AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 115–123

Profile Statistics for Sparse Contingency Tables under Poisson Sampling

Marijus Radavičius[1] and Pavel Samusenko[2]
[1]Institute of Mathematics and Informatics, Vilnius, Lithunia
[2]Vilnius Gediminas Technical University, Vilnius, Lithunia

Abstract: Simple conditions for the inconsistency of classical likelihood ratio (LR) test in case of very sparse categorical data are given. The LR type test based on profile statistics is proposed as an alternative. The performance of both tests for a sparse contingency table is compared by simulations.

1 Introduction

Recently amounts of information are 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 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*. We refer to Agresti (1990, 1999) for examples of sparse contingency tables and the further discussion on this topic.

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 (Smirnoff, 1995), contingency table smoothing by means of generalized log-linear models with random effects (Coull and Agresti, 2003), the parametric and nonparametric bootstrap (von Davier, 1997), Bayes approach (Agresti and Hitchcock, 2005; 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 case of Poisson sampling. In Section 3 a new likelihood ratio type criterion is introduced as an alternative to classical tests in case of very sparse contingency tables. The criterion is derived using the empirical Bayes approach and is based on the profile statistics of the contingency table. In the last section an adaptive procedure for nonparametric testing is described and some simulation results are presented.

2 Inconsistency of Classical Likelihood Ratio Test

In this section simple conditions for the inconsistency of the classical likelihood ratio test 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. Actually, the probability to reject some alternatives tends to 0 as their deviations from the null hypothesis increase.

Let y_j, $j \in J = J(n) := \{1, \dots, n\}$, denote independent Poisson observations. Hence $\underline{y} := (y_1, \dots, y_n) \sim Poisson_n(\underline{\mu})$, where $\underline{\mu} := (\mu_1, \dots, \mu_n) \in \mathcal{M} := [0, M_0]^n$, $M_0 > 0$. We consider very sparse categorical data (contingency tables) $\underline{y} \in \mathbf{Z}_+^n$. Here it means that $\mathbf{E}_{\underline{\mu}}(\underline{y}) = \underline{\mu} = \underline{\mu}(n)$ and as $n \to \infty$

$$\|\underline{\mu}\|_2^2 = o\left(\|\underline{\mu}\|_1\right) . \tag{1}$$

($\|\underline{\mu}\|_q$ denotes the l_q norm of $\underline{\mu}$).

Remark 1. Condition (1) together with $\|\underline{\mu}\|_\infty \le M_0$ implies

$$\|\underline{\mu}\|_1 \le M_0 n\,, \tag{2}$$

and for arbitrary $h > 0$,

$$h^2 |\{j : \mu_j \ge h\}| \le \|\underline{\mu}\|_2^2 = o\left(\|\underline{\mu}\|_1\right) , \tag{3}$$

$$h\|\underline{\mu}\|_1 \le h^2 |\{j : \mu_j \le h\}| + \|\underline{\mu}\|_2^2 \le h^2 n + o\left(\|\underline{\mu}\|_1\right) . \tag{4}$$

Here and later $|A|$ stands for the number of elements of the set A. From (2)–(4), it follows that

$$n_h(\underline{\mu}) := |\{j : \mu_j \ge h\}| = o\,(n)\,, \quad \forall h > 0\,,$$
$$\|\underline{\mu}\|_q^q = o\,(n)\,, \qquad q = 1, 2\,.$$

Consequently, the expected number of the nonzero cells $\mathbf{E}n_h(\underline{y})$, $h \in (0, 1)$, as well as the expected value of the total frequency $\mathbf{E}\|\underline{y}\|_1$ is much smaller than n. Thus, the contingency table $\underline{y}$ contains "a lot of zeros". We refer to Khmaladze (1988) for related notions.

Let us assume for simplicity that a simple hypothesis

$$H_0 : \underline{\mu} = \underline{\mu}^0 \quad \text{versus} \quad H_1 : \underline{\mu} \ne \underline{\mu}^0 \tag{5}$$

with a given $\mu^0 = (\mu_1^0, \dots, \mu_n^0) \in \mathcal{M}_+$, $\mathcal{M}_+ := \mathcal{M} \cap (0, \infty)^n$, is to be tested on the basis of the observed frequencies $\underline{y}$. Consider the *logarithmic likelihood ratio (LLR) statistic*

$$G^2 = G^2(\underline{\mu}^0, \underline{y}) := 2\sum_{j \in J} \left[y_j \log\left(\frac{y_j}{\mu_j^0}\right) + (\mu_j^0 - y_j) \right] =: 2H(\underline{y}) + 2L(\underline{\mu}^0, \underline{y})\,,$$

$$H(\underline{y}) := \sum_{j \in J} y_j \log(y_j)\,,$$

$$L(\underline{\mu}^0, \underline{y}) := \mu_+^0 - \sum_{j \in J} y_j (\log(\mu_j^0) + 1)\,, \qquad \mu_+^0 := \sum_{j \in J} \mu_j^0 = \|\underline{\mu}\|_1\,.$$

It turns out that for sparse data the term $L(\underline{\mu}_0, \underline{y})$ often dominates $H(\underline{y})$.

Lemma 1. *Assume sparsity* (1). *Then* ($n \to \infty$)

$$\begin{aligned}
\mathbf{E}_{\underline{\mu}} G^2(\underline{\mu}^0, \underline{y}) &= 2\mathbf{E}_{\underline{\mu}} L(\underline{\mu}^0, \underline{y}) + O\left(\|\underline{\mu}\|_2^2\right) , \\
\mathrm{StDev}_{\underline{\mu}} G^2(\underline{\mu}^0, \underline{y}) &= 2\mathrm{StDev}_{\underline{\mu}}(L(\underline{\mu}^0, \underline{y})) + O\left(\|\underline{\mu}\|_2\right) , \\
\mathbf{E}_{\underline{\mu}} L(\underline{\mu}^0, \underline{y}) &= \mu_+^0 - \sum_{j \in J} \mu_j (\log(\mu_j^0) + 1) , \\
\mathrm{Var}_{\underline{\mu}}(L(\underline{\mu}^0, \underline{y})) &= \sum_{j \in J} \mu_j (\log(\mu_j^0) + 1)^2 .
\end{aligned}$$

To prove the lemma it suffices to note that for any $\beta > 0$

$$\mathbf{E}_{\underline{\mu}} \sum_{j \in J} (y_j \log(y_j))^\beta = O(\|\underline{\mu}\|_2^2) .$$

Proposition 1. *Suppose that* $\underline{\mu}^0 \in \mathcal{M}_+$, $\underline{\mu} \in \mathcal{M}$,

$$\Delta_n = \Delta_n(\underline{\mu}) := \sum_{i \in J} (\mu_j - \mu_j^0)(\log(\mu_j^0) + 1) \geq 0 , \tag{6}$$

and

$$\|\underline{\mu}\|_2^2 + \|\underline{\mu}^0\|_2^2 = o(D_n^2(\underline{\mu}) + D_n^2(\underline{\mu}^0)) , \qquad D_n^2(\underline{\mu}) := \sum_{i \in J} \mu_j \left(\log(\mu_j^0) + 1\right)^2 . \tag{7}$$

If (1) holds, then

$$\frac{\mathbf{E}_{\underline{\mu}} G^2(\underline{\mu}^0, \underline{y}) - \mathbf{E}_{\underline{\mu}^0} G^2(\underline{\mu}^0, \underline{y})}{\left(\mathrm{Var}_{\underline{\mu}^0} G^2(\underline{\mu}^0, \underline{y}) + \mathrm{Var}_{\underline{\mu}} G^2(\underline{\mu}^0, \underline{y})\right)^{1/2}} = -\frac{\Delta_n + O(\|\underline{\mu}\|_2^2 + \|\underline{\mu}^0\|_2^2)}{\left(D_n^2(\underline{\mu}) + D_n^2(\underline{\mu}^0)\right)^{1/2} (1 + o(1))} .$$

Corollary 1. *For very sparse contingency tables (see (1)), the LR test is inconsistent for testing problem (5) provided (6) and (7) hold and*

$$\|\underline{\mu}\|_2^2 + \|\underline{\mu}^0\|_2^2 = o(\Delta_n) , \qquad D_n^2(\underline{\mu}^0) + D_n^2(\underline{\mu}) \leq (\kappa + o(1))\Delta_n^2, \ \kappa < 1 .$$

When $\kappa = 0$ *we obtain the "reversed consistency": the probability to reject* H_1 *tends to 0 as* $n \to \infty$.

Example. Let $n = 2\tilde{n}$, $\mu_{0i} = \mu_{0i}(n) = o(1)$, $i = 1, 2$; $0 < \mu_{01} < \mu_{02}$, $\rho \in (0, 1)$, and

$$\mu_j^0 = \mu_{01}, \ \forall j \leq \tilde{n}, \qquad \mu_j^0 = \mu_{02}, \ \forall j > \tilde{n},$$

$$\mu_j = (1 - \rho)\mu_{01}, \ \forall j \leq \tilde{n} , \qquad \mu_j = \mu_{02} + \rho\mu_{01}, \ \forall j > \tilde{n} .$$

Then

$$\Delta_n = \frac{\rho\mu_{01}n}{2}\log\left(\frac{\mu_{02}}{\mu_{01}}\right) > 0\,,$$

$$D^2(\mu^0) \asymp D^2(\mu) \asymp n\mu_{02}(\log(\mu_{02}))^2\,.$$

Note that $\|\underline{\mu}^0\|_1 = \|\underline{\mu}\|_1$ and $\|\underline{\mu}^0\|_2^2 + \|\underline{\mu}\|_2^2 \le (\mu_{01} + 2\mu_{02})\ \|\underline{\mu}^0\|_1 = o\left(\|\underline{\mu}^0\|_1\right)$. Thus, the conditions of Corollary are fulfilled if $\mu_{01} \le \rho_1\mu_{02}$, $\rho_1 \in (0,1)$,

$$\mu_{02}^2 = o\left(\mu_{01}|\log(\mu_{01})|\right)\,,$$

$$\frac{\sqrt{\mu}_{02}|\log\left(\mu_{02}\right)|}{\mu_{01}\sqrt{n}} = o(1)\,.$$

Remark 2. Actually, the inconsistency stated in Corollary 1 is not an exceptional feature of the 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

Let us assume that $\{J_m, m = 1,\dots,M\}$ is a partition of J into disjoint subsets such that $\mu_j^0 = \mu_{0m}$, $j \in J_m$, $m = 1,\dots,M$, with some $\mu_{0m} = \mu_{0m}(n) \in (0, M_0]$. Suppose that all alternatives with any $\underline{\mu}$ obtained via permutations of the coordinates within J_m, $m = 1,\dots,M$, are equally likely to occur. Then it is natural to assume that the tests under consideration are invariant with respect to permutations of the coordinates in J_m. This assumption is consistent with the Bayes approach which assumes $\underline{\mu}$ to be a sequence of random variables exchangeable within the each set J_m.

Following the empirical Bayes approach, the parameter $\underline{\mu}$ is treated as random and

$$\{\mu_j, j \in J_m\} \text{ are i.i.d.}\,, \qquad \mu_j \sim G_m\,, \quad j \in J_m\,, \quad m = 1,\dots,M\,.$$

Here $G_m = G_m(\cdot|n)$ are unknown distributions on $[0, M_0]$. Thus, the unknown parameters $\underline{\mu} = (\mu_1,\dots,\mu_n)$ are replaced with the unknown distributions $G = (G_1,\dots,G_M)$. Let

$$\pi_l(G_\ell) := \int_0^{M_0} \pi_l(u)\,dG_\ell(u)\,, \qquad \pi_l(u) := \frac{u^l e^{-u}}{l!}\,, \quad l \in \mathbf{Z}_+\,.$$

In this setting, the null hypothesis in (5) can be restated as follows:

$$H_0^G : G_m = \delta_{\mu_{0m}}\,, \qquad m = 1,\dots,M\,. \tag{8}$$

Here δ_a stands for the degenerate distribution centered on a. The LLR statistic for (8) is given by

$$\ell(G) = 2\sum_{m=1}^{M}\sum_{l\in\mathbf{Z}_+}\eta_l(m)\log\left(\frac{\pi_l(G_m)}{\pi_l(\mu_{0m})}\right)\,,$$

$$\eta_l(m) := |\{y_j, j \in J_m : y_j = l\}|\,. \tag{9}$$

Hence the statistic $\underline{\eta} = \{\eta(m), m = 1, \ldots, M\}$ with $\eta(m) := (\eta_l(m), l \in \mathbf{Z}_+)$, $m = 1, \ldots, M$, is a sufficient statistic for G. Under the Poisson sampling, $\underline{\eta}$ distribution is a product of M multinomial distributions with the infinite number of outcomes, the probabilities of outcomes equal to $\pi_{\mathbf{Z}_+}(\mu_{0m}) := (\pi_l(\mu_{0m}),\ l \in \mathbf{Z}_+)$, and $n_m := |J_m|$ independent trials ($m = 1, \ldots, M$):

$$\eta(m) \sim Multinomial_{\mathbf{Z}_+}(n_m, \pi_{\mathbf{Z}_+}(\mu_{0m}))\,.$$

Components (9) of the statistic $\underline{\eta}$ are called the *profile statistics* of the contingency table. Sometimes they are also referred to as the spectral statistics or frequencies of frequencies. The asymptotic behavior of η_m in the case of multinomial sampling have been investigated, for instance, by Kolchin, Sevastyanov, and Chistyakov (1978). The profile statistics are also related to estimating problem of the structural distribution function of cell probabilities (van Es, Klaassen, and Mnatsakanov, 2003).

Let $\hat{G}$ denote the (nonparametric) maximum likelihood estimator of $G = (G_1, \ldots, G_m)$ (van de Geer, 2003). The inequality given in the next proposition allows one to obtain a conservative critical value for the LLR statistic $\ell(\hat{G})$.

Given $s \in \mathbf{N}$, denote

$$K(s) := \{z = (n - h, z_1, \ldots) \in \mathbf{Z}_+^\infty : h := z_1 + \ldots + z_s \le s;\ z_l = 0,\ \forall l > s\}\,.$$

Proposition 2. *Suppose that $\underline{\mu}^0 \in \mathcal{M}_+$ satisfies sparsity condition (1) and $\mu_j^0 = \mu_{0m}$, $j \in J_m$, $m = 1, \ldots, M$. Then, for any $t = t(n)$, $t/\log(t) > \max(M_0, \|\underline{\mu}^0\|_1)$,*

$$\mathbf{P}_{\underline{\mu}^0}\{\ell(\hat{G}) \ge t\} \le H(t)e^{-t/2} \tag{10}$$

where

$$\begin{aligned} H(t) := &\left| K\left(\left[\frac{t}{\log(t)}\right] + 1\right)\right|^M + n\exp\left\{\frac{t(\log\log(t) + \log(M_0) + 1)}{\log(t)}\right\} \\ &+ \exp\left\{\frac{t(\log(\|\underline{\mu}^0\|_1) + \log\log(t) + 1)}{\log(t)} - \|\underline{\mu}^0\|_1\right\} \end{aligned} \tag{11}$$

and $\log(H(t)) = o(t)$ *provided* $\log(n) = o(t)$.

Proof. Since

$$\ell(\hat{G}) \le \hat{\ell}(\underline{\eta}) := 2\sum_{m=1}^{M}\sum_{l\in\mathbf{Z}_+} \eta_l(m)\log\left(\frac{\eta_l(m)}{n_m\,\pi_l(\mu_{0m})}\right)\,,$$

the inequality

$$\mathbf{P}_{\underline{\mu}^0}\{\ell(\hat{G}) \ge t\} \le \mathbf{P}_{\underline{\mu}^0}\{\hat{\ell}(\underline{\eta}) \ge t\} \tag{12}$$

holds. For $s \in \mathbf{N}$, let $k(m) = (k_l(m),\ l \in \mathbf{Z}_+) \in K(s)$ with $n_m = k_+(m) := \sum_{l=0}^\infty k_l(m)$, $m = 1, \ldots, M$, and $\underline{k} := (k(m), m = 1, \ldots, M)$. Then using Sanov (1957) arguments we obtain the inequality

$$\mathbf{P}_{\underline{\mu}^0}\{\underline{\eta} = \underline{k}\} \le \exp\{-(1/2)\ \hat{\ell}(\underline{k})\}\,.$$

Introduce $\underline{\eta}^+ = (\eta_l^+,\ l \in \mathbf{Z}_+)$ where $\eta_l^+ := \sum_{m=1}^M \eta_l(m),\ l \in \mathbf{Z}_+$. Notice that $\underline{\eta}^+ \in K(s)$ implies $\eta(m) \in K(s),\ \forall\, m = 1, \ldots, M$. Therefore

$$\mathbf{P}_{\underline{\mu}^0}\{\hat{\ell}(\underline{\eta}) \geq t\} \leq |K(s)|^M \exp\{-t/2\} + \mathbf{P}_{\underline{\mu}^0}\{\underline{\eta}^+ \notin K(s)\}\,. \tag{13}$$

Denote

$$\eta_+ := \sum_{l=1}^{\infty} \eta_l^+ = \sum_{j \in J} \mathbf{1}\{y_j > 0\} \leq \|\underline{y}\|_1\,.$$

Since $\|\underline{y}\|_1 \sim Poisson(\|\underline{\mu}^0\|_1)$, for $s > \|\underline{\mu}^0\|_1$,

$$\log\Big(\mathbf{P}_{\underline{\mu}^0}\{\eta_+ > s\}\Big) \leq s \log\left(\frac{\|\underline{\mu}^0\|_1}{s}\right) + s - \|\underline{\mu}^0\|_1\,. \tag{14}$$

Similarly, for $s > M_0$,

$$\mathbf{P}_{\underline{\mu}^0}\left\{\max_{j \in J} y_j > s\right\} \leq \sum_{j \in J} \mathbf{P}_{\underline{\mu}^0}\{y_j > s\} \leq n\ \exp\left\{s \log\left(\frac{M_0}{s}\right) + s\right\}\,. \tag{15}$$

Note that $\eta_+ \leq s$ and $\max_{j \in J} y_j \leq s$ imply $\underline{\eta}^+ \in K(s)$. Hence,

$$\mathbf{P}_{\underline{\mu}^0}\{\underline{\eta} \notin K(s)\} \leq \mathbf{P}_{\underline{\mu}^0}\{\eta_+ > s\} + \mathbf{P}_{\underline{\mu}^0}\{\max_{j \in J} y_j > s\}\,. \tag{16}$$

Take $s = [t/\log(t)] + 1$. Then inequality (10) with $H(t)$ given in (11) follows from (12)–(16). The well-known fact that $\log|K(s)| = O(s)$ as $s \to \infty$ completes the proof.

The proposed LR criterion based on the profile statistics can be viewed as a composite LR test for homogeneous groups of cells obtained via hard clustering. In the next section a flexible and adaptive procedure taking advantage of soft clustering in an auxiliary mixture model is described.

4 Likelihood Ratio Test with Soft Clustering

Here it is assumed that the both parameters, $\underline{\mu}^0$ and $\underline{\mu}$, are sequences of independent identically distributed random variables satisfying a semi-parametric mixture model with a dummy class variable $\nu_j \in \{1, \ldots, M\}$, $j \in J$. Specifically,

$$\mathbf{P}\{\nu_j = m\} = p_m \geq 0\,, \qquad \sum_{m=1}^{M} p_m = 1\,; \tag{17}$$

$$(\mu_j^0 \mid \nu_j = m) \sim LogNormal(a_m, \sigma_m)\,, \qquad j \in J\,, \tag{18}$$

$$(\mu_j \mid \nu_j = m) \sim G_m\,, \qquad j \in J\,, \tag{19}$$

$$(y_j \mid \mu_j) \sim Poisson(\mu_j)\,, \qquad j \in J\,,\ m = 1, \ldots, M\,. \tag{20}$$

Let

$$\theta := (p_m, a_m, \sigma_m, G_m\,,\ m = 1, \ldots, M)$$

be a collection of the parameters of the mixture. Notice that the values of $\underline{\mu}$ are unobservable (latent). The observed data is (y_j, μ_j^0), $j \in J$. Suppose that μ_j^0 and μ_j are conditionally, given ν_j, independent, and y_j, given μ_j, is independent of the rest random variables ($j \in J$). Thus, the parameter θ completely specifies the distribution of the observed data.

The (nonparametric) maximum likelihood method is applied to fit the model to data. Let $\hat{\theta} := (\hat{p}_m, \hat{a}_m, \hat{\sigma}_m, \hat{G}_m\,, m = 1, \dots, M)$ be the maximum likelihood estimator of θ. Obviously, the number of the support points of $\hat{G}_m$ does not exceed $y_{max} := \max_{j\in J} y_j$. For sparse data, y_{max} is small. Thus, the probabilities $\pi_l(\hat{G}_m)\,, l \in \mathbf{Z}_+$, are expressed as the finite mixture of Poisson distributions. Consequently, the initial semi-parametric model defined in (17)–(20) can be approximated and, actually, replaced by a parametric finite mixture model. In order to calculate the maximum likelihood estimator of its parameters, the EM algorithm is used.

Let $p_m(\hat{\theta} \mid y_j, \mu_j^0)$ be the estimated posterior probability of the unobserved class number ν_j, given the observation (y_j, μ_j^0),

$$p_m(\theta \mid y_j, \mu_j^0) := \mathbf{P}_\theta\{\nu_j = m \mid y_j, \mu_j^0\}\,, \qquad j \in J\,,\ m = 1, \dots, M\,.$$

For $m = 1, \dots, M$ and $l \in \mathbf{Z}_+$, set

$$\hat{\eta}_l(m) := \sum_{j\in J} \mathbf{1}\{y_j = l\} p_m(\hat{\theta} \mid y_j, \mu_j^0)\,,$$
$$\hat{\pi}_l^0(m) := \sum_{j\in J} \pi_l(\mu_j^0) p_m(\hat{\theta} \mid y_j, \mu_j^0)\,.$$

The symmetric LLR statistic based on soft clustering and the empirical Bayes approach is defined by

$$\mathcal{L}(\hat{\theta} \mid \underline{y}) := \sum_{m=1}^{M} \sum_{l=1}^{y_{max}} (\hat{\eta}_l(m) - \hat{\pi}_l^0(m)) \Big(\log(\pi_l(\hat{G}_m)) - \log(\pi_l(\exp\{\hat{a}_m\}))\Big)\,. \qquad (21)$$

The performance of the criterion for testing (5) based on $\mathcal{L}(\hat{\theta} \mid \underline{y})$ is illustrated by simulations.

Computer experiment. The framework of the example in Section 1 is adopted. The parameters $\mu_{01} = 0.5$, $\mu_{02} = 1.0$, $\mu_{11} = \mu_{11}(i) = \mu_{01} - 0.05(i-1)$, $\mu_{12} = \mu_{12}(i) = \mu_{02} + 0.05(i-1)$, $i = 1, \dots, 10$, $n = 2\tilde{n} = 40$, the number of simulations is equal to 100. The parameters σ_m are kept fixed, $\sigma_m = 0.5$, $m = 1, \dots, M$. The number of clusters $M = 4$, the maximal number of support points of $\hat{G}_m$ is set to 5.

A critical value for LLR statistic (21) is evaluated by the Monte Carlo method.

The estimated powers of the classical LR test and the proposed criterion based on the statistic $\mathcal{L}$ are presented in Figure 1. The significance level $\alpha = 0.05$. The index $i > 1$, indicates the number of an alternative. The case $i = 1$ corresponds to the null hypothesis. In fact, the power of the proposed test is close to the power of χ^2 test with the additional prior information $\mu_j = \mu_{11}$, $\forall j \le \tilde{n}$, $\mu_j = \mu_{12}$, $\forall j > \tilde{n}$, μ_{11} and μ_{12} are unknown.

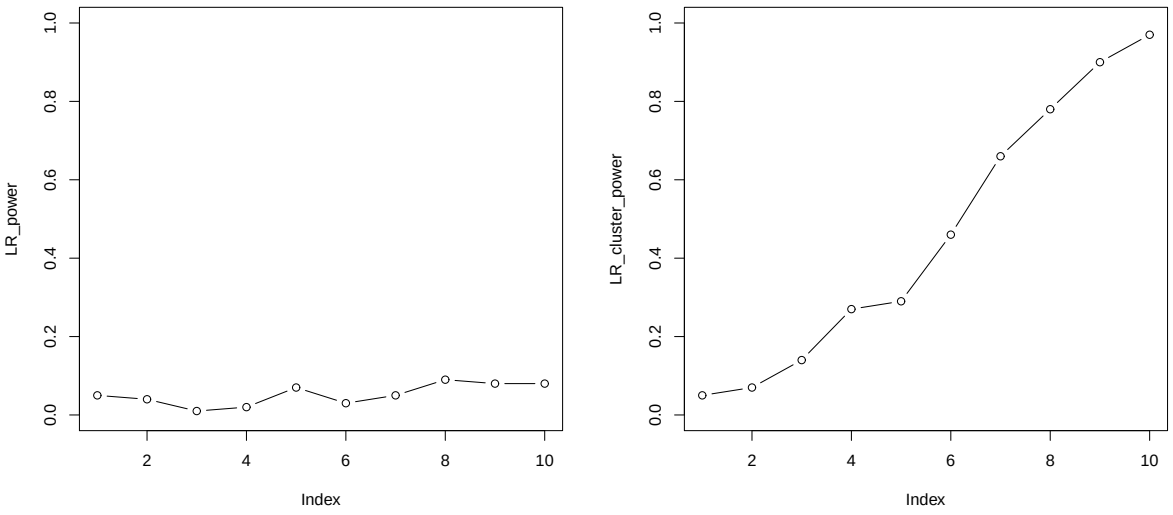

Figure 1: The power of the classical LR test (left) and the test based on $\mathcal{L}$ (right) for alternatives $i = 2, \ldots, 10$.

References

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

Agresti, A. (1999). Exact inference for categorical data: recent advances and continuing controversies. *Statistical Methods and Applications*, *20*, 2709-2722.

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

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

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.

Cressie, N., and Read, T. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society*, *46*, 440-464.

Hu, M. Y. (1999). *Model Checking for Incomplete High Dimensional Categorical Data*. Unpublished doctoral dissertation, University of California, Los Angeles.

Khmaladze, E. (1988). *The statistical analysis of a large number of rare events* (Report No. MS-R8804). Amsterdam.

Kolchin, V. F., Sevastyanov, B., and Chistyakov, V. (1978). *Random Allocations*. New York: Wiley.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, *21*, 3789-3801.

Müller, U. U., and Osius, G. (2003). Asymptotic normality of goodness-of-fit statistics for sparse Poisson data. *Statistics*, *37*, 119-143.

Sanov, I. (1957). On the probability of large deviations of random magnitudes. *Mat. Sb. N. S.*, *42*, 11-44.

Smirnoff, J. S. (1995). Smoothing categorical data. *Journal of Statistical Planning and Inference*, *47*, 41-69.

van de Geer, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*, *41*, 453-464.

van Es, B., Klaassen, C. A. J., and Mnatsakanov, R. M. (2003). Estimating the structural distribution function of cell probabilities. *Austrian Journal of Statistics*, *32*, 85-98.

von Davier, M. (1997). *Bootstraping goodness-of-fit statistics for sparse categorical data. Results of a Monte Carlo study* (Vol. 2). Available from `http//www.pabst-publishers.de/mpr/`

Authors' addresses:

Marijus Radavičius
Department of Applied Statistics
Institute of Mathematics and Informatics
Akademijos 4
LT-08663 Vilnius
Lithuania

E-mail: `mrad@ktl.mii.lt`

Pavel Samusenko
Department of Mathematical Statistics
Faculty of Fundamental Sciences
Vilnius Gediminas Technical University
Sauletekio al. 11
Vilnius
Lithuania

E-mail: `PavelS.vgtu@gmail.com`