Analysis of sparse contingency tables with applications

Marijus Radavičius and Jurgita Židanavičiūtė (Vilnius)

Content

1. *Introduction*
2. *Semiparametric bootstrap and smoothing*
3. *Markov chain and generalized logit model*

1. Introduction

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. Traditionally, the expected (under the null hypothesis) frequency is required to exceed 5 in (almost) all cells of the contingency table. If this condition is violated, the χ^2 approximations of goodness-of-fit statistics may be inaccurate and the table is said to be **sparse** (Agresti (1990)).

Several techniques have been proposed to tackle the problem:

(1) exact tests and alternative approximations (Agresti (1990), Müller and Osius (2003)),

(2) smoothing is applied to ordered data (Simonoff (1995)),

(3) parametric and nonparametric bootstrap (Davier (1997)),

(4) Bayes method (Congdon (2005)),

ect.

They all are not applicable or have some limitations in case of categorical data with very sparse contingency tables containing a large portion of zeros.

We introduce a smoothed bootstrap method based on a special representation form of data under consideration and the method of *semiparametric smoothing* (Faddy and Jones (1998)).

In our study we deal with data of bacteria genoms from the GenBank database. It is supposed that non-coding regions of the genomes are random sequences generated by a finite-order Markov chain and the problem is to estimate the Markov order.

The analogous problem has been investigated via loglinear analysis (Avery and Henderson (1999)).

2. Semiparametric bootstrap and smoothing

Let

$$Y = (Y_l,\ l = 1, \ldots, n) \in \mathcal{A}^n,$$

be a random sequence with unknown distribution G. Here $\mathcal{A}$ is a finite state space (alphabet), $\mathcal{A}^n = \mathcal{A} \times \ldots \times \mathcal{A}$.

On the other hand, let

$$X(t) = (X_l(t),\ l = 1, \ldots, n) \in \mathcal{A}^n, \quad t \in \mathbf{Z}_+,$$

be a discrete time homogeneous Markov chain with a transition probability matrix Π and initial distribution π, $\pi(a) = \mathbf{P}\{X(0) = a\},\ a \in \mathcal{A}^n$.

Under general conditions the Markov chain X is ergodic, and let Q denote its stationary distribution.

The problem: given an observation y from Y, to check

$$H_0 : G = Q \quad \text{versus} \quad H_0 : G \neq Q \tag{1}$$

when only the transition probability matrix Π is available.

Let

$$X^*(t) = X^*(t|y) = \{X_l^*(t),\ l = 1, \ldots, n\},\ t \in \mathbf{Z}_+,$$

be Markov chain with the transition probabilities Π starting from y and

$$Q_t(a) = \mathbf{P}\{X^*(t) = a\}, \quad a \in \mathcal{A}^n.$$

Then $Q_0 = \delta_y$ and $Q_t \to Q$ as $t \to \infty$.

- Thus, for fixed t, Q_t can be treated as "smoothed (toward Q)" empirical distribution of Y.

Consider a vector functional $\lambda = \lambda(Q)$ defined on the measure space and set

$$\eta(i, l) := (\hat{\lambda}_i - \hat{\lambda}_{i+l})^\top \hat{V}^-(\hat{\lambda}_i - \hat{\lambda}_{i+l})$$

where

$$\hat{\lambda}_j = \lambda(\hat{Q}_j),$$

$\hat{Q}_j$ is empirical distribution of $X^*(j)$ based on resampling from $X^*(j)$ (Q_j) given $X^*(0) = y$, $\hat{V}$ is the estimated covariance matrix of $\hat{\lambda}_i$, and $\hat{V}^-$ is generalized inverse of $\hat{V}$.

Testing the null hypothesis is based on statistics

$$\eta^*(l) := \max_i \eta(i,l). \tag{2}$$

The parameter l of the method should be taken as large as possible but is limited by computer resources.

The **resampling methods (bootstrap)** can be applied to estimate the distribution of statistics (2) and to evaluate its critical and p-values.

Remark:
The proposed method of "semiparametric" smoothing can be applied as independent technique of adaptive smoothing.

3. Markov chains and generalized logit model

We apply the logit analysis to assess the Markov property of non-coding regions of genetic (DNA) sequences.

For simplicity, we describe only the case of testing if the Markov order $k = 1$.

Let

$$y = \{y_i, i = 1, ..., 2m + 1\}, \quad y_i \in \mathcal{A},$$

be an observed DNA sequence of nucleotides, $\mathcal{A} = \{A, C, G, T\}$.

For triplets LCR (Left, Centre, Right) in the sequence, denote

$$\{(u_i, v_i, z_i), i = 1, ..., m\} := \{(y_{2i-1}, y_{2i}, y_{2i+1}), i = 1, ..., m\}$$

$$p(uvz) := \mathbf{P}\{(y_{2i-1}, y_{2i}, y_{2i+1}) = (u, v, z)\}, \quad i = \overline{1, m}\}.$$

Condition (C): v_j and $\{(u_i, v_i, z_i), i = 1, ..., m, i \neq j\}$ are conditionally independent given (u_j, z_j), $j = 1, ..., m$, and its conditional distribution is independent of i.

The special form of the data and **Condition (C)** ensures that the standard assumptions of **generalized logit model** hold and standard software can be applied to fit the model and to test hypotheses about the order k of the Markov chain.

Generalized logit model (saturated):

$$\log\left(\frac{p(uvz)}{p(urz)}\right) = \beta_z^C + \beta_{uv}^{LC} + \beta_{vz}^{CR} + \beta_{uvz}^{LCR}. \tag{3}$$

Here β's denote the unknown parameters and r stands for the reference value.

For the first order Markov chain, u_i and z_i are conditionally independent given v_i and hence the null hypothesis

$$H_0 : \beta_{uvz}^{LCR} = 0$$

should be valid.

If y is first order Markov, by making use of "true" generalized logits (3) it is easy to calculate conditional probabilities

$$p(v|u,z) := \mathbf{P}\{y_{2i} = v|(y_{2i-1} = u, y_{2i+1}) = z\}, \quad i = \overline{1,m},$$

and hence generate via Gibbs sampling a homogeneous Markov chain which has distribution of y as its stationary distribution. In practice, logits (3) should be replaced by their estimates.

Final Remarks:

The number of parameters to be estimated in generalized logit model increases very fast as Markov order k increases. However, because of heterogeneity of DNA sequences (rather short coding and non-coding regions) even for k=1 the contingency tables corresponding to (3) are sparse.

However, contingency tables obtained by resampling from Q_j (replications of the sequence $X_l^*(j)$) are not sparse, provided j is sufficiently large and large enough number of replications are generated.

We have not yet obtain conclusive results about the performance of the procedure proposed.