MODEL TESTING FOR HIGH-DIMENSIONAL CONTINGENCY TABLES WITH APPLICATION IN GENETICS

M. Radavičius, J. Židanavičiūtė
Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: mrad@ktl.mii.lt

Abstract

In the paper it is supposed that random sequences is a finite-order Markov chain. The analysis of high order interactions in such sequences leads to large and sparse contingency tables. A special data presentation form and data transformation via multidimensional scaling enables us to apply generalized logit model and reduce the problem to the fitting nonparametric multinomial logistic regression with a few quantitative variables.

1 Introduction

Let $x := \{x_1, \dots, x_n\}$ be a sequence of random elements with values from a finite alphabet $\mathcal{A}$. DNA molecule, for instance, consists of nucleotides of four types: adenine (A), guanine (G), citozine (C), and timine (T). Thus, it is a sequence with the alphabet $\mathcal{A} = \{A, C, G, T\}$.

In applications it is usually assumed that x form a homogenous Markov chain of a finite order k. The number of parameters of the model increases very fast with k. Therefore fitting the model for observed data is a challenging problem even for moderate values of k. A common practice is to perform loglinear analysis however the basic conditions are not met in this case and certain adjustments are necessary (see, e.g. [3]).

In this paper we use a special data presentation form and data transformation via multidimensional scaling for statistical analysis of such random sequences. This enables us to apply generalized logit model [1, 9] and reduce the problem to the fitting nonparametric multinomial logistic regression with a few quantitative variables.

2 Markov property and generalized logit

In this section basic notions of Markov fields are introduced and their relations with generalized logit model is described.

Let $N := \{1, \dots, n\}$. Fix some positive integer $m < n/2$ and define $N^\circ := \{m + 1, \dots, n - m\}$, $\partial N := N \setminus N^\circ$, $U(l) = U_m(l) := [l - m, l + m] \setminus \{l\}$, $l \in N^\circ$ where $[i, j] := \{i, i + 1, \dots, j\}$, $i < j$, $i, j \in N$, is an interval of integers. Given $x \in \mathcal{A}^n$ and a set of indices $I \subset N$ let $x_I := \{x_i, i \in I\}$ denote the corresponding subsequence of x.

Definition 1. A random sequence $x \in \mathcal{A}^n$ is homogeneous Markov chain (random field) of order m if $\forall l \in N^\circ$ and $a \in \mathcal{A}^n$

$$\mathbf{P}\{x_l = a_l | x_j = a_j, j \neq l\} = \mathbf{P}\{x_l = a_l | x_{U_m(l)} = a_{U_m(l)}\} := p(a_l | a_{U_m(l)}). \tag{1}$$

It is supposed that values of the Markov chain x are fixed on the boundary ∂N, $x_{\partial N} = c_{\partial N}$ for some $c \in \mathcal{A}^n$.

Set $\mathcal{X}_+ := \{a \in \mathcal{A}^n : \ a_{\partial N} = c_{\partial N}\}$. If probabilities $\mathbf{P}\{x = a\} > 0$ for all $a \in \mathcal{X}_+$, the distribution of the homogeneous Markov chain of the order m is uniquely determined by *odds ratios* for some reference value, say $b \in \mathcal{A}$, given the values $z = x_{U_m(l)}$ of the m nearest neighbours $U_m(l)$ (Hamersley-Clifford theorem, see [4]):

$$O_{y|b} = Q_{y|b}(z) := \frac{p(y|z)}{p(b|z)}, \quad y \in \mathcal{A}, \ z \in \mathcal{A}^{2m}. \tag{2}$$

Let us introduce the following structure of the observed sequence $x \in \mathcal{A}^n$ with $n = n_m(m+1) - 1$, n_m is an integer,

$$x = \{(y_l, z_l), l \in S\}, \quad S = S_{n,m} = N^\circ \bigcap \{m+1, 2m+1, \ldots, n-m\}, \tag{3}$$

where $y_l := x_{(m+1)l}$ is a target variable and $z_l = x_{U_m(l)}$ is a vector of explanatory variables, $l \in S$.

Assume

(a) $\{y_l, \ l \in S\}$ are conditionally independent given $\{z_l, \ l \in S\}$,

(b) the effect of z on y does not depend on the position l.

Note that these assumptions are fulfilled if x is generated by a homogeneous Markov chain of the order m. They ensure that common conditions of the generalized logit model [1] are satisfied. The Markov property implies the additional conditions on odds ratio (2). Namely, the odds ratio depends on $z = x_{U_m(l)}$ only through interactions $x_{I_j}, \ j = 0, \ldots, m$, where $I_j = U_m(l) \cup [l - m + j, l + j]$. This gives a basis for testing Markovity and selection of the Markov order.

The main problem is the sparsity of the data which takes place even for moderate values of m. A way to overcome this problem is "smoothing" of the data. Since the data is nominal the smoothing can not be made in a standard way. One can proceed as follows.

For simplicity assume that $\mathcal{A} = \{0, 1\}$. Then any sequence $z \in \mathcal{A}^m$ can be identified with a rational (dyadic) number $r = r(z) \in [0, 1)$ via its dyadic representation. Let $\rho(u, v)$ be some natural distance between sequences $u, v \in \mathcal{A}^m$. In computer science, Levenshtein or other edit-type distance is usually used. Further, let $\tau = \tau(M)$ denote a permutation of $M := \{1, \ldots, m\}$ and $z_\tau = (z_{\tau(1)}, \ldots, z_{\tau(m)})$. The permutation τ is to be selected to minimize the discrepancy between the distance $\rho(u, v)$ and $|r(\tau(u)) - r(\tau(v))|$ (multidimensional scaling). Let τ^* denote the solution of this problem. Then the problem of estimation of the odds ratios $O_{a|b}(z), \ y \in \mathcal{A}, \ z \in \mathcal{A}^{2m}$ can be reduced to that of (nonparametric) estimation of the functions $\psi_y : [0, 1) \to [0, \infty)$ defined by

$$\psi_y\big(r\big(\tau^*(z)\big)\big) = O_{y|b}(z), \quad y \in \mathcal{A}, \ z \in \mathcal{A}^{2m}. \tag{4}$$

This task can be solved via kernel smoothing, (orthogonal) series expansions, and another common nonparametric technique. We refer to [7, 5, 6].

3 Applications: DNA sequences

It is natural to consider DNA evolution in time as a homogeneous Markov process. The first stochastic models of DNA evolution assume that the nucleotides in DNA sequence evolve independently from each another. Recently context-dependent evolution models have been proposed (see [8] and references therein). In these models, mutation of each nucleotide depends on its neighbouring nucleotides (context). If DNA sequence evolution in time is reversible and mutation of nucleotides depends on their nearest neighbours, say k from the each side, then stationary distribution of nucleotides in the sequence is k-th order Markov chain (see, e.g., [8]). Is this valid for real DNA sequences?

Avery and Henderson [3] (see also [2]) have considered this problem by using non-coding sequences found within the human (*preproglucagon*) gene. Supposing that DNA sequence is a finite order Markov chain, they have analyzed pairs and triplets of nucleotides, fitted log-linear model to the observed nucleotide frequencies and assessed the Markov chain order. They have showed that the first order Markov chain is not a good fit to the data while the second-order hypothesis is not rejected. However, because of linear and statistical dependencies in the data the standard assumptions of the log-linear model do not hold. Therefore certain adjustment of the χ^2 criterion is used and validated by means of simulation.

In our study we use data of bacteria genoms from the *GenBank* database. The data is presented in the special form described in Section 2. This ensures that the standard assumptions of generalized logit model hold and standard software (SAS system procedure CATMOD, [9]) can be applied to fit the model and test hypotheses provided $m \leq 2$. The hypothesis of second-order Markovity was rejected, e.g., for non-coding regions of bacteria *E.Coli* taken from the *GenBank*.

For larger m and $k \leq m$, the computations are prohibitively expensive. Representation (4) and nonparametric multinomial logistic regression (generalized logit) is applied to tackle the problem. In this case the functions ψ_y depend on two (or four) quantitative variables. The results of the analysis will be presented and discussed.

Bibliography

[1] Agresti A. (1990) *Categorical Data Analysis*. John Wiley & Sons, New York.

[2] Avery P.J. (2002) Fitting interconnected Markov chain models-DNA sequences and test cricket matches. *The Statistician*. Vol. **51**, pp. 267-278.

[3] Avery P.J., Henderson D.A. (2002) Fitting Markov chain models to discrete state series such as DNA sequences. *Appl.Statist.* Vol. **48**, pp. 53-61.

[4] Besag J. (1974) Spacial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. B.* Vol. **36**, pp. 192-236.

[5] Adrian W, Bowman and Adelchi Azzalini Bowman A.W., Azzalini A. (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

[6] Gao F., Wahba G., Klein R., Klein B. (2001) Smoothing spline ANOVA for multivariateBernoulli observations, with applications to ophthalmology data (with discussion) *J. Amer. Statist.Assoc.* Vol. **96**, pp. 127-160.

[7] Hastie T., Tibshirani R. (1990) *Generalized Additive Models*. Chapman and Hall, New York.

[8] Jensen J.L. (2005) Context dependent DNA evoliutionary models. *Research Reports 458*. Department of Mathematical Sciences, University of Aarhus.

[9] Stokes M.E., Davis C.S., Koch G.S. (2001) *Categorical Data Analysis Using the SAS(R) System*. Cary, NC: SAS Institute.