PROFILE STATISTICS FOR SPARSE CONTINGENCY TABLES

M. Radavičius, P.Samusenko
Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: mrad@ktl.mii.lt
Vilnius Gediminas Technical University
Vilnius, LITHUANIA
e-mail: Pavel.Samusenko@gmail.com

Abstract

Simple conditions for the inconsistency of classical goodness-of-fit tests in case of very sparse categorical data are given. The conditions have the following interesting feature of "reversed consistency": the greater deviation from the null hypothesis the less power of the test. In the paper χ^2-type criterion based on profile statistics is introduced as an alternative to classical tests in the case of sparse categorical data.

1 Introduction

Sometimes the inclusion of additional variables in a statistical analysis completely changes the previous conclusions. In the statistical analysis of categorical variables this phenomenon is known as Simpson's paradox. Consequently, all available variables should be included into this kind of analysis. Currently the amount of information is very extensive, therefore problems related to a large dimension and/or sparsity of data arise rather frequently. The sparsity problem is especially topical for categorical data. Relationships between quantitative (continuous) variables are usually described by covariance matrices. Thus, the number of model parameters increases quadratically with n, the dimension of the data. For categorical data, the number of unknown parameters grows exponentially with n. Consequently, even for a moderate number of categorical variables, many cells in the contingency table are empty or have small counts. Traditionally, expected (under the null hypothesis) frequencies in a contingency table are required to exceed 5 in the majority of their cells. 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).

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. Several techniques have been proposed to tackle the problem: exact tests and alternative approximations (Agresti, 1990; Hu, 1999; Müller and Osius, 2003), smoothing of ordered data (Simonoff, 1995), contingency table smoothing by means of generalized log-linear models with random effects (Coull and Agresti, 2003), the parametric and nonparametric bootstrap (Davier, 1997), Bayes approach (Agresti and Hitchcock, 2003; Congdon, 2005), and other methods (see, for instance, Kuss, 2002). They all are not applicable or have some limitations in case of very sparse contingency

tables. In this case, the classical statistical criteria become simply uninformative (inconsistent).

We formalize this statement in the next section. In the last section χ^2-type criterion based on profile statistics is introduced as an alternative to classical tests for very sparse categorical data.

2 Inconsistency of classical tests

In this section simple conditions for the inconsistency of classical goodness-of-fit tests in case of very sparse categorical data are given. Though rather restrictive, the conditions have the following interesting feature ("reversed consistency"): the greater deviation from the null hypothesis the less power of the test.

Let y_j denote an observed frequency of the category $j \in J = J_n := \{1, \dots, n\}$ in a sample of N iid observations. Hence $Y := (y_1, \dots, y_n) \sim Multinomial_n(N, P)$ where $P := (p_1, \dots, p_n) \in \mathcal{P}$,

$$\mathcal{P} := \left\{ q \in \mathbf{R}^n : q_j \geq 0,\, j = 1, \dots, n,\, \sum_{i=1}^{n} q_i = 1 \right\}.$$

We consider very sparse categorical data (contingency tables). Here it means that $n = n(N)$, $P = P(N)$ as $N \to \infty$ and

$$p_{max} := \max_{j \in J} p_j = p_{max}(N) = o(N^{-1}), \quad N \to \infty. \tag{1}$$

Let us assume for simplicity that a simple hypothesis

$$H_0 : P = P_0 \text{ versus } \; H_1 : P \neq P_0 \tag{2}$$

is to be tested on the basis of observed frequencies Y with a given $P_0 = (p_1^0, \dots, p_n^0) \in \mathcal{P}$. Suppose that all components of P_0 are positive and consider the likelihood ratio statistic

$$\begin{aligned} G^2 \;&=\; G^2(P_0, Y) := \sum_{j \in J} y_j \log\left(\frac{y_j}{N p_j^0}\right) =: H(Y) + L(P_0, Y) - N\log(N), \\ &H(Y) := \sum_{j \in J} y_j \log(y_j), \quad L(P_0, Y) := -\sum_{j \in J} y_j \log(p_j^0). \end{aligned}$$

It turns out that for sparse data the term $L(P_0, Y)$ often dominates $H(Y)$.

Proposition 1. *Assume sparsity* (1). *Then*

$$\mathbf{E}_P G^2(P_0, Y) + N\log(N) = \mathbf{E}_P L(P_0, Y) + O(N^2 p_{max}),$$

$$\mathbf{Var}_P G^2(P_0, Y) \leq \left(\sqrt{\mathbf{Var}_P(L(P_0, Y))} + O\left(N\sqrt{p_{max}}\right) \right)^2.$$

Proposition 2. *Let P_0 be a nondecreasing sequence. Suppose that there exists $j_0 \in \{2, \dots, n\}$ such that $\forall j < j_0$ the probabilities $p_j \le p_j^0$, $\forall j \ge j_0$ the probabilities $p_j \ge p_j^0$, and for some $\bar{p} = \bar{p}(N) > 0$ and positive constants Δ and D*

$$\sum_{i \in J} (p_j^0 - p_j) \log(p_j^0) \le -\Delta, \tag{3}$$

$$\sum_{i \in J} p_j^0 \left(\log(p_j^0) - \log(\bar{p}) \right)^2 \le D^2 N. \tag{4}$$

If (1) *holds then*

$$\frac{\mathbf{E}_P G^2(P_0, Y) - \mathbf{E}_{P_0} G^2(P_0, Y)}{\sqrt{\mathbf{Var}_{P_0} G^2(P_0, Y)}} < -\frac{\Delta}{D} + O(N p_{max}). \tag{5}$$

Example. For a given $\beta > 1$ and $\rho \in (0, 1/2)$, set $m = \left[N^\beta \right]$, $n = 2m$, $j_0 = m$,

$$p_j^0 = \rho/m, \ \forall j \le m, \quad p_j^0 = (1 - \rho)/m, \ \forall j > m,$$

$$p_j = 0, \ \forall j \le m, \quad p_j = 1/m, , \ \forall j > m.$$

Then the conditions of Proposition 2 are fulfilled and one can take $\Delta = \frac{\rho}{2} \log \left(\frac{1-\rho}{\rho} \right)$.

Corollary. *Let assumptions of Proposition 2 be valid. Then the likelihood ratio criterion is inconsistent for testing problem* (2).

Remark 1. Note that the asymptotic bound for the decrease in power given in (5) is proportional to $-\Delta$ whereas the constant Δ characterizes via (3) the deviation of the true distribution P from the null hypothesis.

Remark 2. Actually, the inconsistency stated in Corollary 1 is not an exceptional feature of the likelihood ratio statistic G^2. Analogous inconsistency results can be obtained for the other goodness-of-fit criteria, for example tests based on power-divergence statistics (Cressie and Read, 1984).

3 Profile statistics

For a given positive integer k, denote $J_k := \{1, \dots, k\}$. Let $\mathbf{h} : J_N \times (0, 1) \to \mathbf{R}^k$ be a given vector function, $H(j, z) := (h_1(j, z), \dots, h_k(j, z))^\top$. Define the profile statistic as

$$T = T(\hat{H}) := \left(\mathbf{Cov}_P(\hat{H}, \hat{H}) \right)^{-1/2} (\hat{H} - \mathbf{E}_P(\hat{H})) \tag{6}$$

where

$$\hat{H} := \sum_{j=1}^{n} h(Y_j, p_j^0). \tag{7}$$

For the fixed k, under some additional conditions an asymptotic normality of the statistic $T(\hat{H})$ can be proved using standard methods (Kolchin et al. 1978).

The power of the tests based on χ^2-type statistics $X^2 := |T(\hat{H})|^2$ crucially depends on the choice of k and the vector function $H \in \mathbf{R}^k$. Sometimes, for a given structure of q as in the example above, the choice of H is rather obvious. Otherwise, the problem can be reduced to the problem of supervised feature selection and dimensionality reduction. Given a large initial class $\mathcal{H}$ of (linearly independent) functions $h : J_N \times (0,1) \to \mathbf{R}$, the dimension k is chosen and the components of the vector function H are selected from it, for instance, by making use of the reproducing kernel Hilbert spaces and the projection pursuit methods (principal component analysis) (Fukumizu et al., 2004).

References

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

[2] Agresti, A. and Hitchcock, B. D. (2005). Bayes inference for categorical data analysis, *Statistical Methods and Applications* **14**, 297–330.

[3] Coull, B.A. and Agresti, A. (2003). Generalized log-linear models with random effects, with application to smoothing contingency tables, *Statistical Modelling* **3**, 251–271.

[4] Cressie, N. and Read, T. (1984). Multinomial Goodness of Fit Tests, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

[5] Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley & Sons, New York.

[6] von Davier, M. (1997). Bootstraping goodness-of-fit statistics for sparse categorical data. Results of a Monte Carlo study, *Methods of Psychological Research Online* **2**, No.2, Internet: http//www.pabst-publishers.de/mpr/.

[7] Fukumizu, K., Bach, F.R., and Jordan, M.I. (2004). *The Journal of Machine Learning Research archive* **5**, 73–99.

[8] Hu, M.Y. (1999). *Model Checking for incomplete high dimensional categorical data*. The Phd dissertation, University of California, Los Angeles.

[9] Kolchin, V.F., Sevastyanov, B., and Chistyakov V. (1978). *Random allocations*. Wiley, New York.

[10] Kuss, O. (2002). Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data, *Statistics in Medicine* **21**, 3789–3801.

[11] Müller, U.U. and Osius, G. (2003). Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, *Statistics* **37**, No.2, 119–143.

[12] Simonoff, J.S. (1995). Smoothing categorical data, *J.Statist.Plann.Infer.* **47**, 41–69.