# Goodness-of-fit tests for sparse nominal data based on grouping

**Marijus Radavičius**[a], **Pavel Samusenko**[b]

[a]Vilnius University
Naugarduko str. 24, LT-03225 Vilnius, Lithuania
marijus.radavicius@mii.vu.lt

[b]Vilnius Gediminas Technical University
Saulėtekio ave. 11, LT-10223 Vilnius, Lithuania
pavels.vgtu@gmail.com

**Received:** 24 July 2012 / **Revised:** 17 September 2012 / **Published online:** 22 November 2012

**Abstract.** For (very) sparse nominal data, common goodness-of-fit tests usually fail. Alternative goodness-of-fit tests based on extended empirical Bayes approach and grouping are proposed and their consistency is proved. The performance of the tests is illustrated and compared with classical criteria by Monte Carlo simulations.

**Keywords:** contingency table, empirical Bayes, consistency, goodness-of-fit, grouping, latent distribution, sparse asymptotics, structural distribution.

## 1 Introduction

Currently the amount of accessible information is very extensive, therefore problems related to a high dimensionality of data arise rather frequently. For quantitative (continuous) variables, (generalized) *linear models* are usually applied. They describe relationships between the means of these variables or their covariance structures and hence the number of model parameters grows at most as $O(k^2)$ with respect to the dimensionality $k$ of the data. The problem of high dimensionality is especially topical for qualitative (categorical) variables. In this case, the number of model parameters generally *increases exponentially* with $k$. Consequently, even for a moderate number of categorical variables, a corresponding contingency table can be *sparse*, i.e. many cells in the table are empty or have small counts. In fact, for categorical data, the number of cells in the corresponding contingency table is even more important characteristic of sparsity than the dimensionality $k$ itself. Sometimes the number of cells (the number of unknown parameters) is even greater than the sample size (very sparse categorical data).

**Example.** (Cf. [1, p. 16, Case 3].) Suppose a questionnaire consists of $k = 10$ questions, each with 2 possible answers. Then the total number of cells in a contingency table of the answers is $2^k = 2^{10} > 10^3$. Thus, for a sample with $10^3$ respondents, the average of expected frequencies in the contingency table is less than 1.

According to the *rule of thumb* 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 $\chi^2$ approximations of goodness-of-fit statistics may be *inaccurate* and the table is said to be *sparse* [2].

Examples of real sparse categorical data along with their statistical analysis and discussion can be found in ( [1, p. 3], [2, p. 149], [3, p. 3]).

Actually, there are three main problems caused by sparsity in statistical analysis of contingency tables:

1. The standard $\chi^2$ approximation for distributions of classical tests is not sufficiently accurate (see, e.g., [2, 4]). Several techniques have been proposed to tackle this problem: exact tests [2], alternative approximations [5, 6] parametric and nonparametric bootstrap [7], Bayes approach [8, 9] and other methods.
2. The classical tests are not longer (asymptotically) distribution free [1]. The latter property for test implies that the test performance is independent of a null hypothesis to be tested and thus leads to universal decision rules. The lack of this property means that a critical value of every testing problem is a specific problem to be solved.
3. For (very) sparse data, the classical tests become noninformative: they do not anymore measure the goodness-of-fit of a null hypothesis to data. For instance, the classical tests are inconsistent even in cases where a simple consistent test does exist ( [10, 11], see also [1, 12].

The paper is devoted to the third problem. It reveals that possibly there is no sense to solve the former two problems. The goal of the paper is to propose alternative nonparametric criteria to the classical ones which are consistent for sparse categorical (nominal) data as well.

In the next section, we present a brief overview of different approaches to sparsity. We propose the extended empirical Bayes model of sparse asymptotics. This model contains the latent distribution and the structural distribution models as special cases. In Section 3, testing problem is formulated without any assumptions about convergence of distributions. The consistency of tests based on $\phi$-divergences and grouping is proved. Finite-sample performance of these tests is studied using Monte Carlo simulations in Section 4. The proposed tests are compared with the classical criteria.

## 2 Definitions of sparsity

Let $\mathbf{y} := (y_1, \dots, y_n)$ be a contingency table, i.e. a vector of observed frequencies. Set $\mu = \mathbb{E}\mathbf{y}$. Assume that components of $\mathbf{y}$ are independent Poisson random variables,

$$\mathbf{y} \sim \mathrm{Poisson}(\mu).$$

An alternative assumption might be

$$\mathbf{y} \sim \mathrm{Multinomial}_n(N, \mathbf{p}), \quad \mathbf{p} := \frac{\mu}{N}. \tag{1}$$

Consider a simple hypothesis testing problem

$$\mathrm{H}_0\colon \mu = \mu^\circ \quad \text{versus} \quad \mathrm{H}_1\colon \mu \neq \mu^\circ, \tag{2}$$

where $\mu^\circ := (\mu_1^\circ, \dots, \mu_n^\circ)$ is a given vector of positive values.

We are interested in case where contingency tables are *sparse*. Informally it means that the number of cells $n$ is large and expected frequencies of a significant part of cells are small.

There are different ways to define sparsity formally, as well as represent sparsity scale by introducing the corresponding parameters. The definition of sparsity is based on the *sparse asymptotics* (cf. [13, 14]). Denote $\mu_+ := \mathbf{E}y_+$, $y_+ := \sum_{j=1}^n y_j$.

Let $M \to \infty$ be some asymptotic parameter. The sparse asymptotics assumes that $n = n(M) \to \infty$ and $\mu_+ = \mu_+(M) \to \infty$ as $M \to \infty$. In what follows we usually hide the dependence on the asymptotic parameter $M$ though indicate it when introducing new objects and in cases we need to stress this dependence.

### 2.1 Latent distribution model

One of the simplest way to deal with the sparsity is to suppose that the expected frequencies $\mu = (\mu_1, \dots, \mu_n)$ of an *ordered* variable are determined by a *latent distribution function* $F$ on $[0, 1]$ via representation

$$\mu_i = \mu_+\bigl(F(t_i) - F(t_{i-1})\bigr), \tag{3}$$

where $t_0 = 0$, $t_i := i/n$, $i = 1, \dots, n$ (cf. [13, 15]). In this setting, it is usually assumed that there exists rather smooth *latent distribution density* $f$, $f(u) = \mathrm{d}F(u)/\mathrm{d}u$. This assumption implies

$$\mu_i = \mu_i(M) = O\left(\frac{\mu_+}{n}\right), \quad M \to \infty.$$

Thus, in this case the sparsity is expressed by the average expected frequency $\rho = \rho(M) := \mu_+/n$. For multinomial sampling scheme (1) we have $\mu_+ = N$ where $N$ is the sample size of the contingency table $\mathbf{y}$. Hence $\rho(M) = N/n$. A typical assumption for the sparse asymptotics is $\rho = O(1)$. In this case, the number of unknown parameters $n-1$ is proportional to $N$ and hence the consistent estimator of the parameters, in general, does not exist (see, e.g., [16]). The consistent estimator can be constructed under the additional requirements on smoothness of the latent distribution density $f$. Then standard (kernel) smoothing technique can be applied (see, e.g., [13, 15]).

The latent distribution model (3) with uniform with respect to $M$ restrictions on the smoothness of the latent density $f$ is inappropriate for *nominal data*. In this case, the expected frequencies $\mu$ and their sparsity can be described by the *structural distribution function* introduced by Khmaladze [1] to characterize data with a *large number of rare events* (LNRE for short; see also [12, 17]). Thus, LNRE is Khmaladze's definition of sparse categorical data.

### 2.2 Structural distribution

When dealing with testing problem (2), one can suppose that the cell numbering order is irrelevant. It means that the statement $\mu = \mu^\circ$ is replaced by the statement $\{\mu_1, \ldots, \mu_n\} = \{\mu_1^\circ, \ldots, \mu_n^\circ\}$. Actually, it is the same as to require the tests to be invariant with respect to permutations of the cell numbers. Then only permutation invariant hypotheses can be tested. This leads to the testing problem

$$\mathrm{H}_0\colon\ \hat{F}^{(M)} = (F^\circ)^{(M)} \quad \text{versus} \quad \mathrm{H}_1\colon\ \hat{F}^{(M)} \neq (F^\circ)^{(M)}, \tag{4}$$

where $\hat{F}^{(M)}$ is the empirical distribution function of $\{\mu_1, \ldots, \mu_n\}$,

$$\hat{F}^{(M)}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[\mu_i \le u]}, \quad u \in \mathbb{R}_+ := [0, \infty),$$

and $(F^\circ)^{(M)}$ is a given discrete distribution function with $|\mathrm{supp}((F^\circ)^{(M)})| \le n = n(M)$. $|A|$ denotes the number of elements (cardinality) of the set $A$.

Here we explicitly indicate the dependence of the statements on $M$, the key parameter in the sparse asymptotics.

In fact, testing problem (4) as well as (2) is a sequence of statements and it remains some uncertainty how they should be combined. While it is quite natural to take "$\mathrm{H}_0$: $\mu^{(M)} = (\mu^\circ)^{(M)}$ $\forall$ (sufficiently large) $M$", a reasonable definition of $\mathrm{H}_1$ is not so clear. Using ideas of the contiguous alternative approach, the testing problem is expressed through asymptotic characteristics (parameters) of sample distributions.

**Definition 1.** (Cf. [17].) *Suppose that $F_\rho(t) := \hat{F}^{(M)}(\rho t)$ with some scaling factor $\rho = \rho(M)$ converges weakly to some distribution function $F$ as $M \to \infty$. Then $F$ is called a* structural distribution *of the expected cell frequencies $\mu$ (or simply of the table $\mathbf{y}$) with the scaling factor $\rho$.*

In terms of the structural distribution the testing problem states

$$\mathrm{H}_0\colon\ F = F^\circ \quad \text{versus} \quad \mathrm{H}_1\colon\ F \neq F^\circ, \tag{5}$$

where $F^\circ$ is a given distribution function with $\mathrm{supp}(F^\circ) \subset \mathbb{R}_+$. Again, the sparsity scale is determined by $\rho$.

Khmaladze [1] pointed out that the structural distribution can be treated as a latent mixing distribution in the empirical Bayes approach. Below we extend this approach to include the null hypothesis in the Bayes model as well.

### 2.3 Extended empirical Bayes model

Let us suppose that $\{(\mu_i^\circ, \mu_i),\ i = 1, \ldots, n\}$ are independent copies of a random pair $(\gamma^\circ, \gamma)$ taking values in $\mathbb{R}_+^2$ and having distribution $P = P^{(M)}$, $M \to \infty$. Then the marginal distribution $P^\circ$ of $\gamma^\circ$ (respectively, $P^\gamma$ of $\gamma$) coincides with the structural distribution under the null hypothesis $\mathrm{H}_0$ (respectively, under the alternative $\mathrm{H}_1$), see [1].

Fix $M$ or set $M = \infty$. Now the testing problem for structural distribution (5) takes the following form: $\mathrm{H}_0$: $P^\circ = P^\gamma$ versus $\mathrm{H}_1$: $P^\circ \neq P^\gamma$. Thus in this case only the marginal distributions of $P$ are involved.

Let $P^\gamma_\circ$ denote the conditional distribution of $\gamma$ given $\gamma^\circ$:

$$P^\gamma_\circ(\cdot \mid a) := \mathbb{P}\{\gamma \in \cdot \mid \gamma^\circ = a\}, \quad a \in \mathbb{R}_+.$$

Then problem (2) can be extended in terms of $P$ as follows:

$$\mathrm{H}_0\colon\ P^\gamma_\circ(\cdot \mid a) = \delta_a \ \ \forall a \in \Omega \quad \text{versus} \quad \mathrm{H}_1\colon\ P^\gamma_\circ(\cdot \mid a) \neq \delta_a \ \ \forall a \in A.$$

Here $\delta_a$ is the Dirac measure with the support $\{a\}$, $a \in \mathbb{R}_+$, $\Omega$ and $A$ are some measurable sets satisfying, respectively, $P^\circ(\Omega) = 1$ and $P^\circ(A) > 0$.

Note that this extension of (2) can not be tested using the latent distribution model, nor the structural distribution approach. They both suggest some convergence of distributions as $M \to \infty$, i.e. some regularity in the sparse asymptotics of frequency tables. In the next section the testing problem is formulated without any assumptions about convergence of distributions thus providing more flexibility in applications.

## 3 Hypotheses testing under the sparsity

Here we use the extended empirical Bayes framework described in 2.3.

Let $\mathcal{P} = \mathcal{P}^{(M)}$ be a class of probability distributions $P = P^{(M)}$ on $\mathbb{R}^2_+ = \mathbb{R}_+ \times \mathbb{R}_+$, hypothetical distributions of the random pair $(\gamma, \gamma^\circ)$.

Suppose that a discrepancy measure $d(P, Q) = d^{(M)}(P, Q)$ between probability distributions $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ satisfies conditions: $d(P, Q) \geq 0$, $d(Q, Q) = 0$.

Given $Q^{(M)} \in \mathcal{P}^{(M)}$ and $\delta = \delta(M) > 0$, consider the following testing problem:

$$\mathrm{H}_0\colon\ \forall M,\ d\bigl(Q^{(M)}, P^{(M)}\bigr) = 0, \tag{6}$$

versus

$$\mathrm{H}_1\colon\ \forall M,\ d\bigl(Q^{(M)}, P^{(M)}\bigr) \geq \delta(M). \tag{7}$$

Our proofs of the consistency of testing criteria are based on a general result given below.

### 3.1 Main lemma

Given $P^{(M)} \in \mathcal{P}^{(M)}$ for all $M$, let $\mathbb{P}_P = \mathbb{P}_P^{(M)}$ denote the probability distribution of an observed data $\mathcal{D}^{(M)}$ generated by making use of $P^{(M)}$. Let $Q(M) \in \mathcal{P}^{(M)}$ be a hypothetical distribution generating $\mathcal{D}^{(M)}$.

**Assumption C.** Assume that for a given $\delta = \delta(M) > 0$, there exist an estimator $\widehat{d(Q;P)}$ of $d(Q^{(M)}; P^{(M)})$ and $\tau = \tau(M) \in (0,1)$ such that

$$\mathbb{P}_Q^{(M)}\bigl\{\widehat{d(Q;P)} > \bigl(1 - \tau(M)\bigr)\delta(M)\bigr\} \to 0, \quad M \to \infty, \tag{8}$$

and for all $P^{(M)} \in \mathcal{P}^{(M)}$

$$\mathbb{P}_P^{(M)}\bigl\{\widehat{d(Q;P)} \leq d\bigl(Q^{(M)}; P^{(M)}\bigr) - \tau(M)\delta(M)\bigr\} \to 0, \quad M \to \infty. \tag{9}$$

**Lemma 1.** *Assume that Assumption C is valid. Then the criterion*

$$\mathcal{K} := \bigl\{\widehat{d(Q;P)} > \bigl(1-\tau(M)\bigr)\delta(M)\bigr\},$$

*is consistent as $M \to \infty$ for testing* (6) *versus* (7).

*Proof.* Write $\tau = \tau(M)$, $\delta = \delta(M)$ for short. If $\mathrm{H}_0$ is valid,

$$\mathbb{P}_Q^{(M)}(\mathcal{K}) = \mathbb{P}_Q^{(M)}\bigl\{\widehat{d(Q;P)} > (1-\tau)\delta\bigr\} \to 0, \quad M \to \infty,$$

due to (8). If $\mathrm{H}_1$ holds, then $d(Q^{(M)};P^{(M)}) \geq \delta$ and hence

$$\begin{aligned}1-\mathbb{P}_P^{(M)}(\mathcal{K}) &= \mathbb{P}_P^{(M)}\bigl\{\widehat{d(Q;P)} \leq (1-\tau)\delta\bigr\} \\ &\leq \mathbb{P}_P^{(M)}\bigl\{\widehat{d(Q;P)} \leq d\bigl(Q^{(M)};P^{(M)}\bigr) - \tau\delta\bigr\} \to 0, \quad M \to \infty,\end{aligned}$$

by (9). □

In order to apply Lemma 1 we need to specify the discrepancy measure $d$, the class $\mathcal{P}^{(M)}$ of distributions, the estimator $\widehat{d(Q;P)}$, and the critical value $(1-\tau(M))\delta(M)$ for sparse asymptotics $M \to \infty$.

### 3.2 Discrepancy measures

The $\phi$-divergence between two vectors $u, v \in \mathbb{R}_+^n$ is defined by (cf. [18])

$$d_\phi(v;u) := \sum_{i=1}^n v_i \phi\biggl(\frac{u_i}{v_i}\biggr).$$

The function $\phi : \mathbb{R}_+ \to \mathbb{R}$ is convex, strictly convex at 1, $\phi(1) = 0$. The most of $\phi$-divergences widely used to measure distribution discrepancy belong to *power-divergence family* (cf. [4]) with $\phi = \phi_\alpha$:

$$\phi_\alpha(t) := \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, \quad \alpha(\alpha-1) \neq 0, \tag{10}$$

$$\phi_1(t) := t\ln t - t + 1, \quad \phi_0(t) := -\ln t + t - 1. \tag{11}$$

For $\phi = \phi_\alpha$, denote $d_\alpha(v;u) := d_\phi(v;u)$. Taking $\alpha = 1$ and $\alpha = 2$ produce the classical logarithmic likelihood ratio and Pearson $\chi^2$ statistics, respectively,

$$G^2 := d_1(\mu^\circ;y) = 2\sum_{i=1}^n \biggl(y_i \log\frac{y_i}{\mu_i^\circ} - y_i + \mu_i^\circ\biggr),$$

$$X^2 := d_2(\mu^\circ;y) = \sum_{i=1}^n \frac{(y_i - \mu_i^\circ)^2}{\mu_i^\circ},$$

see also Table 1.

Table 1. Goodness-of-fit statistics.

| Statistic | $\alpha$ | Definition |
|---|---|---|
| Small Samples | $5/3$ | $9/5\sum_{i=1}^{k} y_i((y_i/\mu_i^\circ)^{2/3}-1)$ |
| Hellinger Distance | $1/2$ | $4\sum_{i=1}^{k}(\sqrt{y_i}-\sqrt{\mu_i^\circ})^2$ |
| Likelihood Ratio modified | $0$ | $2\sum_{i=1}^{k}(\mu_i^\circ\log(\mu_i^\circ/y_i)+(y_i-\mu_i^\circ))$ |
| $\chi^2$ modified | $-2$ | $\sum_{i=1}^{k}(\mu_i^\circ-y_i)^2/y_i$ |
| Likelihood Ratio symmetrized | | $\sum_{i=1}^{k}(y_i-\mu_i^\circ)\log(y_i/\mu_i^\circ)$ |
| $\chi^2$ symmetrized (Le Cam) | | $2\sum_{i=1}^{k}(y_i-\mu_i^\circ)^2/(y_i+\mu_i^\circ)$ |

However, classical test statistics usually are not appropriate for testing goodness-of-fit in case of sparse contingency tables or LNRE data [1, 10, 11]. A special grouping procedure is applied to increase power of the classical criteria for such data.

*Grouping.* The observed data is $\{(\mu_i^\circ, y_i),\ i=1,\dots,n\}$, where the conditional distribution of $y_i$ given the random pair $(\gamma_i^\circ,\gamma_i)=(\mu_i^\circ,\mu_i)$ is the Poisson distribution with the mean $\mu_i$, and $\{(\gamma_i^\circ,\gamma_i),\ i=1,\dots,n\}$ are i.i.d. with the common distribution $P^{(M)}$.

Let $\Delta=\Delta^{(M)}:=\{\Delta_k,\ k=1,\dots,K\}$ be a partition of $(0,\mu_+^\circ]$ into disjoint intervals $\Delta_k=(t_{k-1},t_k]$ of the length $|\Delta_k|:=t_k-t_{k-1}$ with $t_0=0$, $t_{K-1}<\mu_{+n}^\circ\le t_K<\infty$.

Without loss of generality one can assume that the sequence $(\mu_i^\circ,\ i=1,\dots,n)$ is nondecreasing. Define cumulative empirical sequences, the sequence for initial data,

$$\mu_{+j}^\circ=\sum_{i=1}^{j}\mu_i^\circ,$$

and the sequences determined by the partition $\Delta$,

$$\mu_{k+}=\sum_{i=1}^{n}\mu_i\mathbf{1}_{[\mu_{+i}^\circ\in\Delta_k]},\qquad y_{k+}=\sum_{i=1}^{n}y_i\mathbf{1}_{[\mu_{+i}^\circ\in\Delta_k]}.$$

Suppose that $Q^{(M)}$ and $P^{(M)}$ are the empirical distributions based on the data

$$\bigl\{(\mu_i^\circ,\mu_i^\circ),\ i=1,\dots,n\bigr\},\tag{12}$$

and

$$\bigl\{(\mu_i^\circ,\mu_i),\ i=1,\dots,n\bigr\},\tag{13}$$

respectively. The discrepancy between $Q^{(M)}$ and $P^{(M)}$ is measured by $\phi$-divergence for the grouped data:

$$d\bigl(Q^{(M)};P^{(M)}\bigr)=d_\phi\bigl(Q^{(M)};P^{(M)}\bigr):=\sum_{k=1}^{K}\mu_{k+}^\circ\phi\biggl(\frac{\mu_{k+}}{\mu_{k+}^\circ}\biggr).\tag{14}$$

The straighforward plug-in estimator of $d(Q^{(M)}; P^{(M)})$ is given by

$$\widehat{d(Q;P)} := \sum_{k=1}^{K} \mu_{k+}^{\circ} \phi\left(\frac{y_{k+}}{\mu_{k+}^{\circ}}\right). \tag{15}$$

Let $\eta_u \sim \mathrm{Poisson}(u)$ and suppose that

$$\mathbb{E}\phi^2\left(\frac{\eta_u}{v}\right) < \infty \quad \forall u, v > 0. \tag{16}$$

Denote

$$a(v) := v\mathbb{E}\phi\left(\frac{\eta_v}{v}\right), \tag{17}$$

$$\sigma^2(v;u) := v^2\mathbb{E}\left(\phi\left(\frac{\eta_u}{v}\right) - \phi\left(\frac{u}{v}\right)\right)^2. \tag{18}$$

**Lemma 2.** *Suppose* (16) *is fulfilled. Then*

$$\mathbb{E}_P\widehat{d(Q;P)} \geq d\big(Q^{(M)}; P^{(M)}\big), \tag{19}$$

$$\mathbb{E}_Q\widehat{d(Q;P)} = A(M) := \sum_{k=1}^{K} a(\mu_{k+}^{\circ}), \tag{20}$$

$$\mathrm{Var}_P\, \widehat{d(Q;P)} \leq V^2(M) := \sum_{k=1}^{K} \sigma^2(\mu_{k+}^{\circ}, \mu_{k+}). \tag{21}$$

*Proof* is presented in Appendix.

### 3.3 Consistency

From Lemma 2 it easy to derive the following result.

**Theorem.** *Let $Q^{(M)}$ and $P^{(M)}$ be the empirical distributions based on the data* (12) *and* (13)*, respectively. Suppose* (16) *is fulfilled and the discrepancy measure between $Q^{(M)}$ and $P^{(M)}$ is defined by* (14) *and estimated by* (15)*. If $\delta(M) > A(M)$ and*

$$V_0(M) + V(M) = o\big(\delta(M) - A(M)\big), \quad M \to \infty, \tag{22}$$

*where $A(M)$, $V(M)$ are introduced in* (20)*,* (21) *and*

$$V_0^2(M) := \sum_{k=1}^{K} \sigma^2(\mu_{k+}^{\circ}, \mu_{k+}^{\circ}), \tag{23}$$

*then the criterion*

$$\mathcal{K} := \big\{\widehat{d(Q;P)} > A(M) + \kappa_1\big(\delta(M) - A(M)\big)\big\},$$

*is consistent as $M \to \infty$ for testing* (6) *versus* (7) *with any constant $\kappa_1 \in (0,1)$.*

*Proof* is presented in Appendix.

**Remark 1.** If the partition $\Delta = \Delta^{(M)}$ with $K = K(M) \to \infty$ is such that

$$\Delta_{\min} = \Delta_{\min}^{(M)} := \min_k |\Delta_k| \to \infty, \quad M \to \infty,$$

then the statistic $\widehat{d(Q;P)}$ defined in (15) is asymptotically normal as $M \to \infty$. This fact can be established by arguments of Györfi and Vajda [19] used in the case of multinomial sampling scheme. In the case of sparse asymptotics, however, the power of the test based on the statistic $\widehat{d(Q;P)}$ heavily depends on grouping. Thus, even weaker requirement $\Delta_{\min}^{(M)} \geq \kappa_0$ with a pre-specified constant $\kappa_0 > 0$ may be rather restrictive.

In Section 4 we present (provide) some computer simulation results to illustrate performance of the proposed criterion.

## 4 Computer experiment

In this section the finite-sample ($n = 200$, $\mu_+ \approx 200$) behavior of goodness-of-fit tests based on two different methods of grouping ($K = 10$) is compared with classical criteria. The results of Monte Carlo study with $R = 1000$ replications for two extended Bayes models are presented.

In the first model, named "Bottom split", $\mu$ differs from $\mu^\circ$ in the region of low values of $\mu^\circ$ ("Bottom"), while in the second, named "Top split", $\mu$ differs from $\mu^\circ$ in the region of high values of $\mu^\circ$ ("Top"). The average values of $\mu$ in the both regions are kept close to that of $\mu^\circ$.

The Poisson distribution parameters, i.e. the expected frequencies $\mu$ and the true expected frequencies $\mu^\circ$, are generated as independent Gamma random variables:

$$\mu_i \sim \operatorname{Gamma}\bigl(a(i), v(i)\bigr), \quad \mu_i^\circ \sim \operatorname{Gamma}\bigl(a^\circ(i), v^\circ(i)\bigr), \quad i = 1, \ldots, 200.$$

Here $\operatorname{Gamma}(a, v)$ denotes the Gamma distribution with the mean $a$ and the variance $v$,

$$v^\circ(i) = v(i) = 10^{-4},$$
$$a^\circ(i) = 0.4 + 0.001i + 1.2 \cdot \mathbf{1}_{[i>n/2]},$$

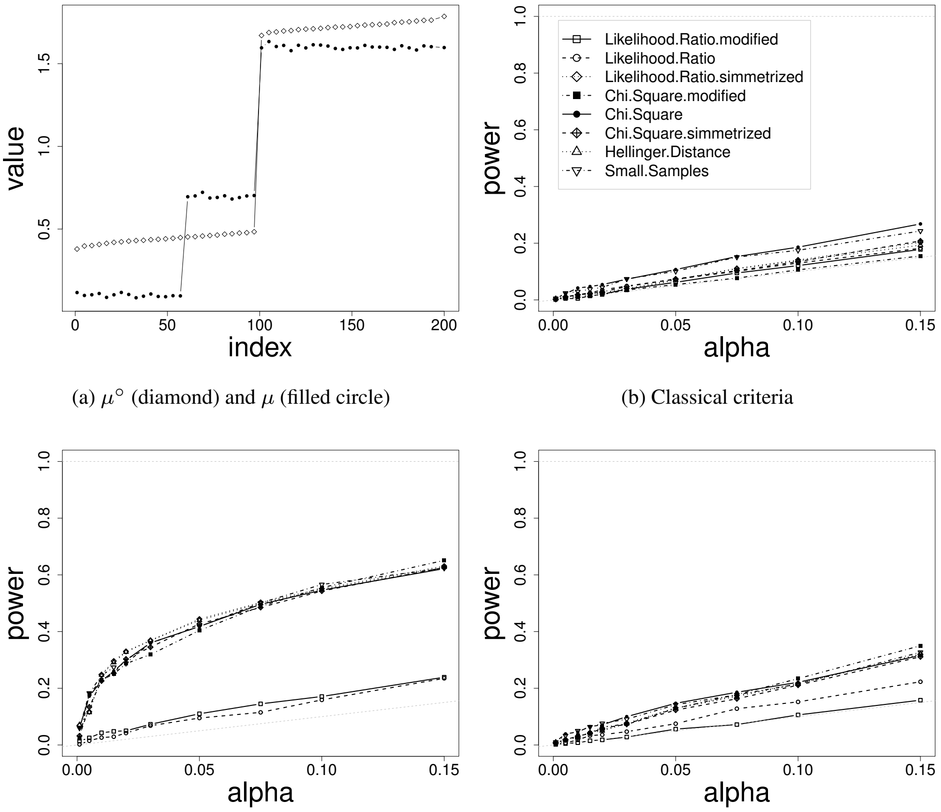
while

$$a(i) = a_{\text{bottom}}(i) := 1.6 - 0.9 \cdot \mathbf{1}_{[i \leq n/2]} - 0.6 \cdot \mathbf{1}_{[i \leq n/4]},$$

and

$$a(i) = a_{\text{top}}(i) := 0.4 + 0.6 \cdot \mathbf{1}_{[i>n/2]} + 1.2 \cdot \mathbf{1}_{[i>3n/4]},$$

in the "Bottom split" model and in the "Top split" model, respectively (see (a) in Fig. 1 and Fig. 2).



(a) $\mu^\circ$ (diamond) and $\mu$ (filled circle)

(b) Classical criteria

(c) Grouping of equal group sizes ($K = 10$)

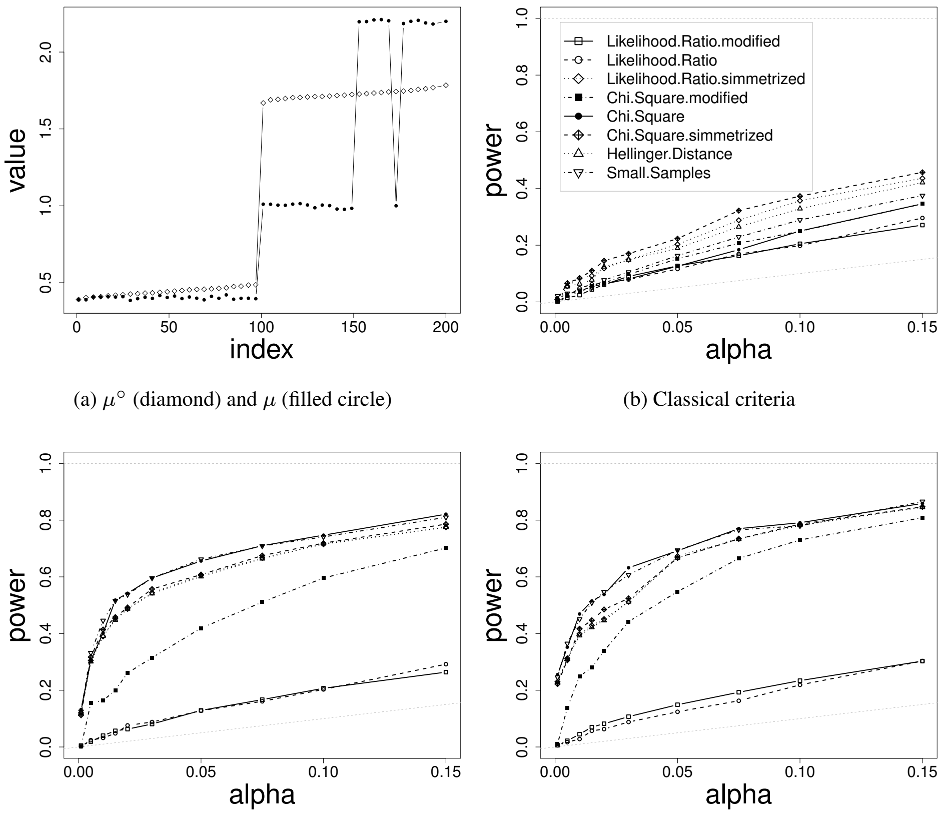
(d) Grouping of equal expected frequencies ($K = 10$)

Fig. 1. Goodness-of-fit tests power for "Bottom split" model. Legend in (b) applies to (c) and (d).

Two simple methods of grouping are applied. By the first method, the groups have equal sizes, i.e. number of elements. In the second method, the groups have equal expected frequencies $\mu^\circ_{k+}$, $k = 1, \ldots, 10$.

In the "Bottom" model, the first grouping method is much better than the second one (Fig. 1(c) and (d)), however, it is slightly worse in the "Top" model (Fig. 2(c) and (d)). Note that expected frequencies in the first grouping in the "Bottom" region are equal $8$ and thus normal approximation for these frequencies fails. Consequently, the performance of the test heavily depends on grouping and hence an adaptive grouping rule can significantly increase the power of the tests. Obviously, the grouping does not help if average values of $\mu^\circ$ and $\mu$ in each group are close.

The classical criteria based on the same $\phi$-divergencies (but without grouping) have very low power, see (b) in Fig. 1 and Fig. 2.



(a) $\mu^\circ$ (diamond) and $\mu$ (filled circle)

(b) Classical criteria

(c) Grouping of equal group sizes ($K = 10$)

(d) Grouping of equal expected frequencies ($K = 10$)

Fig. 2. Goodness-of-fit tests power for “Top split” model. Legend in (b) applies to (c) and (d).

## Appendix

*Proof of Lemma 2.* Since the function $\phi(u/v)$ is convex with respect to $u$ inequality (20) follows from Jensen’s inequality. Further, in view of (18)

$$v^2 \operatorname{Var} \phi\left(\frac{\eta_u}{v}\right) \le v^2 \mathbb{E}\left(\phi\left(\frac{\eta_u}{v}\right) - \phi\left(\frac{u}{v}\right)\right)^2 = \sigma^2(v,u).$$

Consequently,

$$\operatorname{Var}_P \widehat{d(P^\circ; P)} = \sum_{k=1}^{K} (\mu^\circ_{k+})^2 \operatorname{Var}_P \phi\left(\frac{y_{k+}}{\mu^\circ_{k+}}\right) \le \sum_{k=1}^{K} \sigma^2(\mu^\circ_{k+}, \mu_{k+}),$$

since $y_{k+}$, $k = 1, \dots, K$, are mutually independent Poisson random variables (given $\gamma = \mu$). □

*Proof of Theorem.* Let us check the first condition of Lemma 1 (8). Set $\tau(M) = (1-\kappa_1)(1 - A(M)/\delta(M))$. Then (22), (20), (21), (23), and Chebyshev inequality imply

$$\begin{aligned}
&\mathbb{P}_Q\bigl\{\widehat{d(Q;P)} > \bigl(1-\tau(M)\bigr)\delta(M)\bigr\}\\
&\quad\le \mathbb{P}_Q\bigl\{\widehat{d(Q;P)} - A(M) > \kappa_1\bigl(\delta(M) - A(M)\bigr)\bigr\}\\
&\quad\le \frac{V_0^2(M)}{\kappa_1^2(\delta(M) - A(M))^2} \to 0.
\end{aligned}$$

Similarly, for the second condition of Lemma 1 (9), we derive from (22), (19), (21), and Chebyshev inequality

$$\begin{aligned}
&\mathbb{P}_P\bigl\{\widehat{d(Q;P)} < d\bigl(Q^{(M)};P^{(M)}\bigr) - \tau(M)\delta(M)\bigr\}\\
&\quad\le \mathbb{P}_P\bigl\{\widehat{d(Q;P)} - \mathbb{E}_P\widehat{d(Q;P)} < -\tau(M)\delta(M)\bigr\}\\
&\quad\le \mathbb{P}_P\bigl\{\widehat{d(Q;P)} - \mathbb{E}_P\widehat{d(Q;P)} < -(1-\kappa_1)\bigl(\delta(M) - A(M)\bigr)\bigr\}\\
&\quad\le \frac{V^2(M)}{(1-\kappa_1)^2(\delta(M) - A(M))^2} \to 0.
\end{aligned}$$

The proof is complete. □

## References

1. E.V. Khmaladze, The statistical analysis of a large number of rare events, Technical Report, MS-R8804, Centrum Wiskunde & Informatica, Amsterdam, 1988.

2. A. Agresti, *Categorical Data Analysis*, Wiley & Sons, New York, 2007.

3. G. Kvizhinadze, *Large Number of Rare Events: Diversity Analysis in Multiple Choice Questionnaires and Related Topics*, PhD Dissertation, Victoria University of Wellington, 2010.

4. N. Cressie, T. Read, Multinomial goodness of fit tests, *J. R. Stat. Soc., Ser. B.*, **46**, pp. 440–464, 1984.

5. M.Y. Hu, *Model Checking for Incomplete High Dimensional Categorical Data*, University of California, Los Angeles, 1999.

6. U.U. Muller, G. Osius, Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, *Statistics*, **37**, pp. 119–143, 2003.

7. M. von Davier, Bootstraping goodness-of-fit statistics for sparse categorical data. Results of a Monte Carlo study, *Methods of Psychological Research Online*,**2**, pp. 29–48, 1997.

8. A. Agresti, B.D. Hitchcock, Bayes inference for categorical data analysis, *Stat. Methods Appl.*, **14**, pp. 297–330, 2005.

9. P. Congdon, *Bayesian Models for Categorical Data*, John Wiley and Sons, New York, 2005.

10. M. Radavičius, P. Samusenko, Profile statistics for sparse contingency tables under poisson sampling, *Aust. J. Stat.*, **40**, pp. 115–123, 2011.

11. P. Samusenko, Inconsistency of chi-square test for sparse categorical data under multinomial sampling, *Liet. mat. rink. LMD darbai*, **52**, pp. 327–331, 2011.

12. C.A.J. Klaassen, R.M. Mnatsakanov, Consistent estimation of the structural distribution function, *Scand. J. Stat.*, **27**, pp. 733–746, 2000.

13. Y.M. Bishop, S.E. Fienberg, P.W. Holland, *Discrete Multivariate Analysis. Theory and Practice*, The MIT Press, Cambridge, 1975.

14. S.E. Fienberg, P.W. Holland, Simulteneous estimation of multinomial cell probabilities, *J. Am. Stat. Assoc.*, **68**, pp. 683–691, 1973.

15. M. Aerts, I. Augustynas, P. Jansen, Central limit theorem for the total squared error of local polynomial estimators of cell probabilities, *J. Stat. Plann. Inference*, **91**, pp. 181–193, 2000.

16. M. Aerts, I. Augustynas, P. Jansen, Sparse consistency and smoothing for multinomial data, *Stat. Probab. Lett.*, **33**, pp. 41–48, 1997.

17. B. van Es, C.A.J. Klaassen, R.M. Mnatsakanov, Estimating the structural distribution function of cell probabilities, *Aust. J. Stat.*, **32**, pp. 85–98, 2003.

18. F. Liese, I. Vajda, On divergences and informations in statistics and information theory, *IEEE Trans. Inf. Theory*, **52**, pp. 4394–4412, 2006.

19. L. Györfi, I. Vajda, Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models, *Stat. Probab. Lett.*, **56**, pp. 57–67, 2002.