > 0$. Therefore, we still have (A.2) as $\eta \to \infty$. As a result (2.7) follows from the fact that $N/\eta \to \int_0^\tau \lambda(t)dt$ in probability and Slutsky's theorem.

Proof of Theorem 2. Let $c_n = (nh)^{-1/2}$, $\overline{\Lambda}(t) = \Lambda(t_0) + \Lambda'(t_0)(t - t_0)$, $x_t = (1, (t-t_0)/h)^T$, and $\hat{\boldsymbol{\beta}}^* = c_n^{-1}(\hat\beta_1 - \Lambda(t_0), h(\hat\beta_2 - \Lambda'(t_0)))^T$. Suppose $\hat{\boldsymbol{\beta}} = (\hat\beta_1, \hat\beta_2)^T$ maximizes (2.6). Then $\hat{\boldsymbol{\beta}}^*$ maximizes

$$\ell_n(\boldsymbol{\beta}^*) = h\sum_{j=1}^n \Big\{ n_j\Big[\log\{c_n x_{t_j}^T\boldsymbol{\beta}^* + \overline{\Lambda}(t_j)\} - \log\{\overline{\Lambda}(t_j)\}\Big] \\ -(1 - p_0(\hat{\boldsymbol{\theta}}))c_n x_{t_j}^T\boldsymbol{\beta}^*\Big\} \times K_h(t_j - t_0).$$

Using a Taylor series expansion, we obtain that

$$\ell_n(\boldsymbol{\beta}^*) = W_n^T\boldsymbol{\beta}^* + \frac{1}{2}\boldsymbol{\beta}^{*T}A_n\boldsymbol{\beta}^*(1 + o_p(1)),$$

where

$$W_n = hc_n\sum_{j=1}^n \Big\{ n_j\frac{x_{t_j}}{\overline{\Lambda}(t_j)} - (1 - p_0(\hat{\boldsymbol{\theta}}))x_{t_j}\Big\}K_h(t_j - t_0),$$

$$A_n = -hc_n^2 \sum_{j=1}^{n} \left(n_j \frac{x_{t_j} x_{t_j}^T}{\overline{\Lambda}^2(t_j)} \right) K_h(t_j - t_0).$$

It can be shown that, with $u_i = \int_{-\infty}^{\infty} x^i K(x)dx$,

$$A_n = -\tfrac{1-p_0(\boldsymbol{\theta}^0)}{\tau^2 C\lambda(t_0)} \begin{pmatrix} u_0 & u_1 \\ u_1 & u_2 \end{pmatrix} (1 + o_p(1)) \triangleq -A + o_p(1).$$

By applying the Convexity Lemma (see Pollard (1991)), we obtain that $\hat{\boldsymbol{\beta}}^* = A^{-1}W_n + o_p(1)$. Hence the asymptotic normality of $\hat{\boldsymbol{\beta}}^*$ will follow from that of W_n, which we establish next.

By the definition of W_n, it can be shown that $W_n = W_{n1} + W_{n2}$ where

$$W_{n1} = hc_n \sum_{j=1}^{n} \left\{ n_j \frac{x_{t_j}}{\overline{\Lambda}(t_j)} - (1 - p_0(\boldsymbol{\theta}^0))x_{t_j} \right\} K_h(t_j - t_0),$$

$$W_{n2} = hc_n \sum_{j=1}^{n} \left(p_0(\hat{\boldsymbol{\theta}}) - p_0(\boldsymbol{\theta}^0) \right) x_{t_j} K_h(t_j - t_0).$$

By the mean value representation, it can be shown that $W_{n2} = o_p(1)$. Hence the asymptotic normality of W_n follows from that of W_{n1}. It can be proved that (Fan (1992))

$$E[W_{n1}] = \frac{c_n^{-1} h^2 \lambda''(t_0)(1 - p_0(\boldsymbol{\theta}^0))}{2\tau\lambda(t_0)} \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} (1 + o_p(1)),$$

and, since n_j is Poisson, we have

$$Var[W_{n1}] = \frac{1 - p_0(\boldsymbol{\theta}^0)}{C\tau^2\lambda(t_0)} \begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix} (1 + o_p(1)),$$

where $v_i = \int_{-\infty}^{\infty} x^i K^2(x)dx$. Using the assumption of a Poisson distribution for n_j and $\lambda(t_0) > 0$, it can be shown that Liapounov's condition is satisfied, and hence $\hat{\boldsymbol{\beta}}^*$ is asymptotically normal. This establishes Theorem 2.

References

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 939-967.

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87**, 210–217.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45**, 395-413.

Cormack, R. M. and Jupp, P. E. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika* **78**, 911-916.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J., Lin, H. Z. and Zhou, Y. (2006), Local partial-likelihood estimation for life time data. *Ann. Statist.* Vol, 34. 290-325.

Fienberg, S. E. (1972). The multiple recapture census for closed population and incomplete 2^k contingency tables. *Biometrika* **59**, 591-603.

Huggins, R. M., Yang, H. C., Chao, A. and Yip, P. S. F. (2003). Population size estimation using local sample coverage for open populations. *J. Statist. Plann. Inference* **113**, 699-714.

Huggins, R. M. and Yip, P. S. F. (2003). Estimation of the size of the open population from capture-recapture data using weighted martingale methods. *Biometrics* **55**, 387-395.

International Working Group for Disease Monitoring and Forecasting. (1995a). Capture-recapture and multiple-record systems estimation. I: History and theoretical development. *Am. J. Epidemiol.* **142**, 1047-1058.

International Working Group for Disease Monitoring and Forecasting. (1995b). Capture-recapture and multiple-record systems estimation. II: Applications in human diseases. *Am. J. Epidemiol.* **142**, 1059-1068.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econom. Theory* **7**, 186-199.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd edition. Wiley, New York.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92**, 1049-1062.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Ann. Math. Statist.* **43**, 142-152.

Sandland, R. L. and Cormack, R. M. (1984). Statistical inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika* **71**, 27-33.

Tibshirani, R. (1996). Regression shrinkage and selection via LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Yang, H. C. and Huggins, R. M. (2003). The estimation of the size of the open population using local estimating equations. *Statist. Sinica* **13**, 673-689.

Yang, H. C., Huggins, R. M. and Clark, A. S. S. (2003). Estimation of the size of an open population using local estimating equations II: A partially parametric approach. *Biometrics* **59**, 365-374.

Yip, P. S. F. (1991). Conditional inference on a mixture model. *Comm. Statist. A - Theor.* **6**, 2045-2058.

School of Mathematics, Sichuan University, Chengdu, Sichuan 610064, P. R. China.
E-mail: huazhenlin@hotmail.com

Department of Social Work and Social Administration, The University of Hong Kong, Pokfulam Road, Hong Kong.
E-mail: sfpyip@hku.hk

School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia.
E-mail: feng.chen@unsw.edu.au

(Received September 2006; accepted September 2007)