VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Pavel SAMUSENKO

NONPARAMETRIC CRITERIA FOR SPARSE CONTINGENCY TABLES

DOCTORAL DISSERTATION

PHYSICAL SCIENCES,
MATHEMATICS (01P)

Vilnius LEIDYKLA TECHNIKA 2012

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2008–2012.

Scientific supervisor
Assoc Prof Dr Marijus RADAVIČIUS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P).

http://leidykla.vgtu.lt
VGTU leidyklos TECHNIKA 2094-M mokslo literatūros knyga

ISBN 978-609-457-394-1

pavels.vgtu@gmail.com

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Pavel SAMUSENKO

NEPARAMETRINIAI KRITERIJAI RETŲ ĮVYKIŲ DAŽNIŲ LENTELĖMS

DAKTARO DISERTACIJA

FIZINIAI MOKSLAI,
MATEMATIKA (01P)

Vilnius LEIDYKLA TECHNIKA 2012

Disertacija rengta 2008–2012 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas
doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Abstract

In the dissertation, the problem of nonparametric testing for sparse contingency tables is addressed.

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. Traditionally, the expected (under null the 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* . Several techniques have been proposed to tackle the problem: exact tests, alternative approximations, parametric and nonparametric bootstrap, Bayes approach and other methods. However they all are not applicable or have some limitations in nonparametric statistical inference of very sparse contingency tables.

In the dissertation, it is shown that, for sparse categorical data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of null hypotheses to data. Thus, they can be inconsistent even in cases where a simple consistent test does exist.

An improvement of the classical criteria for sparse contingency tables is proposed. The improvement is achieved by grouping and smoothing of sparse categorical data by making use of a new sparse asymptotics model relying on (extended) empirical Bayes approach. Under general conditions, the consistency of the proposed criteria based on grouping is proved. Finite-sample behavior of the criteria is investigated via Monte Carlo simulations.

The dissertation consists of four parts including Introduction, 4 chapters, General conclusions, References and Appendices.

The introduction reveals the importance of the scientific problem, describes the purpose and tasks of the thesis, research methodology, scientific novelty, the practical significance of results The introduction ends in presenting the author's publications on the subject of the defended dissertation, offering the material of made presentations in conferences.

In Chapter 1, an overview of the problem is presented and basic definitions are introduced. Chapter 2 demonstrates the inconsistency of classical tests in case of (very) sparse categorical data for both multinomial and Poisson sampling scheme. In Chapter 3, extended Bayes model is introduced. It provides a basis for smoothing and grouping of sparse (nominal) data. The consistency of criteria based on grouping is proved. Finite-sample behavior of the classical and proposed criteria is studied in Chapter 4. Details of the computer simulation results are given in Appendix.

Santrauka

Disertacijoje sprendžiami neparametrinių hipotezių tikrinimo uždaviniai išretintoms dažnių lentelėms.

Problemos, susijusios su retų įvykių dažnių lentelėmis yra plačiai aptartos mokslinėje literatūroje. Yra pasiūlyta visa eilė metodų: tikslieji testai, alternatyvūs aproksimavimo būdai parametrinė ir neparametrinė saviranka, Bayeso ir kiti metodai. Tačiau jie nepritaikomi arba yra neefektyvūs neparametrinėje labai išretintų dažnių lentelių analizėje.

Disertacijoje parodyta, kad labai išretintiems kategoriniams duomenims tikėtinumo santykio statistika ir Pearsono χ^2 statistika gali pasidaryti neinformatyviomis: jos jau nėra tinkamos nulinės hipotezės ir duomenų suderinamumui matuoti. Vadinasi, jų pagrindu sudaryti kriterijai gali būti net nepagrįsti net tuo atveju, kai egzistuoja paprastas pagrįstas kriterijus.

Darbe yra pasiūlytas klasikinių kriterijų patobulinimas išretintų dažnių lentelėms. Siūlomi kriterijai remiasi išretintų kategorinių duomenų grupavimu ir glodinimu naudojant naują išretinimo asimtotikos modelį, kuris remiasi (išplėstine) empirine Bayeso metodologija. Prie bendrų sąlygų yra įrodytas siūlomų kriterijų, naudojančių grupavimą, pagrįstumas. Kriterijų elgesys baigtinių imčių atveju tiriamas taikant Monte Carlo modeliavimą.

Disertacija susideda iš įvado, 4 skyrių, literatūros sąrašo, bendrų išvadų ir priedo.

Įvade atskleidžiama nagrinėjamos mokslinės problemos svarba, aprašomi darbo tikslai ir uždaviniai, tyrimo metodai, mokslinis naujumas, praktinė gautų rezultatų reikšmė. Įvado pabaigoje pateikiamos autoriaus publikacijų disertacijos tema sąrašas, konferencijose darytų pranešimų medžiaga.

Pirmame skyriuje pateikiama nagrinėjamos temos apžvalga ir pagrindiniai apibrėžimai. Antrame skyriuje yra įrodytas klasikinių kriterijų nepagrįstumas labai išretintiems kategoriniams duomenims polinominėje ir Puasono ėmimo schemoje. Trečiame skyriuje įvedamas išplėstinis Bayeso modelis. Juo remiantis atliekamas išretintų (nominaliųjų) duomenų glodinimas ir grupavimas. Įrodomas kriterijų, sudarytų naudojant grupavimą, pagrįstumas. Klasikinių ir darbe pasiūlytų kriterijų elgesys baigtinių imčių atveju tiriamas ketvirtame skyriuje. Detalesni kompiuterinio modeliavimo rezultatai yra sudėti į Priedą.

Notation

Symbols

N – the number of observation;
n – the number of outcomes;
y_j – the frequency of outcome j;
μ_j – the expected frequency of outcome j;
μ_j° – the possible frequency of outcome j;
p_j – the probability of outcome j;
$\boldsymbol{y}$ – vector of observed frequencies;
$\boldsymbol{\mu}$ – vector of expected frequencies;
$\boldsymbol{\mu}^\circ$ – vector of possible frequencies;
$\boldsymbol{p}$ – vector of probabilities;
X^2 – Pearson χ^2 statistic;
G^2 – likelihood ratio statistic (LR);
$\mathbb{D}(\boldsymbol{y})$ – the variance of y;
$\mathbb{E}(\boldsymbol{y})$ – the expectation of y;
F – the density function;
P – the probability function;
$\mathbb{N}$ – the set of natural numbers;

$\mathbb{Z}$ – the set of integer numbers;
$\mathbb{R}$ – the set of real numbers;
$\|\boldsymbol{\mu}\|_q$ – denotes q norm of μ;
$\mathbb{1}\{A\}$ – the indicator function of the set A;
H_0 – the null hypothesis;
H_1 – the alternative.

Abbreviations

$LNRE$ – Large Number of Rare Events;
$MCMC$ – Markov chain Monte Carlo.

Contents

Introduction

Scientific problem

In the subfield of numerical analysis, a sparse table is a table populated primarily with zeros. The concept of sparsity is useful in combinatorics and application areas such as network theory, which have a low density of significant data or connections. Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. Traditionally, the expected (under null the 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: exact tests (see, overview by Agresti, 2007), alternative approximations (Hu, 1999; Müller and Osius, 2003), parametric and nonparametric bootstrap (von Davier, 1997), Bayes approach (Agresti, Hitchcock, 2005; Congdon, 2005) and other methods, but they all are not applicable or have some limitations in nonparametric statistical inference of very sparse contingency tables.

The dissertation is devoted to the problem of consistent nonparametric testing for sparse contingency tables and the investigation of sparsity effect on the test power.

Topicality of the work

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. 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).

Examples of real sparse categorical data along with their statistical analysis and discussion can be found in (Khmaladze, 1988, p. 3; Kvizhinadze, 2010, p. 3; Agresti, 2007, p. 149; StatXact, 2011, p. 33; see also section 1.2, p. 10)

In general, consistent estimator for expected cell counts of sparse contingency table does not exist. For categorical data of ordered variables, kernel smoothing or grouping enables one to obtain consistent estimators of expected cell counts under additional assumptions of their smoothness. For nominal data, the main object of our study, both kernel smoothing and grouping are not directly applicable.

Research object

The research objects are definitions of sparsity, sparse contingency tables of nominal data, goodness-of-fit statistics based on power divergences, consistency and finite-sample properties of nonparametric tests.

The aim and tasks of the work

The aim of this work is to investigate asymptotic and finite-sample behavior of classical goodness-of-fit statistics and tests for sparse continency tables and to propose improvements of the classical tests.

The tasks of this work are:

1. To propose improvements of the classical tests in order to increase their

sensitivity to deviations from null hypotheses for sparse nominal data.

2. To prove the consistency of the proposed tests.
3. To compare the finite-sample performance of the classical and proposed tests under alternatives of various types by means of computer simulations.

Applied methods

Sparse asymptotics modeling is based on latent distribution model, structural distribution model for large number or rare events (Khmaladze (1988)), and the empirical Bayes approach. Goodness-of-fit is measured by power divergences (Cressie and Reed, 1988). Results for likelihood ratio criterion with profile (spectral) statistics substituted for cell counts rely on general likelihood theory and model-based cluster analysis. Chebyshev type inequalities are used to establish the consistency (inconsistency) of the tests. Finite-sample performance of the tests is studied by Monte-Carlo simulations. All calculations are performed using the `R` software.

Scientific novelty

It is shown that, for (very) sparse data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of null hypotheses to data. For instance, they can be inconsistent even in cases where a simple consistent test does exist.

A new sparse asymptotics model relying on (extended) empirical Bayes approach is introduced. Tests based on MCMC smoothing, on smoothing by grouping or on (finite) mixtures of Poisson distributions are proposed. The consistency of the tests based on grouping is proved.

Practical value of the work results

The proposed nonparametric tests can be applied for statistical inference of high-dimensional categorical (nominal) data frequently met in surveys with large questionnaires, natural language and text processing, genetic data, etc. The tests are easy to implement, they are computationally efficient and do not

impose any specific requirements on the categorical (nominal) data they are applied to.

Statements presented for defence

1. For (very) sparse data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of null hypotheses to data.
2. Sparse asymptotics based on (extended) empirical Bayes approach enables one to apply distribution model to sparse nominal data.
3. In the empirical Bayes setting, MCMC smoothing, smoothing by grouping and modeling by finite mixtures of Poisson distributions can improve the power of classical tests especially for regular alternatives.
4. Under general conditions, the tests based on grouping are consistent.
5. The effect of grouping (smoothing) significantly depends on the grouping (smoothing) method as well as on its parameters (number of groups, number of iterations, etc.).
6. For the irregular alternatives that differ from the null hypothesis by centered independent Gamma random variables ("noise"), the grouping tests which use the discrepancies between both the means and the variances have much better power.

Approval of the work results

On the topic of dissertation there were 4 papers published in reviewed scientific journals. The research results were reported at 5 scientific conferences. The list of conference talks is as follows :

1. Testing problems for sparse contingency tables, *International conference on applied mathematics and approximation theory*, Ankara, Turkey, 2012.
2. Testing problems for sparse contingency tables, *10th international Vilnius conference on probability theory and mathematical statistics*, Vilnius, 2010.
3. Profile statistics for sparse contingency tables, *9th international com-*

puter data and modeling conference on complex stochastic data and systems, Minsk, Belarus, 2010.

4. Goodness-of-fit tests for smoothed categorical data, *LMD 53th conference*, Klaipėda, 2012.
5. Inconsistency of χ^2 test for sparse categorical data under multinomial sampling, *LMD 52th conference*, Vilnius, 2011.

The scope of the scientific work

The dissertation consists of the introduction, three chapters, the conclusions, references, appendix and the list of author's publications. The total scope of the dissertation is 106 pages, 70 mathematical expressions, 29 tables, 31 figures and 62 items of reference.

Acknowledgment

I really appreciate working with doc. dr. Marijus Radavičius for his supervision, patience, helpful and revealing suggestions and constant guidance. He involved me in this interesting project and gave me the chance to learn and experience really much, trusting my capabilities. He always encouraged to do my very best.

Thank you for colleague from Vilnius Gediminas Technical University for their help to provide me encouragement and stimulus to go ahead.

I very much appreciate my parents, my family and my friends. You helped me in this crucial and intensive period. Many thanks for listening to my complaints and frustrations, for supporting me and especially for tolerating my mood behaviors.

1

Historical overview and definitions

The statistical analysis of contingency table is a well studied area and has drawn lots of attention in the statistical literature over the past decades. The developments of appropriate models and test statistics are presented in monographs by Haberman (1974), Bishop et al. (1975), Fienberg (1980), Read and Cressie (1988), Agresti (2002), Congdon (2005), Agresti (2007) and reviewed by Agresti (1992), Fienberg (2000), Agresti and Hitchcock (2005), among others.

Currently the amount of information is very extensive, therefore problems related to a large dimension and/or sparsity 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 $\mathcal{O}(k^2)$ 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 impor-

tant 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. Khmaladze, 1988, 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 χ^2 approximations of goodness-of-fit statistics may be inaccurate and the table is said to be sparse (Cohran, 1954; Agresti, 2007).

Examples of real sparse categorical data along with their statistical analysis and discussion can be found in (Khmaladze, 1988, p. 3; Kvizhinadze, 2010, p. 3; Agresti, 2007, p. 149; StatXact, 2011, p. 33; see also section 1.2, p. 10).

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

1. The standard χ^2 approximation for distributions of classical tests is not sufficiently accurate (see, Agresti, 2007; Cressie, Read, 1984). Several techniques have been proposed to tackle this problem: exact tests (see, overview by Agresti, 2007; Filina, Zubkov, 2008; StatXact, 2011, and references therein), alternative approximations (see, Hu, 1999; Müller, Osius, 2003; Filina, Zubkov, 2011), parametric and nonparametric bootstrap (von Davier, 1997), Bayes approach (Agresti, Hitchcock, 2005; Congdon, 2005) and other methods.

2. The classical tests are not longer (asymptotically) distribution free (see, Khmaladze, 1988). The test is said to be distribution free if its (asymptotic) distribution is independent of the null hypothesis to be tested and thus it leads to universal decision rules. The lack of this property means that calculation of critical value of every testing problem is a specific problem to be solved.

3. For (very) sparse data, the classical criteria become noninformative: their test statistics 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 (Radavičius, Samusenko, 2011; Samusenko, 2011; see also Khmaladze, 1988; Klassen, Mnatsakanov, 2000).

This work is devoted mainly to the third problem. It reveals that possibly there is no sense to solve the former two problems.

In the next section, we present a brief overview of statistical problems for sparse categorical data and different approaches to deal with sparsity. The proposed extended empirical Bayes model of sparse asymptotics contains the latent distribution and the structural distribution models as special cases.

1.1. Categorical data and contingency tables

Categorical variables have two main types of scales, nominal and ordinal. In the dissertation main attention is given to the nominal data. There is only one logic operation (comparison) exists between nominal variables equal / not-equal which does not change information of the data. To determine whether there is a relation between the variables they are structured, summarized and displayed in cross tables (contingency table)

Table 1.1. Contingency table of two variables A and B

		B				
	y_{ij}	1	2	…	j	
	1	y_{11}	y_{12}	…	y_{1j}	y_{1+}
	2	y_{21}	y_{22}	…	y_{2j}	y_{2+}
	·	·	·	…	·	·
A	·	·	·	…	·	·
	·	·	·	…	·	·
	i	y_{i1}	y_{i2}	…	y_{ij}	y_{i+}
		y_{+1}	y_{+2}	…	Y_{+j}	y_{++}

Goodness-of-fit of a model and a contingency table with I number of rows, J columns, and $n = I \times J$ cells is typically measured by Pearson's χ^2 statistic X^2 (Pearson 1900)

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}, \tag{1.1}$$

or the (doubled negative logarithmic) likelihood ratio statistic

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} observed_{ij} \ \log\left(\frac{observed_{ij}}{expected_{ij}}\right). \tag{1.2}$$

Karl Pearson (1900) demonstrated that the large-sample distribution of a test statistic, based on the standardized squared differences between the observed and expected counts of categorical data generated from multinomial, hypergeometric or Poisson distributions, is χ^2 distribution. This work was followed by significant contributions by, among others, Yule (1912), R. A. Fisher (1924), (1934), Yates (1934), Cochran (1936), (1954), Kendall and Stuart (1958), Goodman (1968), see also textbooks (Agresti, 2002) and overview (Fienberg, 2000) and references therein.

As a rule the selection of critical values of goodness-of-fit statistics is based on the convergence of it's distribution to the χ^2 distribution with appropriate degrees of freedom as a samples size tends to infinity. In practice samples sizes are bounded and the question of accuracy of such approximation arises naturally.

1.2. Sparse contingency tables

Goodness-of-fit of a model and a contingency table is typically measured by Pearson's χ^2 statistic X^2 or the (doubled negative) likelihood ratio statistic G^2. When the (expected) frequencies in cells of the contingency table go to ∞ the both statistics have asymptotic χ^2 distribution. The decisions in hypothesis testing rely on this χ^2 approximation, see, e.g., (Neyman and Pearson, 1928; Bishop, et al. 1975; McCullagh and Nelder, 1983; Agresti, 1990). However, the χ^2 approximation usually fails either the table is sparse or the sample size is not large enough. Unfortunately, sparseness of the table is rather often encountered in practice. As Fienberg put it, "The fact remains ... that with extensive questionnaires of modern-day sample surveys, and the detailed and painstaking inventory of variables measured by biological and social scientists, the statistician is often faced with large sparse arrays, full 0's and 1's, in need of careful analysis." (Fienberg, 1980, pp. 174-175). As the dimensionality of categorical data increases, the sparseness is common even when the total sample size is large. The χ^2 approximation can also break down when the table is small but contains very large as well as small cell counts (so-called unbalanced contingency table). Even when the sample size is quite large, recent work has shown that large-sample approximations can be very poor when the contingency table is unbalanced (Haberman, 1988).

Real example of sparse contingency table presented in Table 1.2 It's briefly described and analyzed in "StatXact 5 User Manual" (StatXact, 2011).

Here are cross-tabulated an index of competitiveness against the student to faculty ratio.

Table 1.2. Student/faculty ratio versus competitiveness of state universities

Student/Faculty	Competitiveness of Institution					Row
Ratio	Less	Average	Very	Highly	Most	Total
2	0	0	0	1	0	1
7	0	1	0	1	0	2
8	0	1	0	0	1	2
9	0	1	0	0	0	1
10	1	0	2	0	0	3
11	1	3	0	1	0	5
12	0	2	1	0	0	3
13	1	3	1	0	0	5
14	3	3	1	0	0	7
15	1	5	1	1	0	8
16	1	5	0	0	0	6
17	3	2	1	0	0	6
18	0	2	4	1	0	7
20	0	2	0	0	0	2
21	2	0	0	0	0	2
22	0	0	1	0	0	1
23	0	1	0	0	0	1
24	0	1	1	0	0	2
70	0	1	0	0	0	1
Column Total	13	33	13	5	1	65

Unbalanced categorical data example. The data consists of measurements of heart wall thickness of 947 athletes participating in 25 different sports. The wall thickness $\geq$ 13 mm is indicative of hypertrophic cardiomyopathy. The average proportion of occurrence of the event "wall thickness $\geq$ 13 mm" is about 0,017 (16 occurrences) with the number of athletes in each kind of sport activity ranging from 6 to 95 (38 in average) and the frequency of the event ranging from 0 to 4 with 0 observed 14 times. The problem is to test if probability of the event is related to the kind of sport activity. Example is discussed in Senchaudhuri, Mehta, Patel (1995) and presented in Table 1.3.

Actually, it is not clear what does it practically mean "large expected frequencies" and "large contingency tables". The same is true for "sparse contingency tables". One rule of thumb, due to Cochran (1954), is: The minimum expected cell count for all cells should be at least 5. The problem with this rule is that it can be too conservative. Another rule of thumb, also due to Cochran

Table 1.3. Left ventricular wall thickness versus sporting activity

Sports	Thickness (mm)		Total	Sports	Thickness (mm)		Total
	≥ 13	< 13			≥ 13	< 13	
Weightlifting	1	6	7	Diving	1	10	11
Field wt. events	0	9	9	Boxing	0	14	14
Wrestling/Judo	0	16	16	Cycling	1	63	64
Tae kwon do	1	16	17	Water Polo	0	21	21
Roller Hockey	1	22	23	Yachting	0	24	24
Team Handball	1	25	26	Canoeing	3	57	60
Cross-coun.Ski	1	30	31	Fencing	1	41	42
Alpine Skiing	0	32	32	Tennis	0	47	47
Pentathlon	0	50	50	Rowing	4	91	95
Roller Skating	0	58	58	Swimming	0	54	54
Equestrian	0	28	28	Soccer	0	62	62
Bobsledding	1	15	16	Track	0	89	89
Volleyball	0	51	51				

(1954), is: For tables larger than 2×2, a minimum expected count of 1 is permissible as long as no more than about 20% of the cells have expected values below 5. Various other rules of thumb are recited in (Hu, 1999, p.2).

Simulation studies by Koehler and Larntz (1980) have shown that these rules are far too conservative but it is hopeless to expect simple guidelines to indicate when asymptotic large-sample approximations are adequate (see also Agresti, 2002, for further discussion).

A variety of practical remedies are suggested to solve the sparsity problem (see, e.g., Reiser and VandenBerg (1994), Jöreskog and Moustaki (2001), Kraus, 2012, pp.11,12, and references therein). They include

- collapsing or dropping the contingency table cells with small cell counts;
- adding constants to cells;
- considering only cells with observed or expected frequencies that exceed a certain values.

However, this leads to loss of information and unforeseen results (see, for example, Baglivo et al., 1988).

Other alternatives are to use the exact tests (Agresti, 1992; StatXact, 2011) or (parametric) bootstrap (Davier, 1997). Although this poses no new theoretical challenges it may be computationally expensive or infeasible to do so, especially when the table is large.

1.3. Classical test statistics

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

$$\boldsymbol{y} \sim Poisson_n(\boldsymbol{\mu}). \tag{1.3}$$

A possible alternative might be

$$\boldsymbol{y} \sim Multinomial_n(N, \boldsymbol{p}), \quad \boldsymbol{\mu} := \boldsymbol{p}/N, \tag{1.4}$$

where $\boldsymbol{p}$ is a n-dimensional vector probabilities summing to one.

Consider a simple testing problem

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}^\circ \text{ versus } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}^\circ. \tag{1.5}$$

Here $\boldsymbol{\mu}^\circ := (\mu_1^\circ, \ldots, \mu_n^\circ)$ is a given vector of positive values.

A variety of statistics have been derived for goodness-of-fit testing such as the Pearson chi-squared statistic (X^2), the log likelihood ratio statistic (G^2), the Freeman-Tukey statistic (F^2), the Neyman modified X^2 (X_m^2), and the log likelihood ratio modified G^2 (G_m^2). All of the above statistics are embedded in a family of power divergence statistics thoroughly discussed by Cressie and Read (1984, 1988). For multinomial sampling scheme they are given by

$$D_\lambda = D_\lambda(\boldsymbol{y};\, \boldsymbol{\mu}^\circ) := \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{n} y_i \left[\left(\frac{y_i}{\mu_i^\circ} \right)^\lambda - 1 \right], \tag{1.6}$$

where λ is a real-valued parameter. The cases $\lambda = 0$ and $\lambda = -1$ are defined as the limits $\lambda \to 0$ and $\lambda \to -1$.

When the observed frequencies y and the expected under the null hypothesis H_0 frequencies $\boldsymbol{\mu}^\circ$ match exactly for each possible outcome, the power-divergence statistic is zero (for any choice of λ). In all other cases the statistic is positive and becomes larger as the observed and expected frequencies diverge.

In almost all cases a reasonable choice of λ will lie in the range $\lambda \in (-1,\, 2]$. This conclusion is based on the results on the calculation of significance levels for small samples Cressie and Read (1984, 1988). The two most popular special cases of the power-divergence statistic are Pearson's χ^2 statistic (1.1) (put $\lambda = 1$) and the likelihood ratio statistic (1.2) (the limit as $\lambda \to 0$). The chi-squared statistic X^2 is efficient for testing the equiprobable hypothesis

against certain local alternatives in large sparse tables (Cressie and Read, 1984; Ermakov, 1998). The loglikelihood ratio statistic G^2 is more suitable for testing against certain nonlocal alternatives with some near-zero probabilities. Cressie and Read (1984, 1988) argued that the power-divergence statistic with $\lambda = 2/3$ (we called small sample statistic the best choice for small n when there is little or no knowledge of possible alternative models

$$D_{2/3} = \frac{9}{5} \sum_{i=1}^{n} y_i \left[\left(\frac{y_i}{\mu_i^\circ} \right)^{\frac{2}{3}} - 1 \right].$$

This statistic lies between X^2 and G^2 in terms of the parameter λ.

Various other goodness-of-fit statistics have been proposed. These include the Freeman-Tukey statistic (Freeman and Tukey, 1950; Bishop et al., 1975), which, following Fienberg (1979) and Moore (1986), we define as a power divergence statistics with $\lambda = -1/2$,

$$F^2 = 4 \sum_{i=1}^{n} \left(\sqrt{y_i} - \sqrt{\mu_i^\circ} \right)^2,$$

also the symmetrized (logarithmic) likelihood ratio statistic

$$G_s^2 = \sum_{i=1}^{n} (y_i - \mu_i^\circ) \ \log \left(\frac{y_i}{\mu_i^\circ} \right)$$

and the Le Cam or symmetrized X^2 statistic

$$X_s^2 = 2 \sum_{i=1}^{n} \frac{(y_i - \mu_i^\circ)^2}{y_i + \mu_i^\circ}.$$

Zelterman (1987) proposed a statistic D^2 and compared it with X^2 for testing goodness-of-fit to sparse multinomial distributions, where

$$D^2 = X^2 - \sum\nolimits^{*} (y_i / \hat{\mu}_i),$$

where $\hat{\mu}_i$ is the estimated expected count of the i-th cell, and $\sum^*$ is the summation over all the cells such that $\hat{\mu}_i > 0$. The D_2 statistic is not a member of the family of power divergence statistics.

The simulation experiment strongly recommends to use the D^2 test, in

comparison with the X^2 test, for goodness-of-fit testing with a large sparse contingency table due to its high sensitivity to the sample size and model discrepancy (Kim et al., 2007).

1.4. General discrepancy measures

Here we introduce more general class of goodness-of-fit statistics which includes the family of power divergences and, moreover, is suitable for Poisson sampling scheme.

The Csisźzar ϕ-divergence (Csiszar, (1967); Liese and Vajda, 1987, 2006; Csiszar and Shields, 2004) between two vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{R}_+^n$ is defined by

$$d_\phi(\boldsymbol{v};\, \boldsymbol{u}) := \sum_{i=1}^{n} v_i \phi(u_i/v_i).$$

The function $\phi : \mathbb{R}_+ \to \mathbb{R}$ is convex, strictly convex at 1 with $\phi(1) = 0$. The most of ϕ-divergences widely used to measure distribution discrepancy belong to power-divergence family (cf. Cressie and Read, 1984) with $\phi = \phi_\alpha$ and $\lambda = \alpha - 1$, cf. (1.6):

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

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

For $\phi = \phi_\alpha$, denote $d_\alpha(\boldsymbol{v};\, \boldsymbol{u}) := d_\phi(\boldsymbol{v};\, \boldsymbol{u})$. Taking $\alpha = 1$ and $\alpha = 2$ produce the classical (logarithmic) likelihood ratio statistic G^2 and Pearson χ^2 statistic X^2, respectively (see also Table 1.4).

The selection of critical values of goodness-of-fit statistics are based on their asymptotic distributions as a samples size N tends to infinity. For valid H_0 and fixed n, the number of cells in a contingency table, the goodness-of-fit statistics based on ϕ-divergences with the sufficiently smooth function ϕ, in particular power-divergence statistics, are asymptotically equivalent and χ^2 distributed (with appropriate degrees of freedom). This equivalence is preserved even in case where the number n of cells tends to ∞ provided $n^2 = o(N)$ and behavior of expected cell frequencies in the table are sufficiently regular (see, e.g., Györfi and Vajda, 2002). For contingency tables with higher sparsity this asymptotic equivalence is usually broken.

Table 1.4. Goodness-of-fit statistics

Statistic	λ	α	Definition
$D_{2/3}$	2/3	5/3	$\frac{9}{5}\sum_{i=1}^{k} y_i \left(\left(\frac{y_i}{\mu_i^\circ} \right)^{\frac{2}{3}} - 1 \right)$
F^2	-1/2	1/2	$4\sum_{i=1}^{k} \left(\sqrt{y_i} - \sqrt{\mu_i^\circ} \right)^2$
G_m^2	-1	0	$2\sum_{i=1}^{k} \left(\mu_i^\circ \log \frac{\mu_i^\circ}{y_i} + (y_i - \mu_i^\circ) \right)$
X_m^2	-3	-2	$\sum_{i=1}^{k} \frac{(\mu_i^\circ - y_i)^2}{y_i}$
G_s^2			$\sum_{i=1}^{k} (y_i - \mu_i^\circ) \log \frac{y_i}{\mu_i^\circ}$
X_s^2			$2\sum_{i=1}^{k} \frac{(y_i - \mu_i^\circ)^2}{y_i + \mu_i^\circ}$

In case of very sparse contingency tables, the classical statistics, at least X^2 and G^2, are not appropriate for testing goodness-of-fit as shown by Radavičius and Samusenko (2011) and Samusenko (2011).

1.5. Sparse asymptotics

Bishop et al. (1975) proposed another asymptotic framework, more suitable for modelling sparse contingency tables. This framework, called sparse asymptotics, assumes that the number of cells n in the contingency table goes to infinity along with the sample size N. In fact, testing goodness-of-fit for large contingency tables with the number of cells $n \to \infty$ have been considered earlier, see, for instance, Tumanyan (1954, 1956) and Steck 1957. Results of Haberman (1977) suggest that the usual advices (such as the rule of thumb: expected cell counts should be not less than 5) are far too conservative, at least for the parametric inference. Moris (1975) (see also Tumanyan (1954, 1956)) have shown that, for the Pearson statistic X^2 and the likelihood ratio statistic G^2, χ^2 approximation holds also when $n = n(N) \to \infty$ provided the minimum of the expected cell counts goes to ∞ sufficiently fast (and hence $n = o(N)$). Here χ^2 approximation means that distribution is asymptotically normal with mean n and variance $2n$. Györfi and Vajda (2002) extended this result to goodness-of-fit statistics from the family of Csiszar's ψ-divergencies. Under some regularly conditions, the asymptotic normality of classical statistics is retained for n of the same order as N. The centering and scaling now, however, depends on unknown expected cell frequencies (see, e.g., Medvedev (1977) and survey by Ivanov, Ivchenko and Medvedev, 1984). Consequently, the classical criteria are not distribution free in this case (see also Khmaladze, 1988).

A considerable literature available on limit theorems in allocation or occupancy problems (Kolchin et al., 1978; Ivanov et al., 1984; Gnedin et al., 2007; and references therein).

The asymptotic behavior of criteria based on the number of empty boxes and other occupancy statistics (sometimes called spectral statistics, see Khmaladze, 1988, Kvizhinadze, 2009) is investigated in the book by Kolchin, Sevast'yanov and Chistyakov (1978), see also Ivchenko and Medvedev (1980).

This short outline of asymptotic behavior of test statistics for contingency tables with increasing number cells suggests that the actual effect of sparsity shows up when $N = \mathcal{O}(n),\ n \to \infty$ and/or the expected cell frequencies (under the null hypothesis H_0) show certain irregularity (unbalanced contingency tables). In order to obtain reasonable results in case of sparse asymptotics one needs to make some assumptions about asymptotic behavior of expected cell counts in sparse contingency tables. In the next subsection approaches to sparsity modelling are briefly discussed.

1.5.1. Large number of rare events

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.

The definition of sparsity is based on the sparse asymptotics (cf. Fienberg and Holland,1972; Bishop et al., 1975; Khmaladze 1988). Let $M \to \infty$ be some asymptotic parameter. The sparse asymptotics assumes that $n = n(M) \to \infty$ and the expected cell counts $\boldsymbol{\mu} = \boldsymbol{\mu}(M)$ with $\mu_+ = \mu_+(M) \to \infty$ 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.

Let

$$\boldsymbol{y} \sim Poisson_n(\boldsymbol{\mu})$$

be a contingency table of size n obtained by the Poisson sampling. The multinomial sampling scheme,

$$\boldsymbol{y} \sim Multinomial_n(N, \boldsymbol{p}), \quad \boldsymbol{\mu} := \boldsymbol{p}N,$$

is obtained from the Poisson case by conditioning on the observed total $y_+ := \sum_{i=1}^n y_i$ and assuming $y_+ = N$. Then also $\mu_+ := \mathbb{E}y_+ = N$.

On the other hand, multinomial sampling can be reduced via poissoniza-

tion to the corresponding Poisson sampling scheme which provides a good approximation to the former (in case of sparse asymptotics, see, e.g., Mnatsakanov and Klassen, 2000; van Es et al., 2003). Alternative approximation results are presented in (Čekanavičius, 1999; Čekanavičius and Wang, 2003; Zaitsev, 2005 and references therein).

Khmaladze (1988) introduced a general framework called large number of rare events (LNRE). LNRE can be viewed as a definition of sparse categorical data.

For $k \in \mathbb{Z}_+$, set

$$\begin{aligned} \nu_n(k) &:= \sum_{i=1}^{n} \mathbb{1}\{y_i = k\}, \\ \nu_n &:= \sum_{i=1}^{n} \mathbb{1}\{y_i > 0\} = \sum_{k=1}^{\infty} \nu_n(k). \end{aligned}$$

A vector of observed cell counts y is called contingency table (categorical data) with LNRE(I) and LNRE(II), respectively, if

$$\lim \inf_{M\to\infty} \frac{\mathbb{E}\nu_n(1)}{\mathbb{E}y_+} > 0, \tag{1.7}$$

respectively, if

$$\lim \inf_{M\to\infty} \frac{\mathbb{E}\nu_n(1)}{\mathbb{E}\nu_n} > 0 \tag{1.8}$$

(cf. Khmaladze, 1988, p. 6). Since $y_+ \geq \nu_n$, LNRE(I) implies LNRE(II).

Suppose that, for some constant M_0, $\mu_i \leq M_0 \; \forall i = 1, \ldots, n$. Then

$$\|\mu\|_\infty \leq M_0. \tag{1.9}$$

Here and in the sequel $\|\cdot\|_p$, $p \geq 1$ or $p = \infty$, denotes the l_p-norm. Under restriction (1.9), the LNRE conditions (1.7) and (1.8) are equivqlent for the both sampling schemes, Poisson and multinomial.

Radavičius and Samusenko (2011) considered (very) sparse categorical data (contingency tables). It means that (as $M \to \infty$)

$$\|\boldsymbol{\mu}\|_2^2 = o\left(\|\boldsymbol{\mu}\|_1\right). \tag{1.10}$$

Remark 1.1. Condition (1.10) together with (1.9) implies

$$\|\boldsymbol{\mu}\|_1 \le M_0 n, \tag{1.11}$$

and for arbitrary $h > 0$,

$$\begin{aligned} h^2|\{j : \mu_j \ge h\}| \le \|\boldsymbol{\mu}\|_2^2 &= o\left(\|\boldsymbol{\mu}\|_1\right), &(1.12)\\ h\|\boldsymbol{\mu}\|_1 \le h^2|\{j : \mu_j \le h\}| + \|\boldsymbol{\mu}\|_2^2 &\le h^2 n + o\left(\|\boldsymbol{\mu}\|_1\right). &(1.13)\end{aligned}$$

Here and later $|A|$ stands for the number of elements of the set A.

From (1.11)–(1.13), it follows that

$$\begin{aligned} n_h(\boldsymbol{\mu}) := |\{j : \mu_j \ge h\}| &= o\left(n\right), \quad \forall h > 0,\\ \|\boldsymbol{\mu}\|_q^q &= o\left(n\right), \quad q = 1,\, 2.\end{aligned}$$

Consequently, the expected number of the nonzero cells $\mathbb{E}n_h(\boldsymbol{y})$, $h \in (0, 1)$, as well as the expected value of the total frequency $\mu_+ = \mathbb{E}\|\boldsymbol{y}\|_1$ is much smaller than n. Thus, the contingency table $\boldsymbol{y}$ contains a lot of zeros. LNRE conditions actually means that the contingency table contains a significant part of events observed once in available categorical data (i.e., this part is proportional to the number of observations or to the number of observed different events). Conditions (1.10) and (1.9) insure that LNRE(I) and LNRE(II) are equivalent.

Below we briefly discuss two sparsity models: latent distribution and structural distribution.

1.5.2. Latent distribution model

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

$$\mu_i = \mu_+ \left(F(t_i) - F(t_{i-1})\right), \tag{1.14}$$

where $t_0 = 0,\ t_i := i/n,\ i = 1, \ldots, n$ (cf. Bishop et al., 1975; Aerts et al., 2000). In this setting, it is usually assumed that there exists rather smooth latent distribution density f, $f(u) = dF(u)/du$. This assumption implies

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

Thus, in this case the sparsity is expressed by the average expected frequency $\rho = \rho(M) := \mu_+/n$. For multinomial sampling scheme (1.4) we have $\mu_+ = N$ where N is the sample size of the contingency table $\boldsymbol{y}$. Hence $\rho(M) = N/n$. A typical assumption for the sparse asymptotics is $\rho = \mathcal{O}(1)$. 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., Aerts et al., 1997). 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., Bishop et al., 1975; Aerts et al., 2000).

The latent distribution model (1.14) 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 $\boldsymbol{\mu}$ and their sparsity can be described by the structural distribution function introduced by Khmaladze (1988) (see also Klaassen, Mnatsakanov, 2000; Es, Klaassen, Mnatsakanov, 2003) to characterize data with a large number of rare events (LNRE for short). Thus, LNRE is Khmaladze's definition of sparse categorical data.

1.5.3. Structural distribution model

When dealing with testing problem (1.5), one can suppose that the cell numbering order is irrelevant. It means that the statement $\boldsymbol{\mu} = \boldsymbol{\mu}^\circ$ 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

$$H_0 : \hat{F}^{(M)} = (F^\circ)^{(M)} \ \text{ versus } \ H_1 : \hat{F}^{(M)} \neq (F^\circ)^{(M)}, \tag{1.15}$$

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

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

and $(F^\circ)^{(M)}$ is a given discrete distribution function with $|\mathrm{supp}((F^\circ)^{(M)})| \leq n = n(M)$.

($|A|$ 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 (1.15) as well

as (1.5) is a sequence of statements and it remains some uncertainty how they should be combined. While it is quite natural to take $H_0 : \boldsymbol{\mu}^{(M)} = (\boldsymbol{\mu}^{\circ})^{(M)}\ \forall$ (sufficiently large) M, a reasonable definition of 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 1.1. *(cf. van Es, Klaassen, Mnatsakanov, 2003) Define $F_\rho(t) := \hat{F}^{(M)}(\rho t)$ with some scaling factor $\rho = \rho(M)$ and suppose that F_ρ converges weakly to some distribution function F as $M \to \infty$. Then F is called a structural distribution of the expected cell frequencies μ (or simply of the table $\boldsymbol{y}$) with the scaling factor ρ.*

Again, the sparsity scale is determined by ρ. For the multinomial sampling scheme, setting $\rho = N/n$ yields $\mathrm{supp}(F^{\circ}) \subset [0,\, 1]$.

Khmaladze (1988) have shown in case of $\rho := N/n = \mathcal{O}(1)$, the direct estimation of F by the empirical distribution function of y gives an inconsistent estimator of F. Consistent estimators of F have been constructed by Klaassen & Mnatsakanov (2000) and van Es et al. (2003). Their main assumptions are similar to that in the model of latent distribution (1.14). It reads as follows.

Let $U \sim Uniform[0,\, 1]$ and distribution densities g_M satisfy

$$\hat{F}^{(M)}(\rho t) = \mathbb{P}\{g_n(U) \le t\}, \quad \forall t \in \mathbb{R}_+,$$

and $\|g_n - g\|_\infty \to 0$ for some function g. Here $\|\cdot\|_\infty$ stands for the supremum norm.

Khmaladze (1988) pointed out that the structural distribution can be treated as a latent mixing distribution in the empirical Bayes approach. A consistent estimator of the mixing distribution of Poisson random variables is constructed in (Mnatsakanov and Klassen, (2007)).

In Section 3.1 we extend the empirical Bayes approach to include the null hypothesis in the Bayes model as well

In terms of the structural distribution the testing problem states

$$H_0 : F = F^{\circ} \ \text{ versus } \ H_1 : F \neq F^{\circ}, \tag{1.16}$$

where F° is a given distribution function with $\mathrm{supp}(F^{\circ}) \subset \mathbb{R}_+$.

Khmaladze (1988) 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

Inconsistency of common goodness-of-fit tests

In this chapter simple conditions for the inconsistency of the classical gopdness-of-fit tests in case of very sparse categorical data are given.

In the next section we introduce notation, present some background and specify a sparsity condition. The inconsistency of Pearson's χ^2 test is proved in Section 2.2. A simple example and simulation results provided illustrate the inconsistency and *reversed consistency* phenomena for a finite sample.

Current Chapter results were presented in *Computer Data Analysis and Modeling* conference and could be reviewed in (Radavičius and Samusenko 2010, Samusenko 2011) publications.

2.1. Notation and background

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

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

Let us assume that a simple hypothesis

$$H_0 : \boldsymbol{p} = \boldsymbol{p}^\circ \text{ versus } H_1 : \boldsymbol{p} \in \mathcal{P} \tag{2.1}$$

is to be tested on the basis of the observed frequencies $\boldsymbol{y}$ with a given $\boldsymbol{p}^\circ = (p_1^\circ, \ldots, p_n^\circ) > 0, \boldsymbol{p}^\circ \notin \mathcal{P}$.

We consider very sparse categorical data (contingency tables). Here it means that

$$n = n(M), \quad N = o(n), \quad \boldsymbol{p} = \boldsymbol{p}(M), \quad \boldsymbol{p}^\circ = \boldsymbol{p}^\circ(M) \quad (M \to \infty).$$

We shall also use additional (technical) conditions related to the sparseness, see Theorem 2.1.

In this case Perason's χ^2 statistic

$$X^2 := \sum_{j=1}^{n} \frac{(y_j - Np_j^\circ)^2}{Np_j^\circ} = \sum_{j=1}^{n} \frac{y_j^2}{Np_j^\circ} - N. \tag{2.2}$$

Using moment generation function one can find the means

$$\mathbb{E}X^2 = (N-1)\sum_{j=1}^{n} \frac{p_j^2}{p_j^\circ} + \sum_{j=1}^{n} \frac{p_j}{p_j^\circ} - N, \quad \mathbb{E}_\circ X^2 = n - 1, \tag{2.3}$$

and the variances

$$\begin{aligned}
\mathbb{D}X^2 &= \frac{1}{N}\sum_{j=1}^{n} \frac{p_j}{(p_j^\circ)^2} + 6\left(1 - \frac{1}{N}\right)\sum_{j=1}^{n}\left(\frac{p_j}{p_j^\circ}\right)^2 \\
&+ 4N\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\sum_{j=1}^{n} \frac{(p_j)^3}{(p_j^\circ)^2} \\
&- \frac{1}{N}\left(\sum_{j=1}^{n} \frac{p_j}{p_j^\circ}\right)^2 - \left(1 - \frac{1}{N}\right)\left(\sum_{j=1}^{n} \frac{p_j}{p_j^\circ}\right)\left(\sum_{j=1}^{n} \frac{(p_j)^2}{p_j^\circ}\right) \\
&- (4N-6)\left(1 - \frac{1}{N}\right)\left(\sum_{j=1}^{n} \frac{(p_j)^2}{p_j^\circ}\right)^2, \qquad (2.4) \\
\mathbb{D}_\circ X^2 &= \frac{1}{N}\sum_{j=1}^{n} \frac{1}{p_j^\circ} - \frac{n^2}{N} + 2(n-1)\left(1 - \frac{1}{N}\right) \qquad (2.5)
\end{aligned}$$

of the X^2 statistic. Here and in the sequel $\mathbb{E}, \mathbb{D}$, and $\mathbb{P}$ ($\mathbb{E}_\circ, \mathbb{D}_\circ$, and $\mathbb{P}_\circ$) denote, respectively, the expectation, the variance, and the probability for $Y \sim Multinomial_n(N, \boldsymbol{p})$ (respectively, $Y \sim Multinomial_n(N, \boldsymbol{p}^\circ)$.

2.2. Inconsistency of chi-square test under multinomial sampling

In this section the inconsistency of the X^2 statistic is derived under additional conditions related to and quite natural for (very) sparse categorical data.

Definition 2.1. *Let $T_N := T(S_N)$ be a statistic of a sample S_N with $N = N(M)$ being the sample size. A test (criterion) based on the statistic T_N is said to be* consistent *(as $M \to \infty$) for testing H_0 versus H_1 (2.1) with a given $\mathcal{P} = \mathcal{P}^M$ iff there exists a sequence c_M^* such that*

$$\mathbb{P}_\circ \{T_N > c_M\} + \mathbb{P}\{T_N < c_M\} \to 0, \quad \forall \boldsymbol{p} \in \mathcal{P}^M,\ M \to \infty.$$

Otherwise, the test is called inconsistent.

Theorem 2.1. *Suppose that, for some $\boldsymbol{p} \in \mathcal{P}^M$,*

$$\Delta_M := \mathbb{E}X^2 - \mathbb{E}_\circ X^2 = \sum_{j=1}^n \frac{p_j}{p_j^\circ} + (N-1)\sum_{j=1}^n \frac{p_j^2}{p_j^\circ} - N - (n-1) < 0, \quad (2.6)$$

and the asymptotic relation

$$\rho_M^2 := \frac{\Delta_M^2}{D_M^2} \to \infty \quad (M \to \infty) \quad (2.7)$$

is valid with $D_M := \sqrt{\mathbb{D}_\circ X^2} + \sqrt{\mathbb{D}X^2}$. Then the X^2 test is inconsistent.

On the other hand, the test based on the statistic $T = T_M := |X^2 - (n-1)|$ is consistent with $c_M = |\Delta_M|/2$ provided (2.7) holds for all $\boldsymbol{p} \in \mathcal{P}^M$.

Proof of Theorem 2.1. The Tchebyshev's inequality implies

$$\mathbb{P}_\circ \left\{X^2 \le \mathbb{E}_\circ X^2 - 2\sqrt{\mathbb{D}_\circ X^2}\right\} \le 1/4, \quad (2.8)$$

$$\mathbb{P}\left\{X^2 \ge \mathbb{E}X^2 + 2\sqrt{\mathbb{D}X^2}\right\} \le 1/4. \quad (2.9)$$

Consequently,

$$\begin{aligned}&\mathbb{P}_\circ\left\{X^2 > c_M\right\} + \mathbb{P}\left\{X^2 < c_M\right\} \geq \\ &\geq \mathbb{P}_\circ\left\{X^2 > \max(c_M, c_M^\circ)\right\} + \mathbb{P}\left\{X^2 < \min(c_M, c_M^*)\right\}, \quad (2.10)\end{aligned}$$

where $c_M^\circ := \mathbb{E}_\circ X^2 - 2\sqrt{\mathbb{D}_\circ X^2}$ and $c_M^* := \mathbb{E}X^2 + 2\sqrt{\mathbb{D}X^2}$. Since, in view of (2.6) and (2.7),

$$c_M^* - c_M^\circ = \Delta_M + 2\left(\sqrt{\mathbb{D}_\circ X^2} + \sqrt{\mathbb{D}X^2}\right) < 0$$

for all sufficiently large M, we then get $c_M^\circ \geq c_M^*$ and hence either $\max(c_M,\, c_M^\circ) = c_M^\circ$ or $\min(c_M,\, c_M^*) = c_M^*$. From (2.8), (2.9) and (2.10) we derive inconsistency of X^2 test:

$$\begin{aligned}&\mathbb{P}_\circ\left\{X^2 > c_M\right\} + \mathbb{P}\left\{X^2 < c_M\right\} \geq \\ &\max\left(\mathbb{P}_\circ\left\{X^2 > c_M^\circ\right\}, \mathbb{P}\left\{X^2 < c_M^*\right\}\right) \geq 3/4.\end{aligned}$$

The consistency of T_M follows from (2.3) and the Tchebyshev inequality:

$$\begin{aligned}&\mathbb{P}_\circ\left\{T_M > |\Delta_M|/2\right\} + \mathbb{P}\left\{T_M < |\Delta_M|/2\right\} \\ &\quad\leq \mathbb{P}_\circ\left\{T_M^2 > \Delta_M^2/4\right\} + \mathbb{P}\left\{|X^2 - \mathbb{E}_{\boldsymbol{p}} X^2| > |\Delta_M|/2\right\} \\ &\quad\leq 4\frac{\mathbb{E}_\circ T_M^2 + \mathbb{D}X^2}{\Delta_M^2} = \frac{4}{\rho_M^2} \to 0 \quad (M \to \infty)\end{aligned}$$

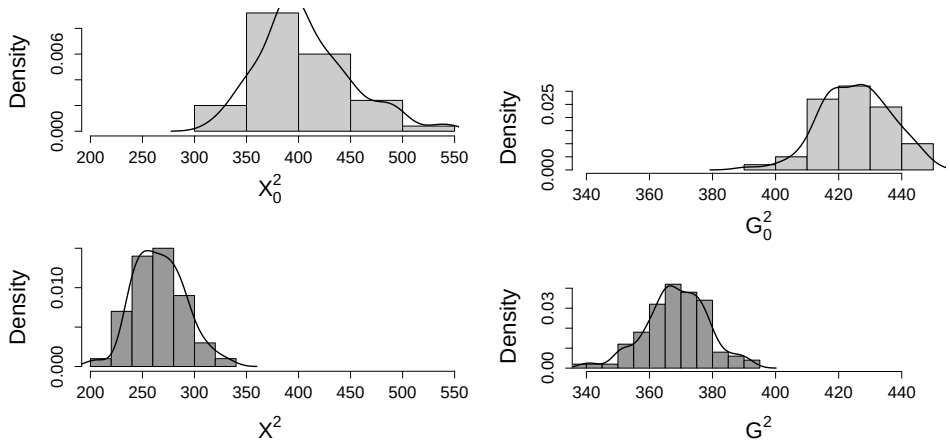
due to (2.7).

Theorem 2.1 shows that (2.6) is the key condition which determines the inconsistency of χ^2 test. When $\boldsymbol{p}^\circ$ is the uniform distribution, $\Delta \geq 0$ for any $\boldsymbol{p}$ and hence, for any $\boldsymbol{p}$, condition (2.6) is not satisfied. In the next section we present a simple example when conditions (2.6) and (2.7) are fulfilled.

Remark 2.1. By definition (2.6)

$$\Delta_M = \sum_{j=1}^n \frac{p_j - p_j^\circ}{p_j^\circ} + (N-1)\sum_{j=1}^n \frac{(p_j - p_j^\circ)^2}{p_j^\circ}. \qquad (2.11)$$

Since the second term in this expression is nonnegative the requirement $\Delta < 0$ implies that the absolute value of the first term in (2.11) should dominate second one.



a) X^2 ($q^\circ = 0{,}2,\ q = 0$) b) G^2 ($q^\circ = 0{,}2,\ q = 0$)

Fig. 2.1. Histograms of statistics under the hypothesis and under the alternative: a) X^2 statistic; b) G^2 statistic

Example

For a given $\beta > 1$ and $q^\circ, q \in (0,\ 1/2)$, set

$$M = N,\ m = \left[N^\beta\right],\ n = 2m,$$

$$p_j^\circ = q^\circ/m,\ \forall j \le m, \quad p_j^\circ = (1-q^\circ)/m,\ \forall j > m,$$
$$p_j = q/m,\ \forall j \le m, \quad p_j = (1-q)/m,\ \forall j > m.$$

Then the conditions of Theorem 2.1 are fulfilled.

If $q = 0$, means (2.3) and variances (2.5), (2.4) are given by

$$\begin{aligned}
\mathbb{E}X^2 &= \frac{N-1}{1-q^\circ} + \frac{m}{1-q^\circ} - N, \quad \mathbb{E}_\circ X^2 = n-1,\\
\mathbb{D}_\circ X^2 &= \frac{m^2}{Nq^\circ(1-q^\circ)} - \frac{n^2}{N} + 2(n-1)\left(1-\frac{1}{N}\right),\\
\mathbb{D}X^2 &= \frac{2(m-1)}{(1-q^\circ)^2}\left(1-\frac{1}{N}\right).
\end{aligned}$$

Consequently,

$$\Delta_N = -\frac{1-2q^\circ}{2-2q^\circ}\, n + \mathcal{O}(N),$$

$\mathbb{D}X^2 = \mathcal{O}(n)$, and

$$\mathbb{D}_{p^\circ} X^2 = \frac{n^2}{N}\left[\frac{1}{4q^\circ(1-q^\circ)} - 1\right] + \mathcal{O}(n).$$

Thus

$$\rho_N = -\sqrt{\frac{Nq^\circ}{1-q^\circ}}\left[1 + \mathcal{O}\left(\frac{N}{n}\right)\right].$$

A computer experiment illustrates the asymptotic findings in case of finite samples. In the simulations, the number of observations $N = 200$, the number of cells $n = 2m = 600$. Two cases are considered: (a) $q^\circ = 0{,}2,\ q = 0$ and (b) $q^\circ = 0{,}2,\ q = 0{,}1$. The number of repetitions is set to 100. The histograms of the X^2 statistic for the null hypothesis H_0 and the alternative H_1 are represented in Figure 2.1.

The figure clearly demonstrates the inconsistency of the X^2 statistic. Actually, in the first case (case (a)), the phenomenon of the reversed consistency is observed: although the values of the X^2 statistic under the null hypothesis H_0 are significantly greater than its values under the alternative H_1 (the data under the alternative *fits* the null hypothesis better than the data under the null hypothesis itself) the latter is evidently separable from the former. Thus Pearson's χ^2 test is completely uninformative in this case.

2.3. Inconsistency of likelihood ratio test under Poisson sampling

In this section 3.4 an adaptive procedure for nonparametric testing is described. and some simulation results are presented.

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, $j \in \{1, \ldots, n\}$, denote independent Poisson observations. Hence $\boldsymbol{y} := (y_1, \ldots, y_n) \sim Poisson_n(\boldsymbol{\mu})$ where $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n) \in \mathcal{M} := [0, M_0]^n$, $M_0 > 0$. We consider very sparse categorical data (contingency tables) $\boldsymbol{y} \in \mathbb{Z}_+^n$. Here it means that n is an asymptotic parameter, $\mathbb{E}_{\boldsymbol{\mu}}(\boldsymbol{y}) =$

$\boldsymbol{\mu} = \boldsymbol{\mu}(n)$ and as $n \to \infty$

$$\|\boldsymbol{\mu}\|_2^2 = o\left(\|\boldsymbol{\mu}\|_1\right). \tag{2.12}$$

Let us assume for simplicity that a simple hypothesis

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}^\circ \text{ versus } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}^\circ \tag{2.13}$$

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

$$\begin{aligned} G^2 \quad = \quad & G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y}) := 2\sum_{j=1}^n \left[y_j \log\left(\frac{y_j}{\mu_j^\circ}\right) + (\mu_j^\circ - y_j) \right], \\ & G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y}) =: 2H(\boldsymbol{y}) + 2L(\boldsymbol{\mu}^\circ, \boldsymbol{y}), \\ & H(\boldsymbol{y}) := \sum_{j=1}^n y_j \log(y_j), \\ & L(\boldsymbol{\mu}^\circ, \boldsymbol{y}) := \mu_+^\circ - \sum_{j=1}^n y_j(\log(\mu_j^\circ) + 1), \quad \mu_+^\circ := \sum_{j=1}^n \mu_j^\circ = \|\boldsymbol{\mu}\|_1. \end{aligned}$$

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

Lemma 2.1. *Assume sparsity* (2.12)*. Then ($n \to \infty$)*

$$\mathbb{E}_{\boldsymbol{\mu}} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y}) = 2\mathbb{E}_{\boldsymbol{\mu}} L(\boldsymbol{\mu}^\circ, \boldsymbol{y}) + \mathcal{O}\left(\|\boldsymbol{\mu}\|_2^2\right),$$

$$\sqrt{\mathbb{D}_{\boldsymbol{\mu}} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y})} = 2\sqrt{\mathbb{D}_{\boldsymbol{\mu}}(L(\boldsymbol{\mu}^\circ, \boldsymbol{y}))} + \mathcal{O}\left(\|\boldsymbol{\mu}\|_2\right),$$

$$\mathbb{E}_{\boldsymbol{\mu}} L(\boldsymbol{\mu}^\circ, \boldsymbol{y}) = \mu_+^\circ - \sum_{j=1}^n \mu_j(\log(\mu_j^\circ) + 1),$$

$$\mathbb{D}_{\boldsymbol{\mu}}(L(\boldsymbol{\mu}^\circ, \boldsymbol{y})) = \sum_{j=1}^n \mu_j(\log(\mu_j^\circ) + 1)^2.$$

Proof of Lemma 2.1. To prove the lemma it suffices to note that for any $\beta > 0$

$$\mathbb{E}_{\boldsymbol{\mu}} \sum_{j=1}^n \left(y_j \log(y_j)\right)^\beta = \mathcal{O}(\|\boldsymbol{\mu}\|_2^2).$$

Proposition 2.1. *Suppose that* $\boldsymbol{\mu}^\circ \in \mathcal{M}_+$, $\boldsymbol{\mu} \in \mathcal{M}$,

$$\Delta_n = \Delta_n(\boldsymbol{\mu}) := \sum_{j=1}^{n} (\mu_j - \mu_j^\circ)(\log(\mu_j^\circ) + 1) \geq 0, \tag{2.14}$$

and

$$\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu}^\circ\|_2^2 = o(D_n^2(\boldsymbol{\mu}) + D_n^2(\boldsymbol{\mu}^\circ)), \tag{2.15}$$

$$D_n^2(\boldsymbol{\mu}) := \sum_{j=1}^{n} \mu_j \left(\log(\mu_j^\circ) + 1\right)^2.$$

If (2.12) *holds, then*

$$\frac{\mathbb{E}_{\boldsymbol{\mu}} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y}) - \mathbb{E}_{\boldsymbol{\mu}^\circ} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y})}{\left(\mathbb{D}_{\boldsymbol{\mu}^\circ} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y}) + \mathbb{D}_{\boldsymbol{\mu}} G^2(\boldsymbol{\mu}^\circ, \boldsymbol{y})\right)^{1/2}} = -\frac{\Delta_n + \mathcal{O}(\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu}^\circ\|_2^2)}{\left(D_n^2(\boldsymbol{\mu}) + D_n^2(\boldsymbol{\mu}^\circ)\right)^{1/2}(1 + o(1))}.$$

Corollary 2.1. *For very sparse contingency tables* (see (2.12)), the LR test is inconsistent for testing problem (2.13) provided (2.14) and (2.15) hold and

$$\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu}^\circ\|_2^2 = o(\Delta_n), \quad D_n^2(\boldsymbol{\mu}^\circ) + D_n^2(\boldsymbol{\mu}) \leq (\kappa + o(1))\Delta_n^2, \ \kappa < 1.$$

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

Example. Let $n = 2\tilde{n}$, $\mu_i^\circ = \mu_i^\circ(n) = o(1)$, $i = 1, 2$; $0 < \mu_1^\circ < \mu_2^\circ$, $\rho \in (0, 1)$, and

$$\mu_j^\circ = \mu_1^\circ, \ \forall j \leq \tilde{n}, \quad \mu_j^\circ = \mu_2^\circ, \ \forall j > \tilde{n},$$

$$\mu_j = (1 - \rho)\mu_1^\circ, \ \forall j \leq \tilde{n}, \quad \mu_j = \mu_2^\circ + \rho\mu_1^\circ, \ \forall j > \tilde{n}.$$

Then

$$\Delta_n = \frac{\rho\mu_1^\circ n}{2} \log\left(\frac{\mu_2^\circ}{\mu_1^\circ}\right) > 0,$$

$$D^2(\boldsymbol{\mu}^\circ) \asymp D^2(\boldsymbol{\mu}) \asymp n\mu_2^\circ(\log(\mu_2^\circ))^2.$$

Note that $\|\boldsymbol{\mu}^\circ\|_1 = \|\boldsymbol{\mu}\|_1$ and $\|\boldsymbol{\mu}^\circ\|_2^2 + \|\boldsymbol{\mu}\|_2^2 \leq (\mu_1^\circ + 2\mu_2^\circ)\ \|\boldsymbol{\mu}^\circ\|_1 = o\left(\|\boldsymbol{\mu}^\circ\|_1\right)$. Thus, the conditions of Corollary are fulfilled if $\mu_1^\circ \leq \rho_1\mu_2^\circ$, $\rho_1 \in (0, 1)$,

$$(\mu_2^\circ)^2 = o\left(\mu_1^\circ |\log(\mu_1^\circ)|\right),$$

$$\frac{\sqrt{\mu_2^\circ}|\log(\mu_2^\circ)|}{\mu_1^\circ\sqrt{n}} = o(1).$$

Remark 2.2. Actually, the inconsistency stated in Corollary 2.1 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, Read (1984).

2.4. Conclusions of the second chapter

1. The classical tests are no longer (asymptotically) distribution free.
2. The classical tests became noninformative: they are inconsistent even in cases where a simple consistent tests does exist.
3. For very sparse categorical data, common goodness-of-fit tests may be inconsistent and hence there is no sense to approximate their distributions.

3

Hypotheses testing for sparse categorical data

The goal of this chapter is to propose consistent nonparametric criteria in case of sparse categorical data.

In the next section, an extended empirical Bayes model of sparse asymptotics is introduced. This model contains the latent distribution and the structural distribution models as special cases. In Section 2, the testing problem is formulated without any assumptions about convergence of distributions. The consistency of tests based on phi-divergences and grouping is proved. In Section 3.3 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.

Proposed consistent nonparametric criteria in case of sparse categorical data and modeling results were published in (Radavičius and Samusenko 2011).

3.1. 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 $P = P^{(M)}$ where M is an asymptotic parameter, $M \to \infty$. Then the marginal distribu-

tion P° of γ° (respectively, P^γ of γ) coincides with the structural distribution under the null hypothesis H_0 (respectively, under the alternative H_1), see (Khmaladze, 1988).

Fix M or set $M = \infty$. Now the testing problem for structural distribution (1.16) takes the following form:

$$H_0 : P^\circ = P^\gamma \text{ versus } H_1 : P^\circ \neq P^\gamma.$$

Thus in this case only the marginal distributions of P are involved.

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

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

Then problem (1.5) can be extended in terms of P as follows:

$$H_0 : P^\gamma_\circ(\cdot \mid a) = \delta_a \; \forall a \in \Omega \text{ versus } H_1 : P^\gamma_\circ(\cdot \mid a) \neq \delta_a \; \forall a \in A.$$

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

Note that this extension of (1.5) 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.

3.2. Goodness-of-fit criteria based on grouping

Here we use the extended empirical Bayes framework described in 3.1.

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:

$$H_0 : \forall\, M,\ d(Q^{(M)},\, P^{(M)}) = 0, \tag{3.1}$$

versus

$$H_1 : \forall\ M,\ d(Q^{(M)},\ P^{(M)}) \geq \delta(M). \tag{3.2}$$

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

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

$$\mathbb{P}_Q^{(M)} \left\{ \widehat{d}(Q;\ P) > (1 - \tau(M))\delta(M) \right\} \to 0, \quad M \to \infty, \tag{3.3}$$

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

$$\mathbb{P}_P^{(M)}\ \left\{ \widehat{d}(Q;\ P) < d(Q^{(M)};\ P^{(M)}) - \tau(M)\delta(M) \right\} \to 0, \quad M \to \infty. \tag{3.4}$$

Lemma 3.1. *Assume that Assumption C is valid. Then the criterion*

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

is consistent as $M \to \infty$ for testing (3.1) versus (3.2).

Proof of Lemma 3.1. Write $\tau = \tau(M),\ \delta = \delta(M)$ for short. If H_0 is valid,

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

due to (3.3). If H_1 holds, then $d(Q^{(M)};\ P^{(M)}) \geq \delta$ and hence

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

by (3.4).

In order to apply Lemma 3.1 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 $M \to \infty$.

Grouping

The observed data is

$$\mathcal{D} = \mathcal{D}^{(M)} := \{(\mu_i^\circ,\, y_i),\ i = 1, \ldots,\, n\},$$

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 μ_i, and $\{(\gamma_i^\circ,\, \gamma_i),\ i = 1, \ldots,\, n\}$ are iid with the common distribution $P^{(M)}$.

Let $\Delta = \Delta^{(M)} := \{\Delta_k,\, k = 1, \ldots,\, K\}$ 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 $t_0 = 0,\ t_{K-1} < \mu_+^\circ \le t_K < \infty$.

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

$$\mu_{+j}^\circ = \sum_{i=1}^{j} \mu_i^\circ, \quad \mu_{+0}^\circ = 0,$$

and the sequences determined by the partition Δ,

$$\begin{aligned}
\mu_{k+}^\circ &= \sum_{i=1}^{n} \mu_i^\circ\, \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}, \\
\mu_{k+} &= \sum_{i=1}^{n} \mu_i\, \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}, \\
y_{k+} &= \sum_{i=1}^{n} y_i\, \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}.
\end{aligned}$$

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

$$\{(\mu_i^\circ, \mu_i^\circ),\ i = 1, \ldots,\, n\}, \tag{3.5}$$

and

$$\{(\mu_i^\circ, \mu_i),\ i = 1, \ldots,\, n\}, \tag{3.6}$$

respectively. The discrepancy between $Q^{(M)}$ and $P^{(M)}$ is measured by ϕ-

divergence for the grouped data:

$$d(Q^{(M)};\, P^{(M)}) = d_\phi(Q^{(M)};\, P^{(M)}) := \sum_{k=1}^{K} \mu^\circ_{k+} \phi(\mu_{k+}/\mu^\circ_{k+}). \tag{3.7}$$

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

$$\widehat{d}(Q;\, P) := \sum_{k=1}^{K} \mu^\circ_{k+}\, \phi(y_{k+}/\mu^\circ_{k+}). \tag{3.8}$$

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

$$\mathbb{E}\phi^2(\eta_u/v) < \infty \quad \forall\, u, v > 0. \tag{3.9}$$

Denote

$$a(v) := v\, \mathbb{E}\phi(\eta_v/v), \tag{3.10}$$

$$\sigma^2(v;\, u) := v^2\, \mathbb{E}(\phi(\eta_u/v) - \phi(u/v))^2. \tag{3.11}$$

Lemma 3.2. *Suppose (3.9) is fulfilled. Then*

$$\mathbb{E}_P\, \widehat{d}(Q;\, P) \geq d\left(Q^{(M)};\, P^{(M)}\right), \tag{3.12}$$

$$\mathbb{E}_Q\, \widehat{d}(Q;\, P) = A(M) := \sum_{k=1}^{K} a(\mu^\circ_{k+}), \tag{3.13}$$

$$\mathbb{D}_P\, \widehat{d}(Q;\, P) \leq V^2(M) := \sum_{k=1}^{K} \sigma^2(\mu^\circ_{k+},\, \mu_{k+}). \tag{3.14}$$

Proof of Lemma 3.2.

Since the function $\phi(u/v)$ is convex with respect to u inequality (3.13) follows from Jensen's inequality. Further, in view of (3.11)

$$v^2 \mathbb{D}\phi(\eta_u/v) \leq v^2 \mathbb{E}(\phi(\eta_u/v) - \phi(u/v))^2 = \sigma^2(v,\, u).$$

Consequently,

$$\begin{aligned}\mathbb{D}_P\, \widehat{d}(P^\circ;\, P) &= \sum_{k=1}^{K} \left(\mu^\circ_{k+}\right)^2 \mathbb{D}_P\, \phi^2(y_{k+}/\mu^\circ_{k+}) \\ &\le \sum_{k=1}^{K} \sigma^2(\mu^\circ_{k+},\, \mu_{k+}),\end{aligned}$$

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

Consistency

From Lemma 3.2 it easy to derive the following result.

Theorem 3.1. *Let $Q^{(M)}$ and $P^{(M)}$ be the empirical distributions based on the data (3.5) and (3.6), respectively. Suppose (3.9) is fulfilled and the discrepancy measure between $Q^{(M)}$ and $P^{(M)}$ is defined by (3.7) and estimated by (3.8). If*

$$V_0(M) + V(M) = o(\delta(M) - A(M)), \quad M \to \infty, \tag{3.15}$$

where $A(M)$, $V(M)$ are introduced in (3.13), (3.14) and

$$V_0^2(M) := \sum_{k=1}^{K} \sigma^2(\mu^\circ_{k+}, \mu^\circ_{k+}), \tag{3.16}$$

then the criterion

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

is consistent as $M \to \infty$ for testing (3.1) versus (3.2) with any constant $\kappa_1 \in (0,\, 1)$.

Proof of Theorem 3.1.

Let us check the first condition of Lemma 1 (3.3). Set $\tau(M) = (1-\kappa_1)(1-A(M)/\delta(M))$. Then (3.15), (3.13), (3.14), (3.16), and Chebyshev inequality

imply

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

Similarly, for the second condition of Lemma 1 (3.4), we derive from (3.15), (3.12), (3.14), and Chebyshev inequality

$$\begin{aligned} &\mathbb{P}_P\left\{\widehat{d}(Q;\,P) < d(Q^{(M)};\,P^{(M)}) - \tau(M)\delta(M)\right\} \\ &\quad\leq \mathbb{P}_P\left\{\widehat{d}(Q;\,P) - \mathbb{E}_P\,\widehat{d}(Q;\,P) < -\tau(M)\delta(M)\right\} \\ &\quad\leq \mathbb{P}_P\left\{\widehat{d}(Q;\,P) - \mathbb{E}_P\,\widehat{d}(Q;\,P) < -(1-\kappa_1)(\delta(M)-A(M))\right\} \\ &\quad\leq \frac{V_0^2(M)}{(1-\kappa_1)^2(\delta(M)-A(M))^2} \to 0. \end{aligned}$$

The proof is complete.

Remark 3.1. 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 (3.8) is asymptotically normal as $M \to \infty$. This fact can be established by arguments of Györfi, Vajda, (2002) 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 4.4 we present (provide) some computer simulation results to illustrate performance of the proposed criterion.

3.3. Profile statistics

Let us assume that $\{J_m,\, m = 1, \ldots,\, M\}$ is a partition of set $\{1, \ldots,\, n\}$ into disjoint subsets such that $\mu_j^\circ = \mu_m^\circ,\ j \in J_m, m = 1, \ldots,\, M$, with some $\mu_m^\circ = \mu_m^\circ(n) \in (0,\, M_0]$. Suppose that all alternatives with any $\boldsymbol{\mu}$ obtained via permutations of the coordinates within $J_m,\, m = 1, \ldots,\, M$, 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 $\boldsymbol{\mu}$ to be a sequence of random variables exchangeable within the each set J_m.

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

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

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

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

In this setting, the null hypothesis in (1.5) can be restated as follows:

$$H_0^G :\ G_m = \delta_{\mu_m^\circ}, \quad m = 1, \ldots,\, M. \tag{3.17}$$

Here δ_a stands for the degenerate distribution centered on a. The LLR statistic for (3.17) is given by

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

$$\eta_l(m) := |\{y_j, j \in J_m : y_j = l\}|. \tag{3.18}$$

Hence the statistic $\boldsymbol{\eta} = \{\eta(m),\ m = 1, \ldots,\, M\}$ with $\eta(m) := (\eta_l(m),\ l \in \mathbb{Z}_+),\ m = 1, \ldots,\, M$, is a sufficient statistic for G. Under the Poisson sampling, $\boldsymbol{\eta}$ distribution is a product of M multinomial distributions with the infinite number of outcomes, the probabilities of outcomes equal to $\pi_{\mathbb{Z}_+}(\mu_m^\circ) :=$

$(\pi_l(\mu_m^\circ),\ l \in \mathbb{Z}_+)$, and $n_m := |J_m|$ independent trials $(m = 1, \ldots, M)$:

$$\eta(m) \sim Multinomial_{\mathbb{Z}_+}(n_m,\, \pi_{\mathbb{Z}_+}(\mu_m^\circ)).$$

Components (3.18) of the statistic $\boldsymbol{\eta}$ 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 behaviour of η_m in the case of multinomial sampling have been investigated, for instance, by Kolchin, Sevastyanov, Chistyakov (1978). The profile statistics are also related to estimating problem of the structural distribution function of cell probabilities Es, Klaassen, Mnatsakanov (2003).

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

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

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

Theorem 3.2. *Suppose that* $\boldsymbol{\mu}^\circ \in \mathcal{M}_+$ *satisfies sparsity condition* (1.10) and $\mu_j^\circ = \mu_m^\circ, j \in J_m,\ m = 1, \ldots, M$. Then, for any $t = t(n),\ t/\log(t) > \max(M_0,\ \|\boldsymbol{\mu}^\circ\|_1)$,

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\ell(\hat{G}) \ge t\} \le H(t)\ \mathrm{e}^{-t/2} \tag{3.19}$$

where

$$\begin{aligned} H(t) \;&:=\; \left| K\left(\left[\frac{t}{\log(t)}\right] + 1\right)\right|^M + n \exp\left\{\frac{t(\log\log(t) + \log(M_0) + 1)}{\log(t)}\right\} \\ &+\; \exp\left\{\frac{t(\log(\|\boldsymbol{\mu}^\circ\|_1) + \log\log(t) + 1)}{\log(t)} - \|\boldsymbol{\mu}^\circ\|_1\right\} \end{aligned} \tag{3.20}$$

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

Proof of Theorem 3.2. Since

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

the inequality

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\ell(\hat{G}) \ge t\} \le \mathbb{P}_{\boldsymbol{\mu}^\circ}\{\hat{\ell}(\boldsymbol{\eta}) \ge t\} \tag{3.21}$$

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

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\boldsymbol{\eta} = \boldsymbol{k}\} \le \exp\{-(1/2)\ \hat{\ell}(\boldsymbol{k})\}.$$

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

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\hat{\ell}(\boldsymbol{\eta}) \ge t\} \le |K(s)|^M \exp\{-t/2\} + \mathbb{P}_{\boldsymbol{\mu}^\circ}\{\boldsymbol{\eta}^+ \notin K(s)\}. \tag{3.22}$$

Denote

$$\eta_+ := \sum_{l=1}^{\infty} \eta_l^+ = \sum_{j=1}^{n} \mathbb{1}\{y_j > 0\} \le \|\boldsymbol{y}\|_1.$$

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

$$\log\left(\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\eta_+ > s\}\right) \le s \log\left(\frac{\|\boldsymbol{\mu}^\circ\|_1}{s}\right) + s - \|\boldsymbol{\mu}^\circ\|_1. \tag{3.23}$$

Similarly, for $s > M_0$,

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\left\{\max_{j\in\{1,\ldots,n\}} y_j > s\right\} \le \sum_{j=1}^{n} \mathbb{P}_{\boldsymbol{\mu}^\circ}\{y_j > s\} \le n\ \exp\left\{s \log\left(\frac{M_0}{s}\right) + s\right\}. \tag{3.24}$$

Note that $\eta_+ \le s$ and $\max_{j\in\{1,\ldots,n\}} y_j \le s$ imply $\boldsymbol{\eta}^+ \in K(s)$. Hence,

$$\mathbb{P}_{\boldsymbol{\mu}^\circ}\{\boldsymbol{\eta} \notin K(s)\} \le \mathbb{P}_{\boldsymbol{\mu}^\circ}\{\eta_+ > s\} + \mathbb{P}_{\boldsymbol{\mu}^\circ}\{\max_{j\in\{1,\ldots,n\}} y_j > s\}. \tag{3.25}$$

Take $s = [t/\log(t)] + 1$. Then inequality (3.19) with $H(t)$ given in (3.20) follows from (3.21)–(3.25). The well-known fact that $\log|K(s)| = \mathcal{O}(s)$ as $s \to \infty$ completes the proof.

In the next section a flexible and adaptive procedure taking advantage of soft clustering in an auxiliary mixture model is described.

3.4. Likelihood ratio test with soft clustering

Here it is assumed that the both parameters, $\boldsymbol{\mu}^\circ$ and $\boldsymbol{\mu}$, are sequences of independent identically distributed random variables satisfying a semi-parametric

mixture model with a dummy class variable $\nu_j \in \{1, \dots, M\},\ j \in \{1, \dots, n\}$. Specifically,

$$\mathbb{P}\{\nu_j = m\} = p_m \geq 0, \quad \sum_{m=1}^{M} p_m = 1; \tag{3.26}$$

$$\begin{aligned} (\mu_j^\circ \mid \nu_j = m) &\sim LogNormal(a_m,\, \sigma_m), \quad j \in \{1, \dots, n\}, && (3.27)\\ (\mu_j \mid \nu_j = m) &\sim G_m, \quad j \in \{1, \dots, n\}, && (3.28)\\ (y_j \mid \mu_j) &\sim Poisson(\mu_j), \quad j \in \{1, \dots, n\}, && (3.29)\\ & \quad m = 1, \dots, M. && (3.30) \end{aligned}$$

Let

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

be a collection of the parameters of the mixture. Notice that the values of $\boldsymbol{\mu}$ are unobservable (latent). The observed data is $(y_j,\, \mu_j^\circ),\ j \in \{1, \dots, n\}$. Suppose that μ_j° and μ_j are conditionally, given ν_j, independent, and y_j, given μ_j, is independent of the rest random variables ($j \in \{1, \dots, n\}$). 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 $y_{max} := \max_{j \in \{1, \dots, n\}} y_j$. For sparse data, y_{max} is small. Thus, the probabilities $\pi_l(\hat{G}_m),\ l \in \mathbb{Z}_+$, are expressed as the finite mixture of Poisson distributions. Consequently, the initial semi-parametric model defined in (3.26)–(3.30) 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 $p_m(\hat{\theta} \mid y_j,\, \mu_j^\circ)$ be the estimated posterior probability of the unobserved class number ν_j, given the observation $(y_j,\, \mu_j^\circ)$,

$$p_m(\theta \mid y_j,\, \mu_j^\circ) := \mathbb{P}_\theta\{\nu_j = m \mid y_j,\, \mu_j^\circ\}, \quad j \in \{1, \dots, n\}, \quad m = 1, \dots, M.$$

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

$$\begin{aligned} \hat{\eta}_l(m) &:= \sum_{j=1}^{n} \mathbb{1}\{y_j = l\}\, p_m(\hat{\theta} \mid y_j, \mu_j^{\circ}), \\ \hat{\pi}_l^{\circ}(m) &:= \sum_{j=1}^{n} \pi_l(\mu_j^{\circ})\, p_m(\hat{\theta} \mid y_j,\, \mu_j^{\circ}). \end{aligned}$$

The symmetric logarithmic likelihood ratio (LLR) statistic based on soft clustering and the empirical Bayes approach is defined by

$$\mathcal{L}(\hat{\theta} \mid \boldsymbol{y}) := \sum_{m=1}^{M} \sum_{l=1}^{y_{max}} (\hat{\eta}_l(m) - \hat{\pi}_l^{\circ}(m)) \left(\log(\pi_l(\hat{G}_m)) - \log(\pi_l(\exp\{\hat{a}_m\})) \right). \tag{3.31}$$

The performance of the criterion for testing (1.5) based on $\mathcal{L}(\hat{\theta} \mid \boldsymbol{y})$ is illustrated by simulations.

Computer experiment. The framework of the example in Section 2.3 is adopted. The parameters $\mu_1^{\circ} = 0{,}5,\ \mu_2^{\circ} = 1,\ \mu_1 = \mu_1(i) = \mu_1^{\circ} - 0{,}05(i-1)$, $\mu_2 = \mu_2(i) = \mu_2^{\circ} + 0{,}05(i-1),\ i = 1, \ldots, 10$, $n = 2\tilde{n} = 40$, the number of simulations is equal 100. The parameters σ_m are kept fixed, $\sigma_m = 0{,}5,\ m = 1, \ldots, M$. 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 (3.31) 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 (Radavičius, Samusenko, 2011, p. 122). 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.

3.5. Conclusions of the third chapter

1. A extended empirical Bayes model of sparse asymptotics introduced in this chapter contains the latent distribution and the structural distri-

bution models as special cases.

2. Under general conditions the nonparametric criteria based on based on phi-divergences and grouping as well as likelihood ratio criterion based on profile statistics and maximum likelihood estimator are consistent.

3. The proposed likelihood ratio criterion based on the profile statistics can be viewed as a composite likelihood ratio test for homogeneous groups of cells obtained via hard clustering.

4

Modeling results

The goal of this chapter is to compare the finite-sample behavior of some classical goodness-of-fit tests and the proposed nonparametric criteria based on the grouping method as well as MCMC smoothing algorithm.

There is a vast literature devoted to simulation studies of sparse contingency tables. Some of them are mentioned in Section 4.1 In Section 4.2, goodness-of-fit criteria to be compared are specified and two new criteria based on gamma (gamma-weighted) grouping and MCMC smoothing, respectively, are introduced. In Section 4.3, specific models for sparse categorical data simulation are described. The main simulation results are presented and discused in Section 4.4.

Comparison of some classical goodness-of-fit tests, proposed nonparametric criteria and MCMC smoothing algorithm with some modeling results were published in (Radavičius and Samusenko 2012).

4.1. Overview of experiments done before

There are many sources where Monte Carlo algorithms have been applied to show positive and negative effects of new method in comparison with other already known methods. It enables one to enumerate a random subset

of all the possible outcomes in the reference set when the exact approach is computationally infeasible or it's execution time is unpredictably long. Monte Carlo sampling is a good compromise for handling large, sparse tables to estimate precisely the inferential characteristics of interest, such as exact p-values and confidence intervals. Some examples with simulating the conditional hypergeometric Sampling distribution can be found in (Agresti, Wackerly and Boyett, 1979; Boyett, 1979; Cox and Plackett, 1980; Patefield, 1981, 1982; Kreiner, 1987; StatXact, 2011).

Accuracy of various approximations to finite-sample distributions of (classical) goodness-of-fit test has been investigated via simulations by Koehler and Larntz, (1980), Cressie and Reed, (1984,1988), Haberman (1988), Hu, (1999), Finkler (2010), among others.

Additionally we can find some examples described in statistical applications user manuals, textbooks and on internet. Below we introduce some of them.

Several statistical applications created to compute an exact p-value for a data set whose sample size can be very high. StatXact 5 is one which has great written user manual with real examples inside. It's introduces you to Exact Nonparametric Inference (also known as Permutational Inference) by discussing how exact tests are defined, why they are difficult to compute and the role of Monte Carlo methods in producing reliable inference (StatXact, 2011).

Some of textbooks on nonparametric methods, for example, Manly (1991), Sprent (1993), Good (1993), Edgington (1995) and Agresti (1990) devote considerable space to exact and Monte Carlo methods of inference for categorical data.

Introduction to our simulations

Classical Goodness-of-fit criterions and MCMC smoothing methods are based on means. In our presented grouping methods means, variance or their combinations could be used. In case of *goodness-of-fit of means* we will call them (means), *goodness-of-fit of variances* (variances) and *goodness-of-fit of means and variances* (both).

4.2. Compared goodness-of-fit tests

Goodness-of-fit tests considered in this section is based on power-divergence statistics with $\alpha \in \{1, 0, -1, -2\}$, respectively $\lambda \in \{2, 1, 0, -1\}$. The proposed criteria use the same power-divergence statistics, however they

used grouped or smoothed categorical data. To improve classical tests (see formula 3.8) three grouping rules are applied to cells of a contingency tables ordered in μ° increasing order:

1. Grouping into groups with (approximately) equal expected counts (referred to as GC method).
2. Grouping into groups with (approximately) equal number of cells (referred to as GQ method; group bounds are quantiles of the empirical structural distribution function F°).
3. grouping into groups with approximately equal number of cells and weighting with weights determined by Gamma distribution (referred to as GG method, see Subsection 4.2.1).

Usually goodness-of-fit statistics are based on discrepancies between average values in each cell (category) or group. However, for alternatives with (only) irregular deviations from null hypotheses (see *Two step with variance* model in Subsection 4.3), criteria using such statistics have low power which (as results of Section 4.4 shows) even decreases when smoothing effect increases. This suggests to use discrepancies between group variances instead of averages when calculating goodness-of-fit statistics and also to construct the omnibus test by summing the both goodness-of-fit statistics. These modifications of the tests based on grouping are referred to by *mean*, *var*, and *both*, respectively.

The goodness-of-fit test based on the MCMC (Markov chain Monte Carlo) smoothing is a specific non-random (i.e. averaged) version of the semiparametric smoothing algorithm, see Faddy and Jones, (1998), Radavičius and Židanavičiūtė, (2009). Further comments and details of the algorithm are presented in Subsection 4.2.2.

The methods for improving the classical tests are summarized in Table 4.1.

4.2.1. Grouping and gamma weighing

The goodness-of-fit statistics for the test GG is constructed in the same way as for the other tests based on grouping. The difference is only in using weights when calculating observed and expected values to be compared.

More precisely, let Δ be the partition consisting of K disjoint intervals with approximately equal number of μ° values and let m_k be the medians of $\{\mu_i^\circ : \mu_{+i}^\circ \in \Delta_k,\ i = 1, \ldots, n\}$, $k = 1, \ldots, K$. Let $g(\cdot \mid a, v), a, v > 0$, denote a probability density of gamma distribution with a mean a and a

Table 4.1. Method used in computer experiments section

Name	Description
GC	Grouping of equal expected frequencies • K – number of groups
GQ	Grouping of equal group sizes • K – number of groups
GG	Gama kernel smoothing for groups of equal group sizes • K – number of groups
MCMC	smoothing (averaging) algorithm • $q \in (0,\, 1)$ – percent of data to be updated • k – number of updating (averaging) iterations

variance v. Fix a smoothing parameter $h > 0$ and define

$$\begin{aligned} w_k(i) &:= W_k^{-1} g(\mu_i^\circ \mid m_k,\, m_k h), \quad \mu_{+i}^\circ \in \Delta_k,\ i = 1, \ldots,\, n, \\ W_k &:= \sum_{i=1}^{n} g(\mu_i^\circ \mid m_k,\, m_k h) \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}, \quad k = 1, \ldots,\, K, \\ \bar{\mu}_k &:= \sum_{i=1}^{n} \mu_i w_k(i)\, \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}, \\ \bar{y}_k &:= \sum_{i=1}^{n} y_i w_k(i)\, \mathbb{1}\{\mu_{+i}^\circ \in \Delta_k\}. \end{aligned}$$

Then the goodness-of-fit statistics for the test GG is given by (for a specified function ϕ)

$$\widehat{d(Q;\, P)} := \sum_{k=1}^{K} \bar{\mu}_k^\circ\, \phi(\bar{y}_k / \bar{\mu}_k^\circ). \tag{4.1}$$

4.2.2. Test based on Markov chain Monte Carlo smoothing

The construction of the test using MCMC (Markov chain Monte Carlo) smoothing follows the general line used for the previous tests. Actually, it is just another way to obtain smoothed values of μ° and y. The procedure to calculate these values are obtained in two steps:

- Markov chain Monte Carlo (a local Gibbs) sampler is built up which (asymptotically) generates random contingency tables with the expected counts μ°,
- instead of generating a new random value in the Markov chain, its con-

dition distribution is used to calculate the conditional expectation of this value.

The MCMC smoothing procedure depends on r ir k:

- $r \in (0,\, 1)$, percent of data will be update.
- k number of cycles.

MCMC smoothing algorithm:

1. Observed and expected frequencies ordered ascending by expected (under the H_0) frequencies μ°.
2. *Temporal values* calculation:

$$y_+ := y_j + y_{j+1}, \quad \mu_+ := \mu_j^\circ + \mu_{j+1}^\circ, \quad p := \mu_j^\circ / \mu_+^\circ, \quad j = 1 \div n-1.$$

3. *New values* y^{new} calculation:

$$\begin{aligned} y_j^{new_1} &:= (1-r)y_j + pry_+, \quad j = 1 \div n-1, \\ y_{j+1}^{new_2} &:= (1-r)y_j + (1-p)ry_+. \end{aligned}$$

4. *Update* original values y:

$$y_j := \frac{y_j^{new_1} + y_j^{new_2}}{2},\ j = 2 \div n-1; \quad y_1 := y_1^{new_1},\ y_n := y_n^{new_2}.$$

5. Steps 2, 3 and 4 repeat k times.

Overall smoothing effect depends on the product $r \cdot k$.

4.3. Models for sparse contingency table simulation

Three different types of models (see Table 4.2) have been chosen to compare the power of the classical tests and their improved versions presented in the previous subsections.

For all simulated models, the expected cell counts both for the null hypothesis and the alternative are generated as independent Gamma random variables:

$$\mu_i \sim Gamma(a(i),\, v(i)), \quad \mu_i^\circ \sim Gamma(a^\circ(i),\, v^\circ(i)), \quad i = 1, \ldots,\, n. \tag{4.2}$$

Table 4.2. Descriptions of models used in computer experiments.

Name		Description
Two step	(2S)	both μ and μ° are take only two different values, and have close overall averages
Top split	(TS)	μ differs from μ° in the region of high values of μ° ("Top")
Bottom split	(BS)	μ differs from μ° in the region of low values of μ° ("Bottom")
Two step with variance (irregularity)	(2SV)	averages of μ are the same as μ° but μ has high (random) variability (irregularity)

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

"Two step" model is one of the simplest sparse data models used in Monte Carlo studies (see, for instance, Finkler (2010)). It is used in Chapter 2 to illustrate inappropriateness (non-informativity) of the classical tests.

In "Two step" model, the expected cell counts are approximated (a small noise is added) by two-step functions both under the null hypothesis H_0 and the alternative H_1. More precisely, (4.2) is applied with $v^\circ(i) = v(i) \equiv 0{,}01^2$,

$$a^\circ(i) = 0{,}5 \cdot \mathbb{1}\{i \le n/2\} + 1{,}6 \cdot \mathbb{1}\{i > n/2\},$$

$$a(i) = a_{2S03}(i) := 0{,}2 \cdot \mathbb{1}\{i \le n/2\} + 1{,}9 \cdot \mathbb{1}\{i > n/2\},$$

$$a(i) = a_{2S05}(i) := 0{,}001 \cdot \mathbb{1}\{i \le n/2\} + 2{,}1 \cdot \mathbb{1}\{i > n/2\},$$

Graphical illustration of "Two step" model is given in (a) of Fig 4.1.

"Top split" and "Bottom split" models allow one to compare power of the tests in cases where the alternative H_1 differs from the null hypothesis H_0 in a region of larger expected cell counts ("Top split" model) versus cases where this difference is in a region of smaller expected cell counts ("Bottom split" model). In "Top split" model an additional step is added in the region of larger expected cell counts (keeping the average value in the region approximately the same). In "Bottom split" model an additional step is added in the region of smaller expected cell counts (keeping the average value in the region approximately the same). Thus, (4.2) is applied with the the same as in (4.3)

average a°, $v^\circ(i) = v(i) \equiv 0{,}01^2$,

$$a^\circ(i) = 0{,}5 \cdot \mathbb{1}\{i \leq n/2\} + 1{,}6 \cdot \mathbb{1}\{i > n/2\} + 0{,}001i,$$

$$a(i) = a_{BS03}(i) := 1{,}6 - 0{,}8 \cdot \mathbb{1}\{i \leq n/2\} - 0{,}6 \cdot \mathbb{1}\{i \leq n/4\},$$

and

$$a(i) = a_{TS06}(i) := 0{,}5 + 0{,}5 \cdot \mathbb{1}\{i > n/2\} + 1{,}2 \cdot \mathbb{1}\{i > 3n/4\}.$$

Graphical illustration of "Bottom split" and "Top split" models are given in (a) of Fig 4.6 and Fig 4.4, respectively.

"Two step with variance" models are examples with irregular (random) behavior of the expected cell frequencies under the alternative H_1 while keeping the averages of the expected cell frequencies under the alternative H_1 close to that under the null hypothesis H_0. This is achieved by setting the average a° and the variance a° the same as in the previous models and

$$a(i) = a_{2SV025}(i) := 0{,}5 \cdot \mathbb{1}\{i \leq n/2\} + 1{,}6 \cdot \mathbb{1}\{i > n/2\},$$

with variance $v(i) \equiv 0{,}5^2$. Graphical illustration of "Two step with variance" model is given in (a) of Fig. 4.10.

In computer simulations, the size n of contingency tables to be generated is taken to be equal 200 and the Poisson sampling scheme is applied.

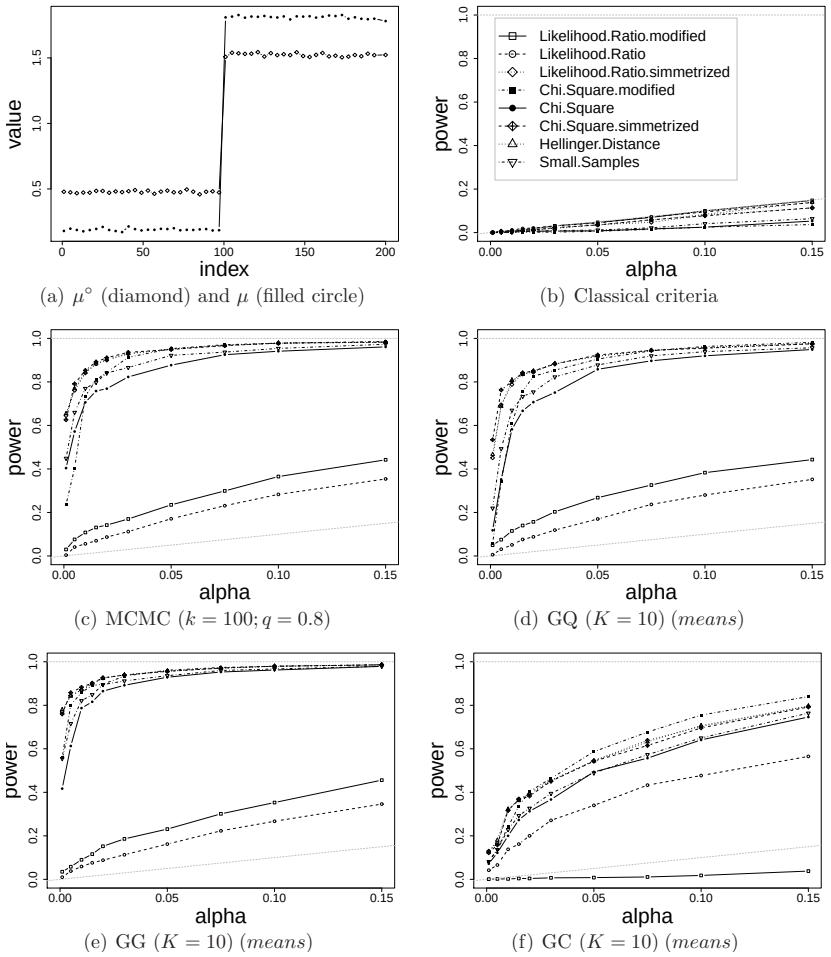
4.4. Computer experiment results

Only a part of the computer simulation results are presented here. Four the most successful in some simulated models criteria are chosen: G_s^2, X_s^2, F^2 and X_m^2. Tables of results, for these criteria, of Monte Carlo study with $R = 1000$ replications are given in Appendix, some of them are discussed in the next subsections.

4.4.1. Two step models

In the "Two step" model 2S03, the expected counts μ under the alternative H_1 differ from the expected counts μ° under the null hypothesis H_0 by $0{,}3$ for all cells (Fig. 4.1(a)). The powers of the tests based on mean discrepan-

cies (mean) and on both mean and variance discrepancies (both) are shown in Figures 4.1 and 4.2, respectively.



(a) μ° (diamond) and μ (filled circle)

(b) Classical criteria

(c) MCMC ($k = 100; q = 0.8$)

(d) GQ ($K = 10$) ($means$)

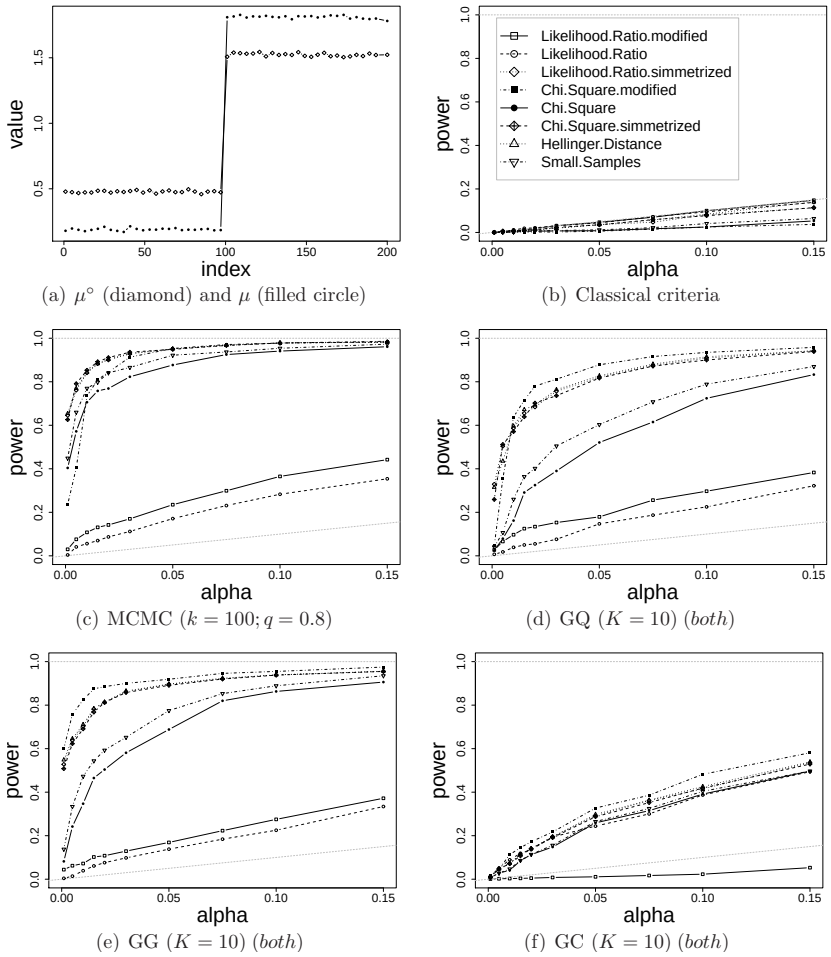
(e) GG ($K = 10$) ($means$)

(f) GC ($K = 10$) ($means$)

Fig. 4.1. Goodness-of-fit (means) tests power of MCMC, GQ, GG and GC are compared in 2S03 model

The classical tests have very low power (Figures 4.1(b) and 4.2(b)). The best methods are GG grouping (mean) (Fig. 4.1(e)) and MCMC smoothing (Fig. 4.1(c)). The goodness-of-fit test based on the symmetrized χ^2 statistic X_s^2, the symmetrized likelihood ratio statistic G_s^2, and the Hellinger distance F^2 show the similar performance and are the best. The numerical values of power of symmetrized χ^2 statistic X_s^2 and symmetrized likelihood ratio statistic G_s^2 are presented in Table A.2 and Table A.1, respectively. The bold figures

in the tables indicate the three highest powers.



(a) μ° (diamond) and μ (filled circle)

(b) Classical criteria

(c) MCMC ($k = 100; q = 0.8$)

(d) GQ ($K = 10$) ($both$)

(e) GG ($K = 10$) ($both$)

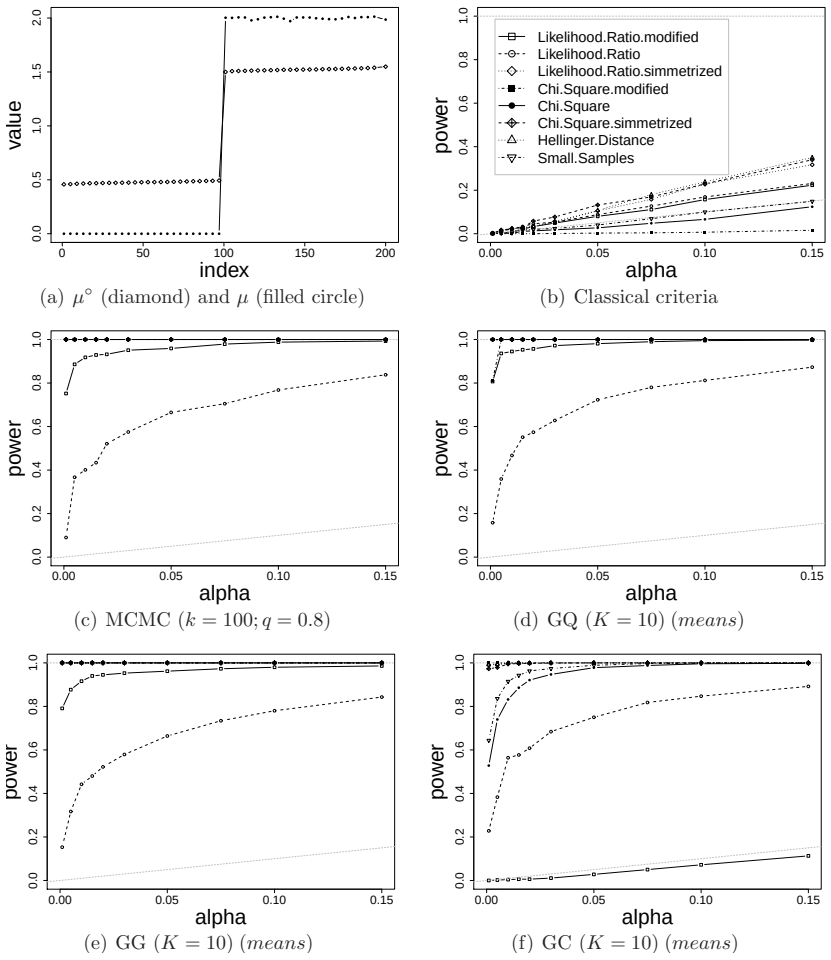
(f) GC ($K = 10$) ($both$)

Fig. 4.2. Goodness-of-fit (both) tests power of MCMC, GQ, GG and GC are compared in 2S03 model

In this case, deviations of the alternative H_1 from the null hypothesis H_0 are rather regular. As a result, variance discrepancies are irrelevant for testing goodness-of-fit: powers of the tests based on both mean and variance discrepancies are significantly worse than powers of the tests based on mean discrepancies alone. For goodness-of-fit statistics (both), the modified χ^2 statistic X^2_m seems to be the best (Fig. 4.2, line with filled squares; Table A.4).

The dependence of the tests power on MCMC smoothing parameters k and q is illustrated in Fig. A.1. The dependence of the power of the GG tests

(i.e., tests based on grouping by GG method) on the group number K is shown in Fig. A.3.



(a) μ° (diamond) and μ (filled circle)

(b) Classical criteria

(c) MCMC ($k = 100; q = 0.8$)

(d) GQ ($K = 10$) ($means$)

(e) GG ($K = 10$) ($means$)

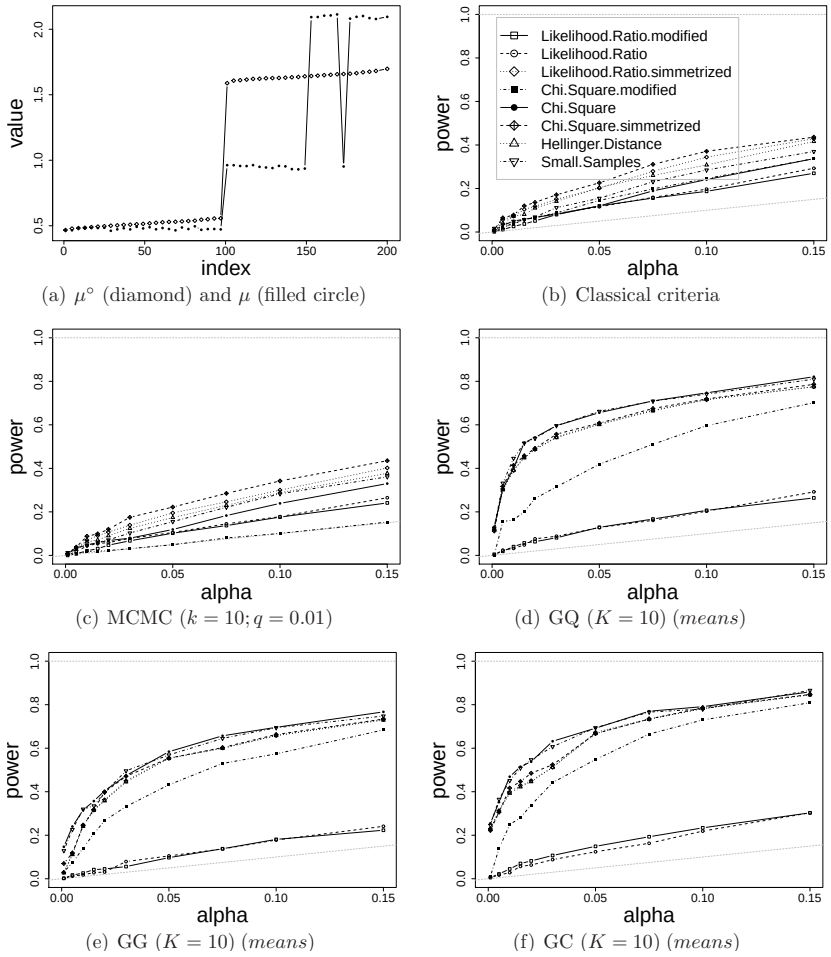
(f) GC ($K = 10$) ($means$)

Fig. 4.3. Goodness-of-fit (means) tests power of MCMC, GQ, GG and GC are compared in 2S05 model

In the "Two step" model 2S05, the expected counts μ under the alternative H_1 differ from the expected counts μ° under the null hypothesis H_0 by $0{,}5$ for all cells (Fig. 4.3(a)). The simulation results in this case are similar as in the model 2S03 but powers are much higher, see Figure 4.3 and Table A.5 as an example.

4.4.2. Split models

"Top split" model TS06. In the "Top split" model TS06, the values of μ differs from that of μ° by 0,6 for all cells with high values of μ° and μ is generated with the same variance $v(i) \equiv 0{,}01^2$ as μ° (Fig. 4.4(a)).



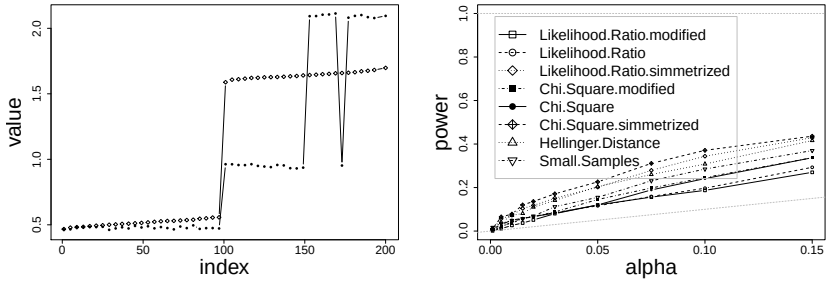
(a) μ° (diamond) and μ (filled circle) (b) Classical criteria

(c) MCMC ($k = 10; q = 0.01$) (d) GQ ($K = 10$) ($means$)

(e) GG ($K = 10$) ($means$) (f) GC ($K = 10$) ($means$)

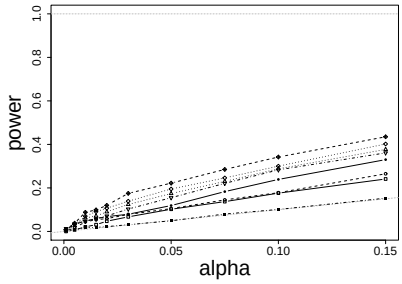
Fig. 4.4. Goodness-of-fit (means) tests power of MCMC, GQ, GG, GC are compared for TS06 model

The classical tests and the tests using MCMC smoothing have very low power (Figures 4.4(b,c) and 4.5(b,c)). The best method is GC grouping (mean) (Fig. 4.4(e)). The goodness-of-fit test based on the χ^2 statistic X^2, the symmetrized χ^2 statistic X_s^2, the symmetrized likelihood ratio statistic G_s^2, and

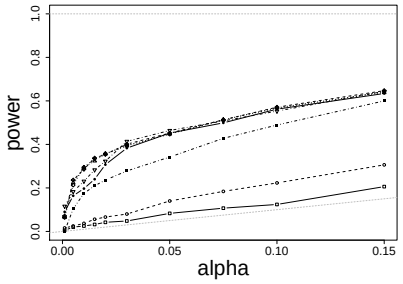


(a) μ° (diamond) and μ (filled circle)

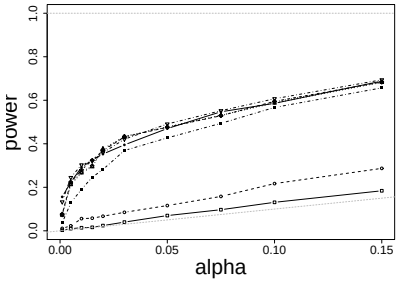
(b) Classical criteria



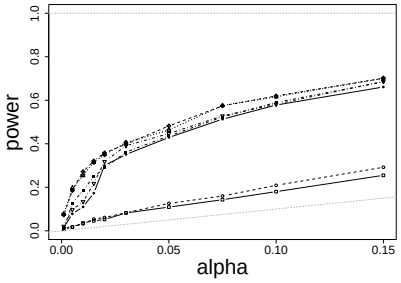
(c) MCMC ($k = 10; q = 0.01$)



(d) GQ ($K = 10$) ($both$)



(e) GG ($K = 10$) ($both$)



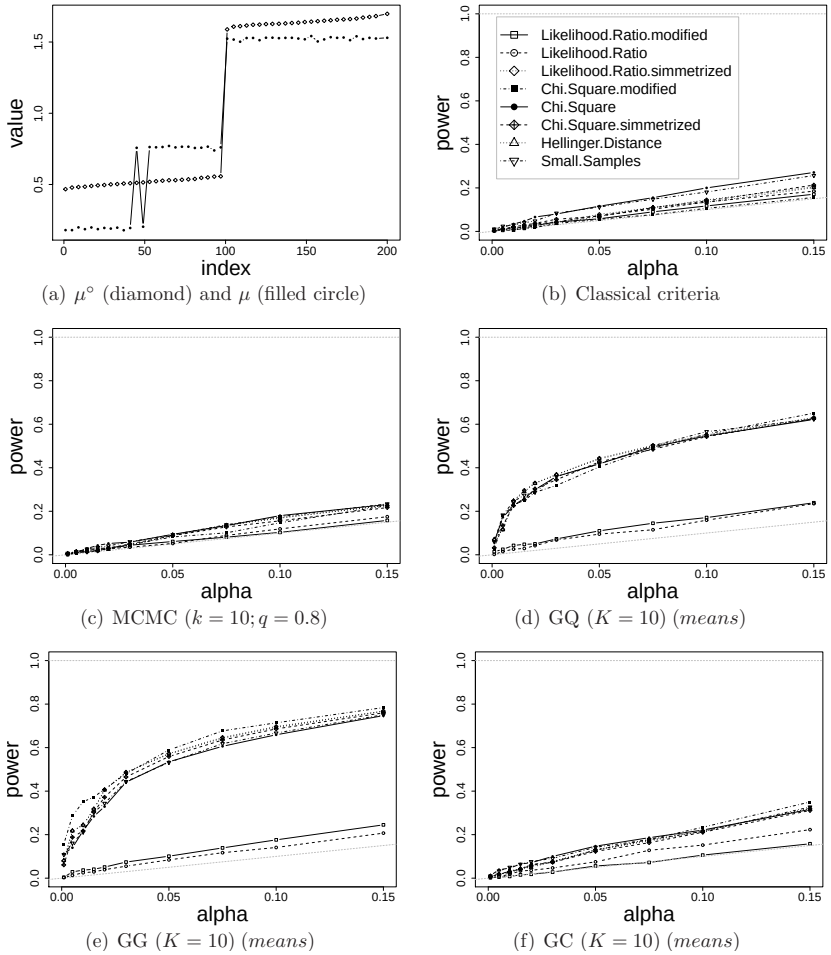
(f) GC ($K = 10$) ($both$)

Fig. 4.5. Goodness-of-fit (both) tests power of MCMC, GQ, GG, GC are compared for TS06 model

the Hellinger distance F^2 show the similar performance with X^2 being the best. The numerical values of power of χ^2 statistic X^2 and symmetrized likelihood ratio statistic G_s^2 are presented in Tables B.1 and B.2, respectively. The powers of the tests based on both mean and variance discrepancies are significantly worse than the powers of the tests based on mean discrepancies alone (Fig. 4.4). The dependence of the power of the GC tests on the group number K is shown in Figure B.4.

"Bottom split" model BS03. In the "Bottom split" model BS03, the values of μ differs from that of μ° by 0,3 for all cells with low values of μ° and μ

is generated with the same variance $v(i) \equiv 0{,}01^2$ as μ° (Fig. 4.6(a)).



(a) μ° (diamond) and μ (filled circle) (b) Classical criteria

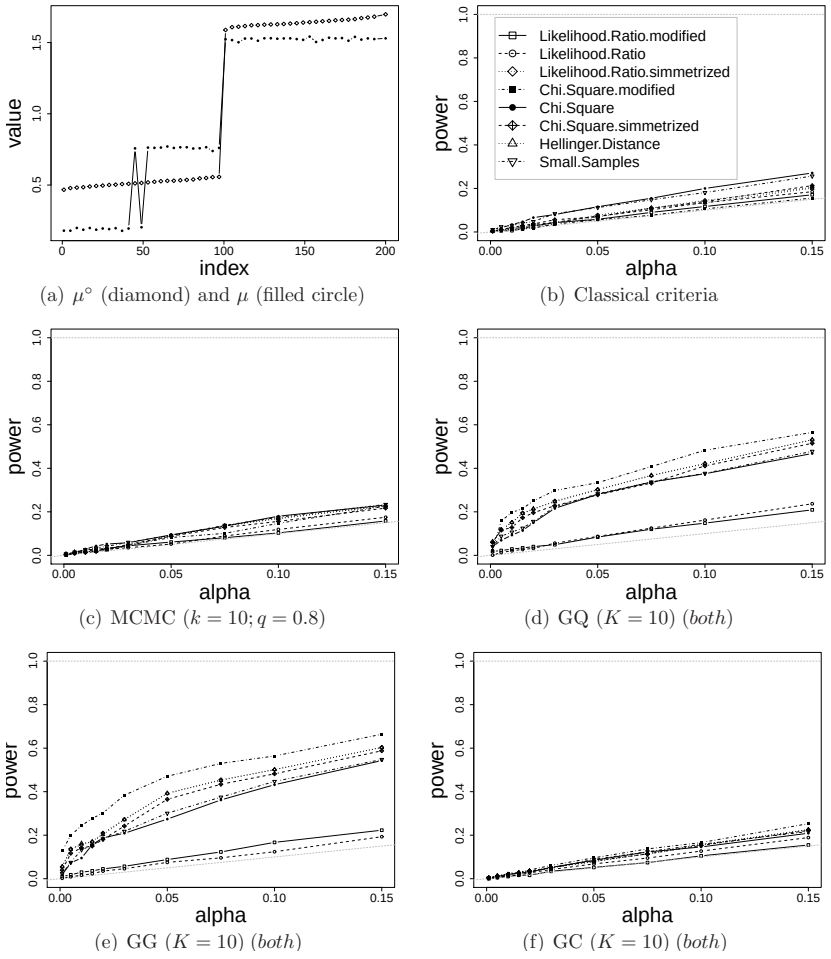
(c) MCMC ($k = 10; q = 0.8$) (d) GQ ($K = 10$) ($means$)

(e) GG ($K = 10$) ($means$) (f) GC ($K = 10$) ($means$)

Fig. 4.6. Goodness-of-fit (means) tests power of MCMC, GQ, GG, GC are compared for BS03 model

The classical tests and the tests using MCMC smoothing have very low power (Figures 4.6(b,c) and 4.7(b,c)). The best method is GG grouping (mean) (Fig. 4.6(e)). The goodness-of-fit test based on the modified χ^2 statistic X_m^2, the symmetrized χ^2 statistic X_s^2, the symmetrized likelihood ratio statistic G_s^2, and the Hellinger distance F^2 show the similar performance with X_m^2 being the best. The numerical values of power of modified χ^2 statistic X_m^2 and symmetrized likelihood ratio statistic G_s^2 are presented in Table B.9 and Table B.6, respectively. The powers of the tests based on both mean and variance dis-

crepancies are significantly worse than the powers of the tests based on mean discrepancies alone (Fig. 4.6).



(a) μ° (diamond) and μ (filled circle)

(b) Classical criteria

(c) MCMC ($k = 10; q = 0.8$)

(d) GQ ($K = 10$) ($both$)

(e) GG ($K = 10$) ($both$)

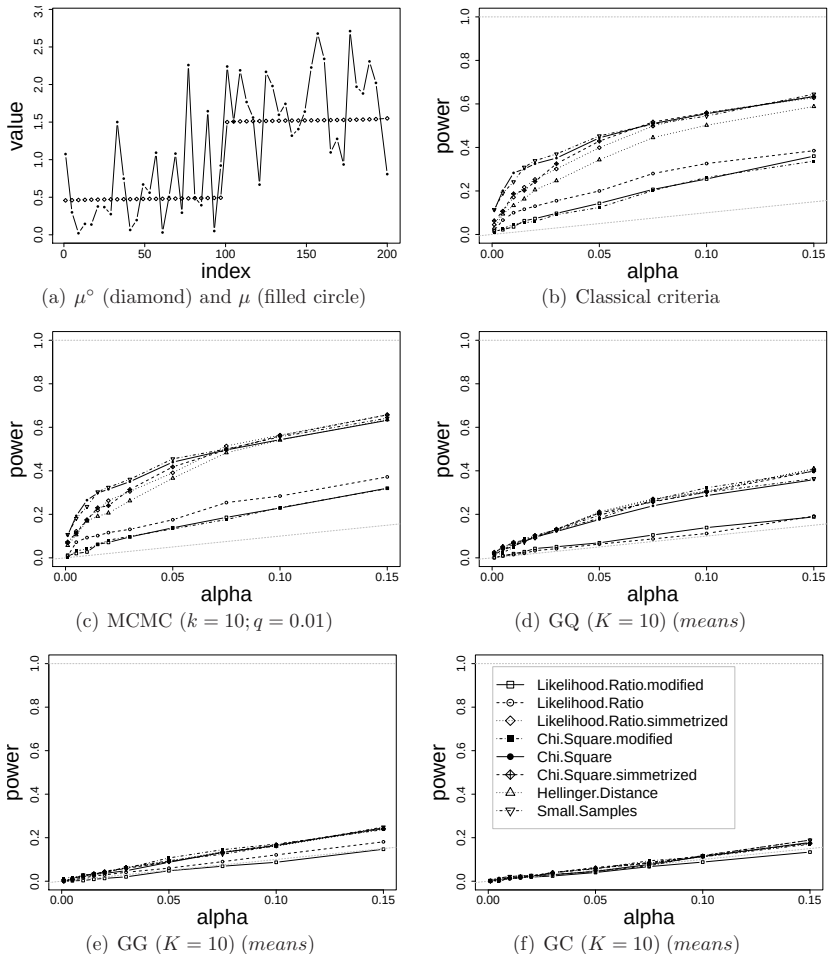
(f) GC ($K = 10$) ($both$)

Fig. 4.7. Goodness-of-fit (both) tests power of MCMC, GQ, GG, GC are compared for BS03 model

The dependence of the power of the GG tests on the group number K is shown in Figure B.7.

4.4.3. Irregular model

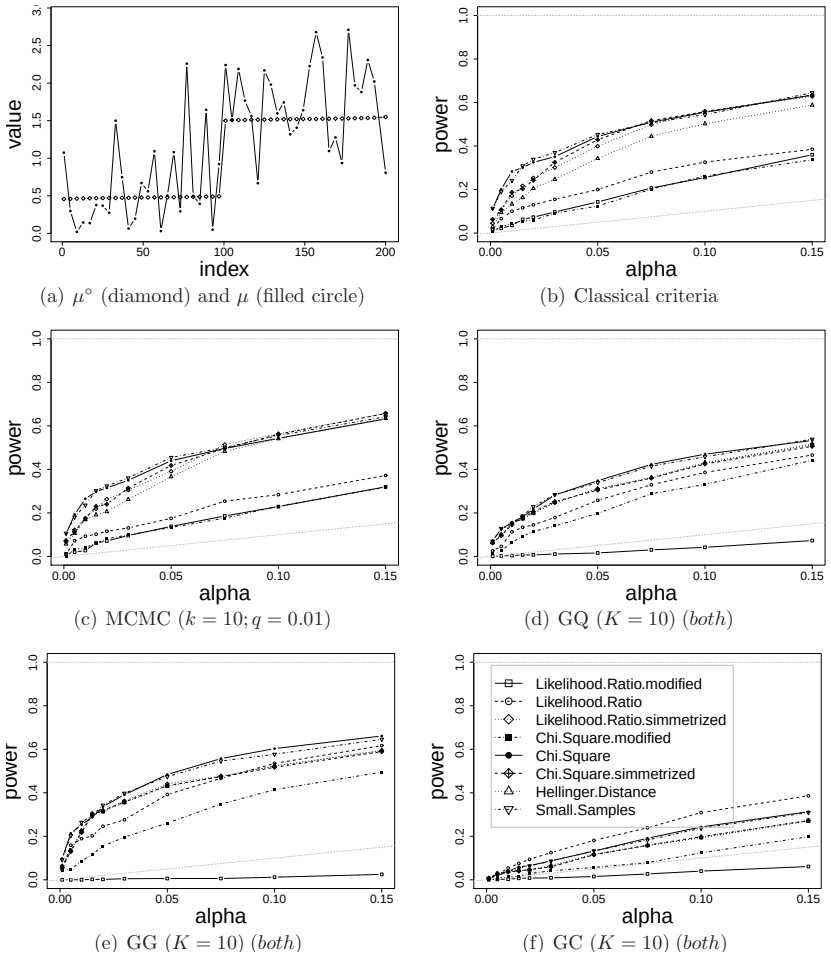
In the "Two step with variance" model 2SV025 the average values of μ and μ° are the same ($a(i) \equiv a^\circ(i)$, see Subsection 4.3) but μ has the variance $v(i) = 0{,}5^2$ (Fig. 4.8(a)). The powers of the tests based on mean discrepancies (mean) and on both mean and variance discrepancies (both) are shown in Figures 4.8 and 4.9, respectively. Here the number of groups in the tests based on grouping $K = 10$.



(a) μ° (diamond) and μ (filled circle)
(b) Classical criteria
(c) MCMC ($k = 10; q = 0.01$)
(d) GQ ($K = 10$) ($means$)
(e) GG ($K = 10$) ($means$)
(f) GC ($K = 10$) ($means$)

Fig. 4.8. Goodness-of-fit (means) tests power of classical criteria MCMC, GQ, GG and GC with $K = 10$ are compared in 2SV025 model

In contrast to the previous simulation results, the classical tests outper-



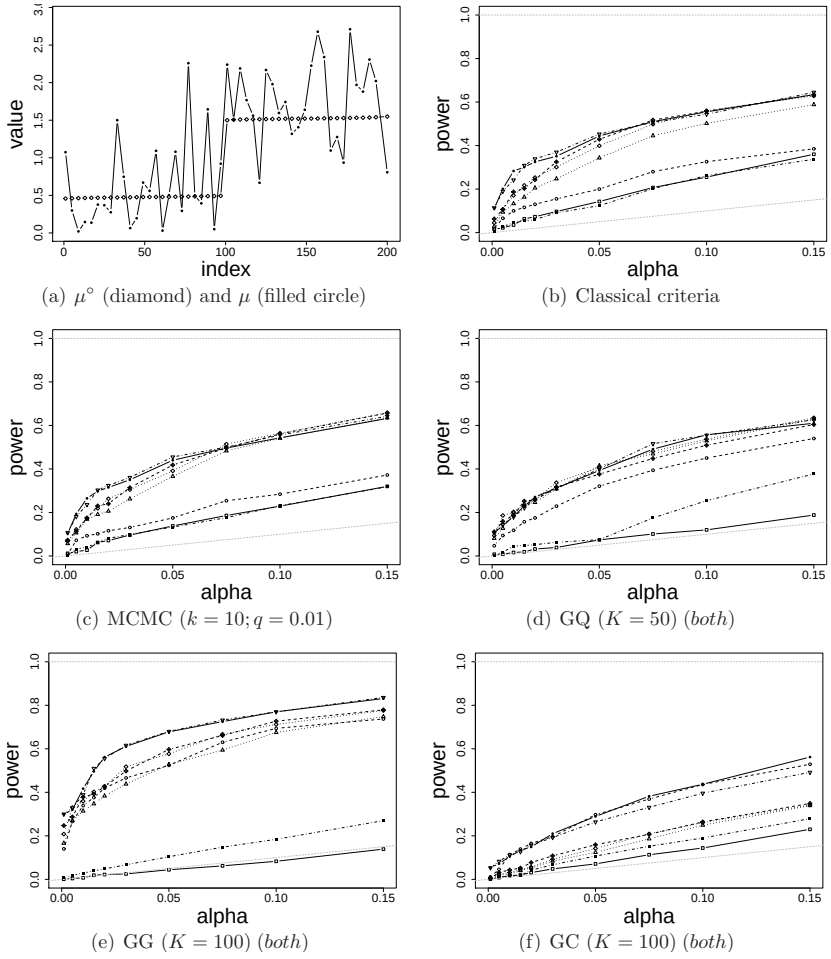
(a) μ° (diamond) and μ (filled circle)
(b) Classical criteria
(c) MCMC ($k = 10; q = 0.01$)
(d) GQ ($K = 10$) ($both$)
(e) GG ($K = 10$) ($both$)
(f) GC ($K = 10$) ($both$)

Fig. 4.9. Goodness-of-fit (both) tests power of classical criteria MCMC, GQ, GG and GC with $K = 10$ are compared in 2SV025 model

forms the tests which use grouping and the discrepancies only between the means (Fig. 4.8(b) versus Fig. 4.8(d,e,f)). Note that in this case the MCMC smoothing is the best method (Fig. 4.8(c)). Taking into account also the discrepancies between the variances significantly improves the power of tests based on grouping (Fig. 4.9). The GG grouping (both) (i.e., based on the discrepancies between both the means and variances) is slightly than the MCMC smoothing (Fig. 4.9 (e) versus Fig. 4.9 (c)).

The dependence of the tests power on MCMC smoothing parameters k and q is presented in Fig. C.1. In turn, the dependence of the power of the GG

grouping tests (GC grouping tests) on the group number K and the type of the discrepancies (mean), (var) or (both) is shown in Fig. C.3 (Fig. C.3).



(a) μ° (diamond) and μ (filled circle)
(b) Classical criteria
(c) MCMC ($k = 10; q = 0.01$)
(d) GQ ($K = 50$) ($both$)
(e) GG ($K = 100$) ($both$)
(f) GC ($K = 100$) ($both$)

Fig. 4.10. Goodness-of-fit (means) tests power of classical criteria, MCMC, GQ, GG and GC are compared in 2SV025 model

In Figure 4.10, the power graphs of the classical tests, the MCMC tests, and the grouping tests GQ (both), GG (both), GQ (both) with the best number of groups $K \in \{10, 50, 100\}$ are presented. The greatest power has the grouping tests GG (both), followed by the grouping tests GQ (both) and the MCMC smoothing test.

In the 2SV025 model, the χ^2 statistic X^2 (lines with filled circles) and the small sample statistic $D_{2/3}$ (lines with point down triangles) seems to be

the best, the symmetrized likelihood ratio statistic G_s^2 (lines with diamonds) and symmetrized χ^2 statistic X_s^2 (lines with filled diamonds) have the notably lower power, see also Tables C.1, C.2, C.3 and C.4.

4.5. Conclusions of the forth chapter

The overview of the simulation results lead to the following concluding remarks:

1. For very sparse contingency tables, the classical tests may have very low power.
2. For regular alternatives, when the expected cell counts under the alternative smoothly depends on the expected cell counts under the null hypothesis, grouping or smoothing may considerably increase the power of the goodness-of-fit tests.
3. The effect of grouping (smoothing) significantly depends on the grouping (smoothing) method as well as on its parameters (number of groups, number of iterations, etc.). Three grouping methods have been considered: grouping into groups of equal expected counts (GC), grouping into groups of equal size (GQ), and grouping into groups of equal size and weighting by making use of Gamma density (GG). Among these grouping methods, the GG method demonstrates the best overall performance.
4. For the irregular alternatives that differ from the null hypothesis by centered independent Gamma random variables (noise), the grouping tests which use the discrepancies only between the means have low power which usually decreases when the number of groups decreases. In this case, the grouping tests which use the discrepancies between the variances have much better power. This suggests omnibus tests which take into account the discrepancies between both the means and the variances.
5. The MCMC smoothing, as distinct from the grouping methods, works well for the irregular alternatives, however fails for the split models.
6. As noted in the previous studies, for sparse data unlike the standard case, The goodness-of-fit criteria based on various divergences may have quite different power. Our results confirm this observation.

General conclusions

Having solved the tasks listed in the introduction the following results were obtained:

1. For (very) sparse data, the likelihood ratio statistic and Pearson's χ^2 statistic may become noninformative: they do not anymore measure the goodness-of-fit of null hypotheses to data.
2. Sparse asymptotics based on (extended) empirical Bayes approach enables one to apply distribution model to sparse nominal data.
3. Under general conditions, the tests based on grouping are consistent.
4. The effect of grouping (smoothing) significantly depends on the grouping (smoothing) method as well as on its parameters (number of groups, number of iterations, etc.).
5. In the empirical Bayes setting, MCMC smoothing, smoothing by grouping and modeling by finite mixtures of Poisson distributions can improve the power of classical tests especially for regular alternatives.
6. For the irregular alternatives that differ from the null hypothesis by centered independent Gamma random variables (noise), the grouping tests which use the discrepancies between both the means and the variances have much better power.

References

Aerts, M.; Augustynas, I.; Jansen P. 1997. Sparse consistency and smoothing for multinomial data, *Statistics and Probability Letters* 33: 41–48.

Aerts, M.; Augustynas, I.; Jansen P. 2000. Central limit theorem for the total squared error of local polynomial estimators of cell probabilities, *Journal of Statistical Planning and Inference* 91: 181–193.

Agresti, A. 1990. *Categorical Data Analysis*. Wiley & Sons, New York. 558 p.

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

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

Agresti, A. 1992. A Survey of Exact Inference for Contingency Tables, *Statistical Science* 7.1: 131–153.

Agresti, A.; Hitchcock, B. D. 2005. Bayes inference for categorical data analysis, *Statistical Methods and Applications* 14: 297–330.

Agresti, A.; Wackerly, D.; Boyett, J. 1979. Exact conditional tests for cross-classifications: Approximation of attained significance levels, *Psychometrika* 44: 75–83.

Finkler, A. 2010. Goodness of fit statistics for sparse contingency tables, *Mathematics Statistics Theory* 14: 297–330. `http://arxiv.org/pdf/1006.2620v1.pdf`

Bishop, Y. M.; Fienberg, S. E.; Holland, P. W. 1975. *Discrete Multivariate Analysis.*

Theory and Practice. The MIT Press, Cambridge.

Boyett, J. 1979. Random R $\times$ C tables with given row and column totals, *Journal of the Royal Statistical Society* 28: 329–332.

Coull, B. A.; Agresti, A. 2003. Generalized log-linear models with random effects, with application to smoothing contingency tables, *Statistical Modelling* 3: 251–271.

Cressie, N.; Read, T. 1984. Multinomial Goodness of Fit Tests, *Journal of the Royal Statistical Society* 46: 440–464.

Congdon, P. 2005. *Bayesian Models for Categorical Data*. New York: John Wiley and Sons, Inc. 425 p.

Cox, M. A.; Plackettr, L. 1980. Small samples in contingency tables, *Biometrika* 67: 1–13.

Čekanavičius, V.; Wang, Y. H. 2003. Compound Poisson approximations for sums of discrete nonlattice variables, *Advances in Applied Probability* 35: 228–250.

Čekanavičius, V. 1999. On compound Poisson approximations under moment restrictions, *Teor. Veroyatnost. i Primenen.* 44: 74–86.

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. `http://www.pabst-publishers.de/mpr/`

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

Edgington, E. S. 1995. *Randomization tests*. Marcel Dekker, New York.

Faddy, M. J. Jones, M. C. 1998. Semiparametric smoothing for discrete data, *Biometrika* 85: 131–138.

Fienberg, S. E. 2000. Contingency tables and log-linear models: Basic results and new developments., *Journal of the American Statistical Association* 95: 643–647.

Fienberg, S. E.; Holland, P. W. 1973. Simulteneous estimation of multinomial cell probabilities, *Journal of the American Statistical Association* 68: 683–691.

Filina, M. V.; Zubkov, A. M. 2008. Exact computation of Pearson statistics distribution and some experimental results, *Austrian Journal of Statistics* 37: 129–135.

Filina, M. V.; Zubkov, A. M. 2011. Tail Properties of Pearson Statistics Distributions, *Austrian Journal of Statistics* 40: 47–54.

Fukumizu, K.; Bach, F. R.; Jordan, M. I. 2004. Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces, *The Journal of Machine Learning Research archive* 5: 73–99.

van de Geer, S. 2003. Asymptotic theory for maximum likelihood in nonparametric

mixture models, *Computational Statistics and Data Analysis* 41: 453–464.

Good, P. 1993. *Permutation Tests*. Springer Verlag, New York.

Gnedin, A.; Hansen, B.; Pitman, J. 2007. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws, *Probability Surveys* 4: 146–171. `http://www.i-journals.org/ps/viewarticle.php?id=92`

Györfi, L., Vajda, I. 2002. Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. *Statistics and Probability Letters* 56: 57–67.

Haberman, S. J. 1974. *The Analysis of Frequency Data*. The university of Chicago Press.

Hall, P.; Titterington, D. M. 1987. The effect of simulation order on level accuracy and power of Monte Carlo tests, *Journal of the Royal Statistical Society* 51: 459–467.

Hu, M. Y. 1999. *Model Checking for incomplete high dimensional categorical data*. University of California. Los Angeles. 101 p.

Ivchenko, G. I.; Medvedev, Yu. I. 1987. Separable statistics and checking hypotheses for grouped data, *Teor. Veroyatn. Primen.* 3: 549–560.

Kim, S. H.; Choi, H.; Lee, S. 2007. Estimate-based goodness-of-fit test for large sparse multinomial distributions, *Applied Mathematics Research report 07–08* Department of Mathematical Sciences, KAIST, Daejeon, S. Korea.

Khmaladze, E. V. 1988. The statistical analysis of a large number of rare events, in *Technical Report, MS-R8804, Amsterdam*.

Klaassen, C.; Mnatsakanov, R. 2000. Consistent estimation of the structural distribution function, *Scandinavian Journal of Statistics* 27: 733–746.

Kolchin, V. F.; Sevastyanov, B.; Chistyakov, V. 1978. *Random allocations*. Washington: Wiley.

Kreiner, S. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies, *Scandinavian Journal of Statistics* 14: 97–112.

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

Kvizhinadze, G. 2010. *Large number of rare events: Diversity analysis in multiple choice questionnaires and related topics* Doctoral Dissertation. Victoria University of Wellington.

Liese, F.; Vajda, I. 2006. On divergences and informations in statistics and information theory, *Transactions on Information Theory* 52: 4394–4412.

Manly, B. 1991. *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.

Medvedev, Yu. I. 1977. Divisible statistics in a polynomial scheme I, II, *Theory of Probability and Its Applications* 22: 607–615.

Muller, U. U.; Osius, G. 2003. Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, *Statistics* 37.2: 119–143.

Patefield, W. M. 1981. An efficient method of generating random R $\times$ C tables with given row and column totals, *Journal of the Royal Statistical Society* 30: 91–97.

Patefield, W. M. 1982. Exact tests for trends in ordered contingency tables, *Journal of the Royal Statistical Society* 31: 32–43.

Radavičius, M.; Židanavičiūtė, J. 2009. Semiparametric smoothing of sparse contingency tables, *Journal of Statistical Planning and Inference* 139(11): 3900–3907. `http://www.sciencedirect.com/science/article/pii/S0378375809001566`

Rao, C. R. 1965. *Linear Statistical Inference and its Applications*. John Wiley, New York.

Read, T.; Cressie, N. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York.

Sanov, I. N. 1957. On the probability of large deviations of random magnitudes, *Mat. Sb. N. S.* 42.84: 11–44.

Senchaudhuri, P.; Mehta, C. R.; Patel, N. R. 1995. Estimating exact p-values by the method of control variates, or Monte Carlo rescue, *Journal of American Statistical Association.* 90.430: 640–648.

Simonoff, J. S. 1995. Smoothing categorical data, *Journal of Statistical Planning and Inference* 47: 41–69.

Sprent, P. 1993. *Applied Nonparametric Statistical Methods*. Chapman and Hall, London.

StatXact 2011. *StatXact 5, Statistical Software for Exact Nonparametric Inference, User Manual*, Cytel Software, Cambridge, Mass. `http://www.cytel.com`.

Steck, G. P. 1957. Limit theorems for conditional distributions, *University of California Publications in Statistics* 12: 237–284.

Sun, Wei Guo, Elizabeth, A. T. 1989. Analysis of sparse contingency tables: Monte Carlo estimation of exact *p*-value, *Technical report No. 187* University of Washington, Seattle.

Tumanyan, S. Kh. 1954. On the asymptotic distribution of the χ^2 criterion, *Dokl. Akad. Nauk SSSR* 6: 1011–1012 (in Russian).

Tumanyan, S. Kh. 1956. The asymptotic distribution of the χ^2 criterion under simultaneous increase of the volume of the observations and the number of groups,

Theory of Probability and Its Applications 1: 131–145 (in Russian).

Zaitsev, A. Yu. 2005. On approximation of the sample by a Poisson point process, *Journal of Mathematical Sciences* 128: 2556–2563.

Zelterman, D. 1987. Goodness-of-fit tests for large sparse multinomial distributions, *Journal of American Statistical Association* 82: 624–629.

List of author's publications on the topic of dissertation

In the reviewed scientific journals

Radavičius, M.; Samusenko, P. 2012. Nonparametric testing for sparse nominal data, *Nonlinear Analysis: Modelling and Control*. ISSN 1392-5113. (Thomson ISI Web of Science). (Accepted for publication).

Radavičius, M.; Samusenko, P. 2011. Profile statistics for sparse contingency tables under poisson sampling, *Austrian Journal of Statistics* 40: 115–123. ISSN 1026-597X.

Samusenko, P. 2011. Inconsistency of chi-square test for sparse categorical data under multinomial sampling, *Lietuvos Matematikos Rinkinys: LMD darbai* 52: 327–331. ISSN 0132-2818.

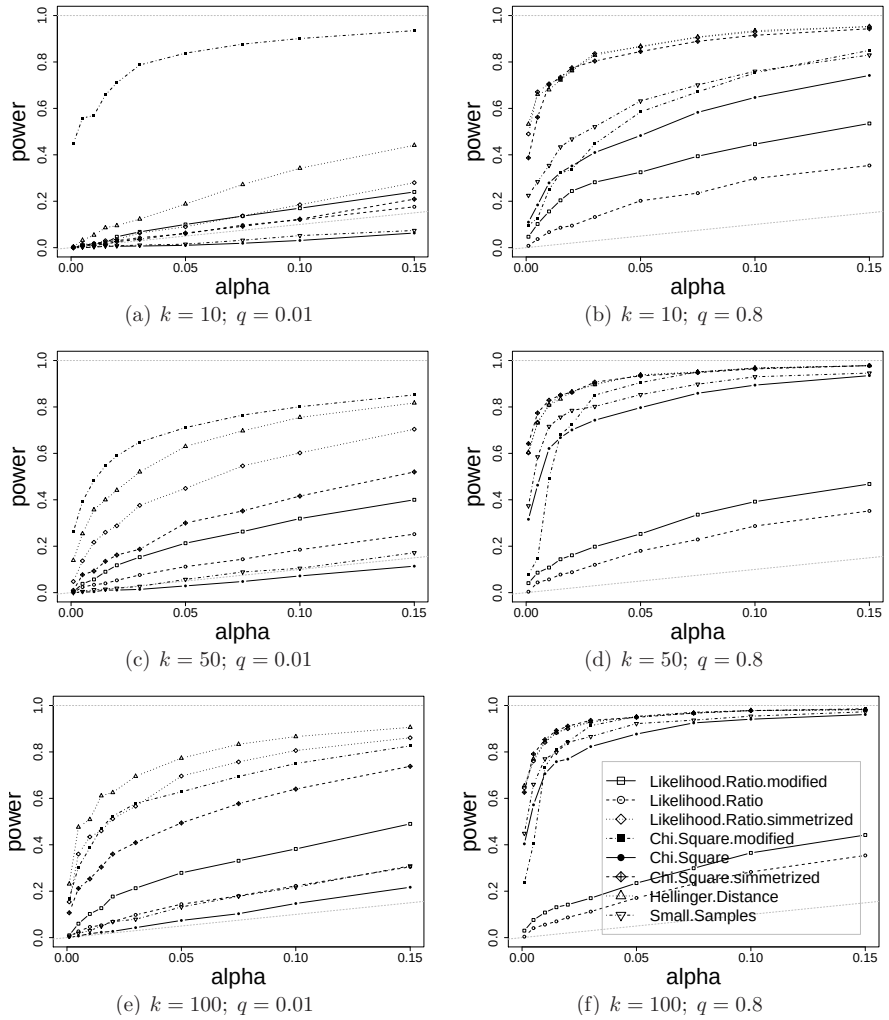
Radavičius, M.; Samusenko, P. 2010. Profile statistics for sparse contingency tables, *Computer Data Analysis and Modeling. Complex Stochastic Data and Systems, held in Minsk, Belarus* Minsk. BSU. 9(2): 55–58. ISBN 978-985-476-848-9.

Appendices

Appendix A. Two step models

Table A.1. Goodness-of-fit tests power of Likelihood ratio symmetrized statistic for 2S03 model

α	Likelihood ratio symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.113	**0.982**	**0.976**	**0.986**	0.795	0.942	0.955	0.535
0.1	0.081	**0.978**	**0.958**	**0.979**	0.703	0.909	0.939	0.415
0.075	0.047	**0.968**	**0.945**	**0.970**	0.639	0.875	0.921	0.362
0.05	0.036	**0.951**	**0.918**	**0.956**	0.546	0.821	0.896	0.291
0.03	0.021	**0.928**	**0.884**	**0.939**	0.452	0.756	0.861	0.196
0.02	0.010	**0.900**	**0.844**	**0.926**	0.391	0.684	0.813	0.139
0.015	0.008	**0.883**	**0.836**	**0.900**	0.370	0.655	0.782	0.115
0.01	0.008	**0.841**	**0.788**	**0.871**	0.317	0.587	0.702	0.072
0.005	0.007	**0.761**	**0.695**	**0.845**	0.166	0.504	0.638	0.045
0.001	0.000	**0.645**	0.451	**0.770**	0.130	0.329	**0.528**	0.010

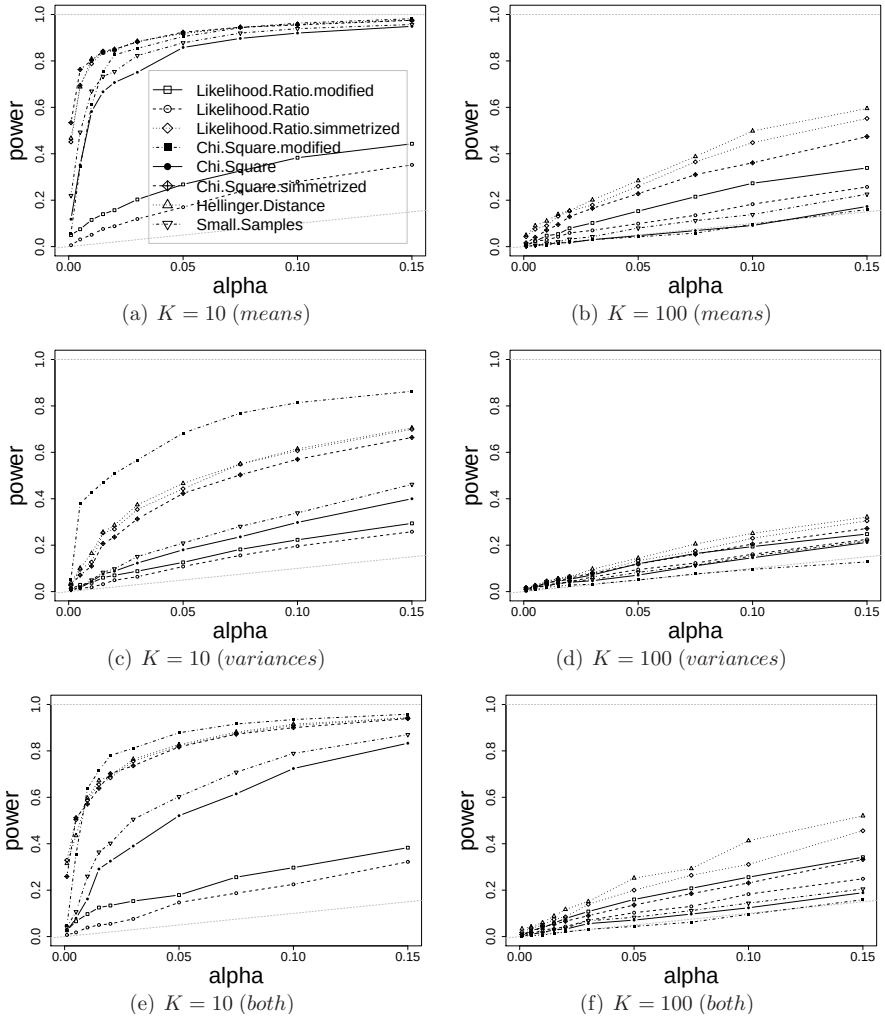


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. A.1. Goodness-of-fit tests power of MCMC smoothing dependance on its parameters k and q in 2S03 model

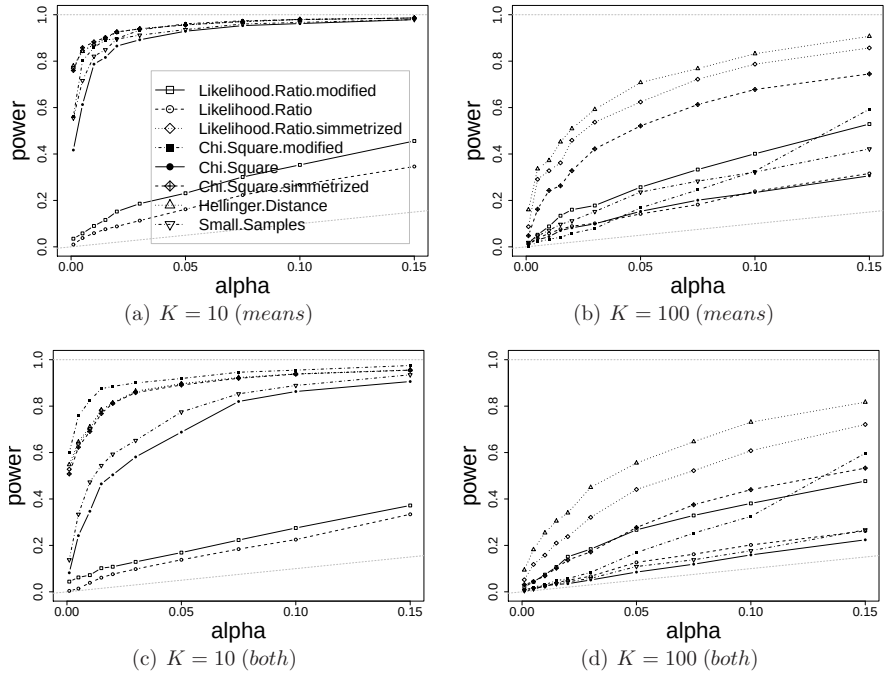


(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. A.2. Goodness-of-fit tests power of GQ for $K \in \{10, 100\}$ in 2S03 model



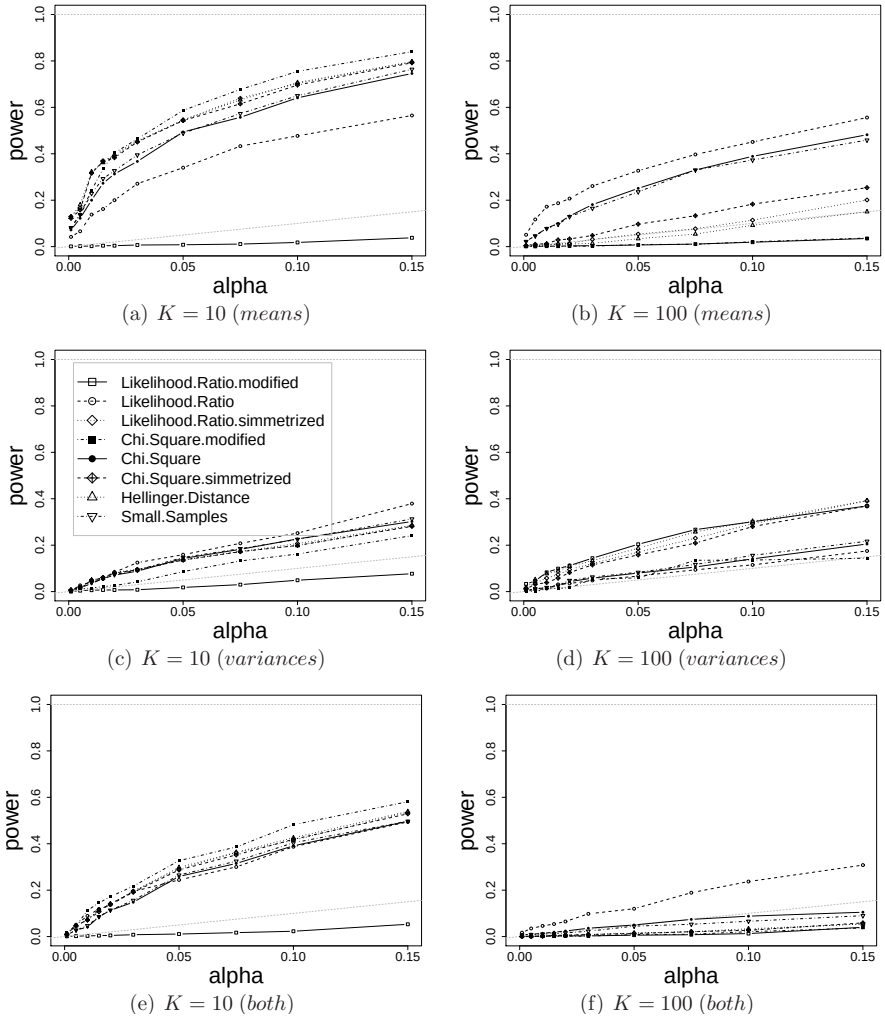
(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(both)$ (d) $K = 100$ $(both)$

Fig. A.3. Goodness-of-fit tests power of GG for $K \in \{10, 100\}$ in 2S03 model

Table A.2. Goodness-of-fit tests power of Chi-Square symmetrized statistic for 2S03 model

α	Chi-Square symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.114	**0.983**	**0.975**	**0.986**	0.792	0.939	0.955	0.529
0.1	0.077	**0.978**	**0.955**	**0.980**	0.696	0.900	0.938	0.419
0.075	0.059	**0.967**	**0.945**	**0.971**	0.615	0.872	0.920	0.352
0.05	0.035	**0.949**	**0.924**	**0.956**	0.543	0.817	0.891	0.287
0.03	0.023	**0.936**	**0.882**	**0.939**	0.451	0.736	0.858	0.191
0.02	0.012	**0.911**	**0.850**	**0.926**	0.383	0.702	0.812	0.141
0.015	0.010	**0.892**	**0.842**	**0.902**	0.366	0.639	0.768	0.107
0.01	0.007	**0.853**	**0.802**	**0.883**	0.317	0.571	0.691	0.071
0.005	0.005	**0.791**	**0.763**	**0.858**	0.158	0.512	0.623	0.048
0.001	0.000	**0.626**	**0.534**	**0.760**	0.122	0.259	0.508	0.009



(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

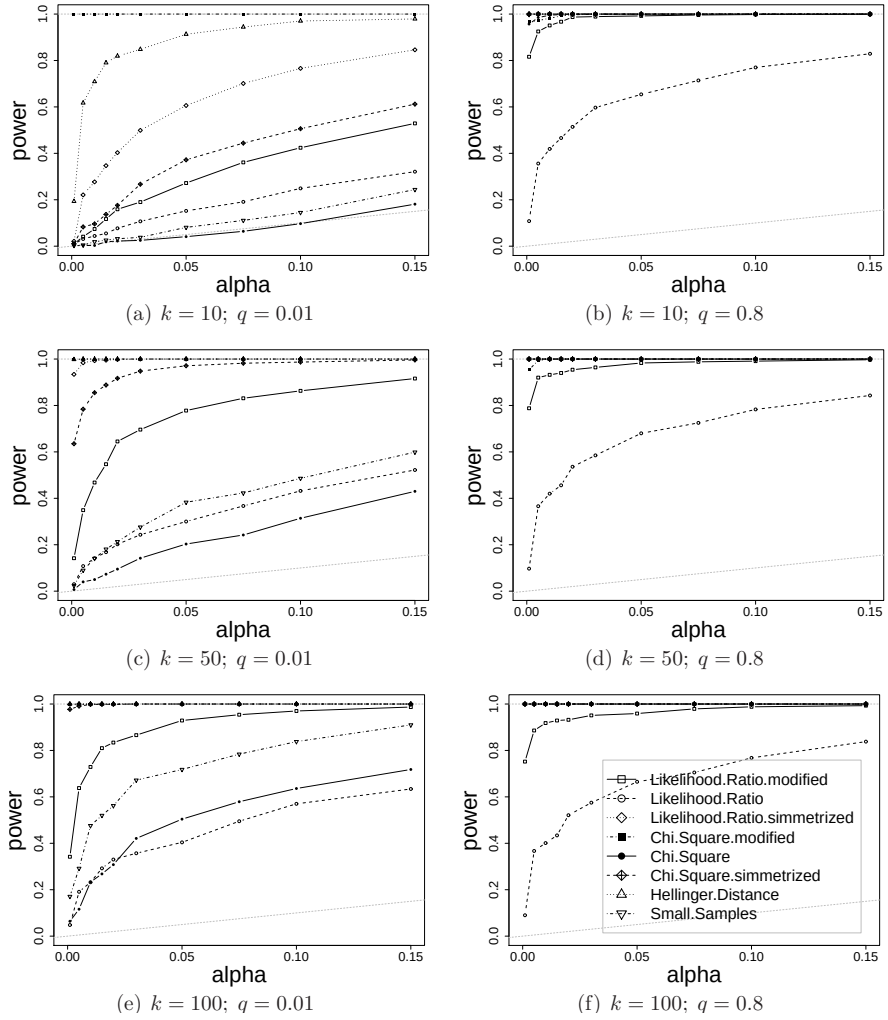
Fig. A.4. Goodness-of-fit tests power of GC for $K \in \{10, 100\}$ in 2S03 model

Table A.3. Goodness-of-fit tests power of Hellinger distance statistic for 2S03 model

α	Hellinger distance statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.140	**0.982**	**0.976**	**0.986**	0.797	0.943	0.956	0.539
0.1	0.084	**0.979**	**0.958**	**0.980**	0.707	0.915	0.939	0.426
0.075	0.057	**0.967**	**0.945**	**0.971**	0.632	0.881	0.924	0.363
0.05	0.036	**0.951**	**0.919**	**0.958**	0.544	0.827	0.897	0.298
0.03	0.019	**0.931**	**0.885**	**0.940**	0.456	0.764	0.865	0.194
0.02	0.014	**0.903**	**0.850**	**0.924**	0.397	0.693	0.814	0.141
0.015	0.010	**0.883**	**0.835**	**0.898**	0.362	0.672	0.784	0.113
0.01	0.008	**0.844**	**0.809**	**0.871**	0.323	0.596	0.711	0.074
0.005	0.005	**0.769**	**0.688**	**0.842**	0.180	0.435	0.646	0.047
0.001	0.000	**0.654**	0.467	**0.780**	0.127	0.315	**0.548**	0.010

Table A.4. Goodness-of-fit tests power of Chi-Square modified statistic for 2S03 model

α	Chi-Square modified statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.037	**0.986**	**0.982**	**0.986**	0.840	0.958	0.975	0.581
0.1	0.025	**0.978**	**0.963**	**0.979**	0.755	0.935	0.955	0.482
0.075	0.015	**0.971**	0.943	**0.974**	0.677	0.916	**0.946**	0.386
0.05	0.006	**0.953**	0.905	**0.961**	0.587	0.878	**0.919**	0.327
0.03	0.002	**0.913**	0.854	**0.934**	0.464	0.812	**0.900**	0.217
0.02	0.002	**0.844**	0.826	**0.897**	0.404	0.780	**0.885**	0.172
0.015	0.002	**0.809**	0.756	**0.891**	0.335	0.715	**0.875**	0.145
0.01	0.002	**0.735**	0.611	**0.859**	0.243	0.638	**0.824**	0.113
0.005	0.002	**0.404**	0.344	**0.801**	0.138	0.355	**0.758**	0.046
0.001	0.001	**0.236**	0.058	**0.562**	0.083	0.046	**0.600**	0.017

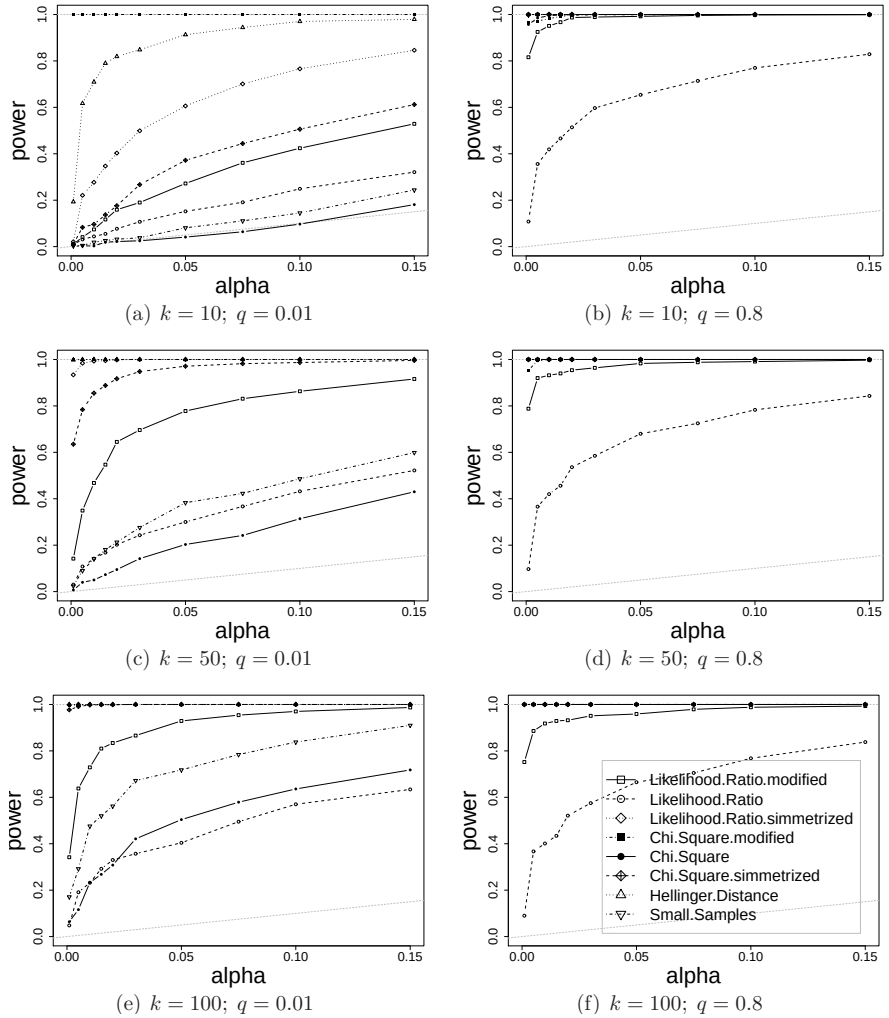


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. A.5. Goodness-of-fit tests power of MCMC dependance on its parameters k and q in 2S05 model

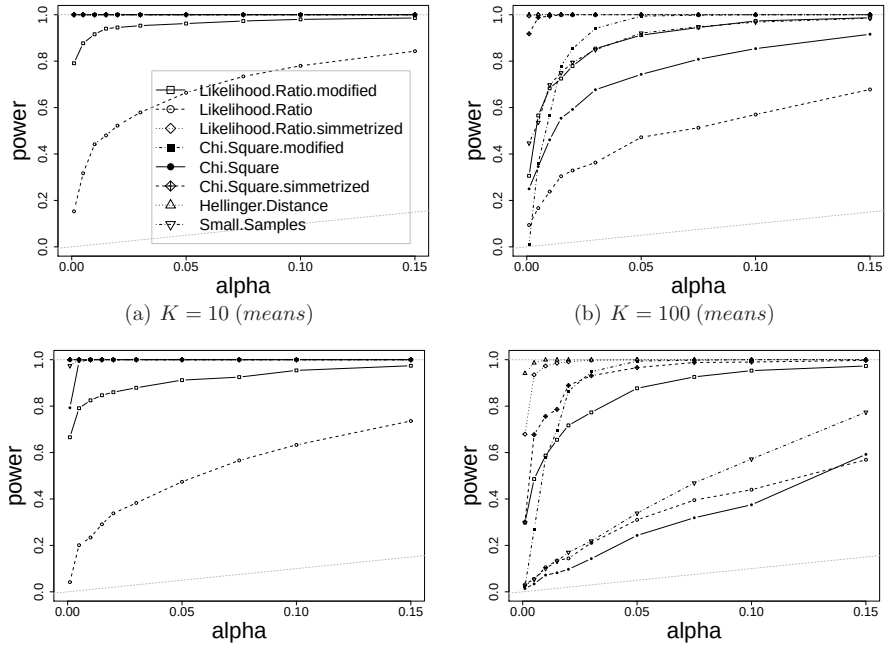


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. A.6. Goodness-of-fit tests power of GQ for $K \in \{10, 100\}$ in 2S05 model



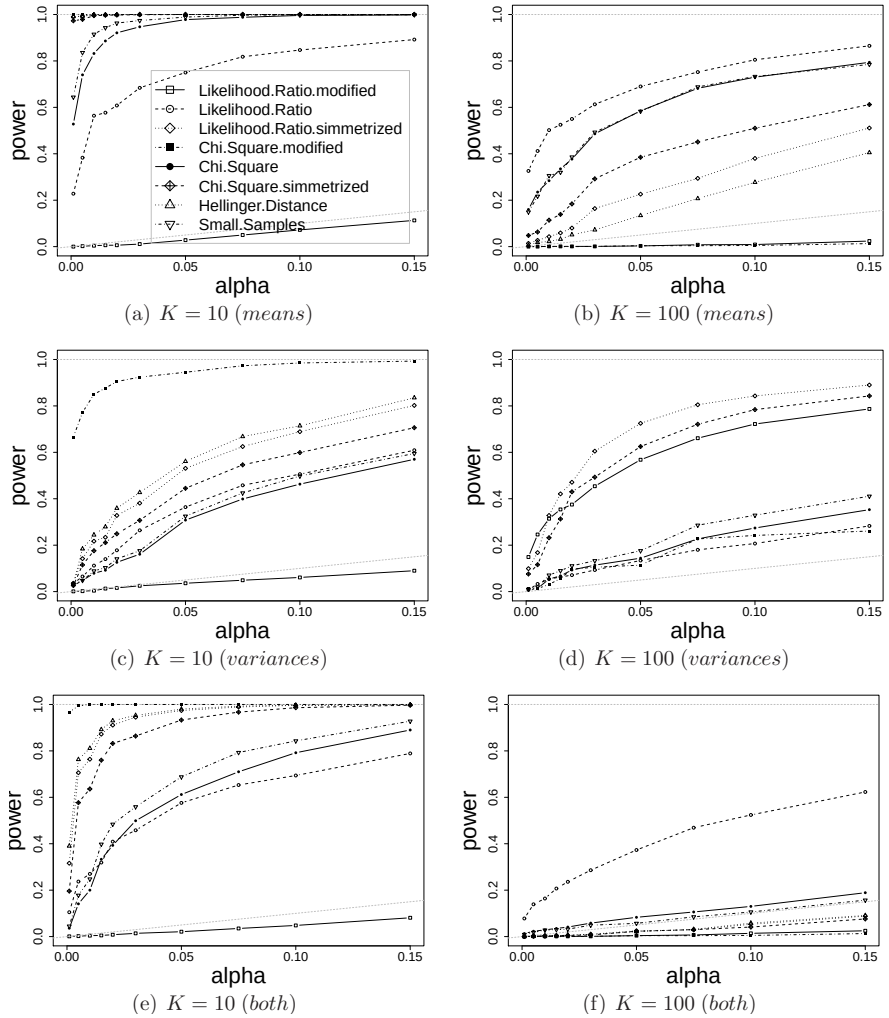
(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(both)$ (d) $K = 100$ $(both)$

Fig. A.7. Goodness-of-fit tests power of GG for $K \in \{10, 100\}$ in 2S05 model

Table A.5. Goodness-of-fit tests power of Likelihood ratio symmetrized statistic for 2S05 model

α	Likelihood ratio symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.317	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.998
0.1	0.228	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.996
0.075	0.158	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.989
0.05	0.105	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.974
0.03	0.057	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.945
0.02	0.034	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.911
0.015	0.026	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.872
0.01	0.024	**1.000**	**1.000**	**1.000**	0.998	**1.000**	**1.000**	0.764
0.005	0.015	**1.000**	**1.000**	**1.000**	0.993	**1.000**	**1.000**	0.706
0.001	0.000	**1.000**	**1.000**	**1.000**	0.990	**1.000**	**1.000**	0.316



(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. A.8. Goodness-of-fit tests power of GC for $K \in \{10, 100\}$ in 2S05 model

Table A.6. Goodness-of-fit tests power of Chi-Square symmetrized statistic for 2S05 model

α	Chi-Square symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.340	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.997
0.1	0.229	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.986
0.075	0.169	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.967
0.05	0.132	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.933
0.03	0.077	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.864
0.02	0.058	**1.000**	**1.000**	**1.000**	0.998	**1.000**	**1.000**	0.832
0.015	0.029	**1.000**	**1.000**	**1.000**	0.997	**1.000**	**1.000**	0.760
0.01	0.023	**1.000**	**1.000**	**1.000**	0.995	**1.000**	**1.000**	0.636
0.005	0.017	**1.000**	**1.000**	**1.000**	0.979	**1.000**	**1.000**	0.577
0.001	0.002	**1.000**	**1.000**	**1.000**	0.973	**1.000**	**1.000**	0.196

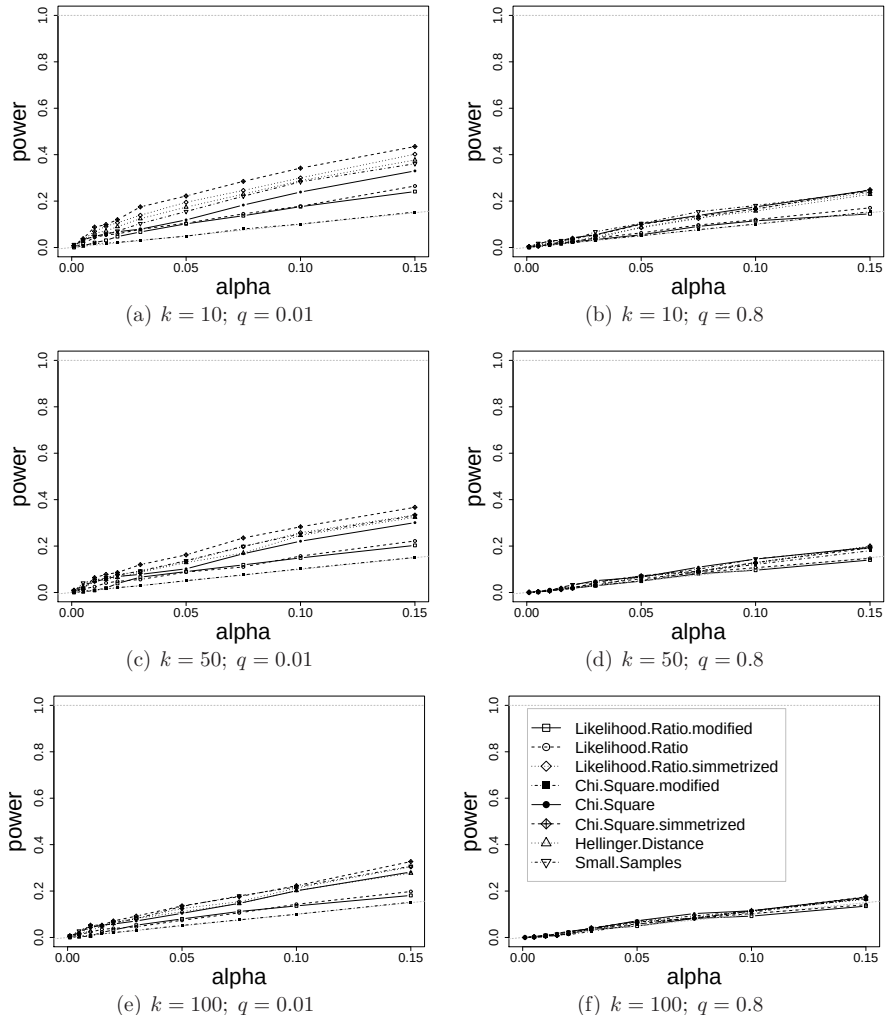
Table A.7. Goodness-of-fit tests power of Hellinger distance statistic for 2S05 model

α	Hellinger distance statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.350	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.999
0.1	0.238	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.997
0.075	0.180	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.993
0.05	0.107	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.980
0.03	0.051	**1.000**	**1.000**	**1.000**	1.000	**1.000**	**1.000**	0.953
0.02	0.033	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.930
0.015	0.029	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.892
0.01	0.023	**1.000**	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.811
0.005	0.014	**1.000**	**1.000**	**1.000**	0.993	**1.000**	**1.000**	0.763
0.001	0.000	**1.000**	**1.000**	**1.000**	0.991	**1.000**	**1.000**	0.390

Table A.8. Goodness-of-fit tests power of Chi-Square modified statistic for 2S05 model

α	Chi-Square modified statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.016	**1.000**	1.000	**1.000**	1.000	1.000	**1.000**	1.000
0.1	0.007	**1.000**	1.000	**1.000**	1.000	1.000	**1.000**	1.000
0.075	0.004	**1.000**	1.000	**1.000**	1.000	1.000	**1.000**	1.000
0.05	0.003	**1.000**	1.000	**1.000**	1.000	1.000	**1.000**	1.000
0.03	0.001	**1.000**	1.000	**1.000**	0.999	1.000	**1.000**	1.000
0.02	0.000	**1.000**	1.000	**1.000**	0.998	1.000	**1.000**	1.000
0.015	0.000	**1.000**	1.000	**1.000**	0.997	1.000	**1.000**	1.000
0.01	0.000	**1.000**	1.000	**1.000**	0.995	1.000	**1.000**	1.000
0.005	0.000	**1.000**	1.000	**1.000**	0.979	1.000	**1.000**	0.994
0.001	0.000	**1.000**	0.812	**1.000**	0.973	0.811	**1.000**	0.967

Appendix B. Split models

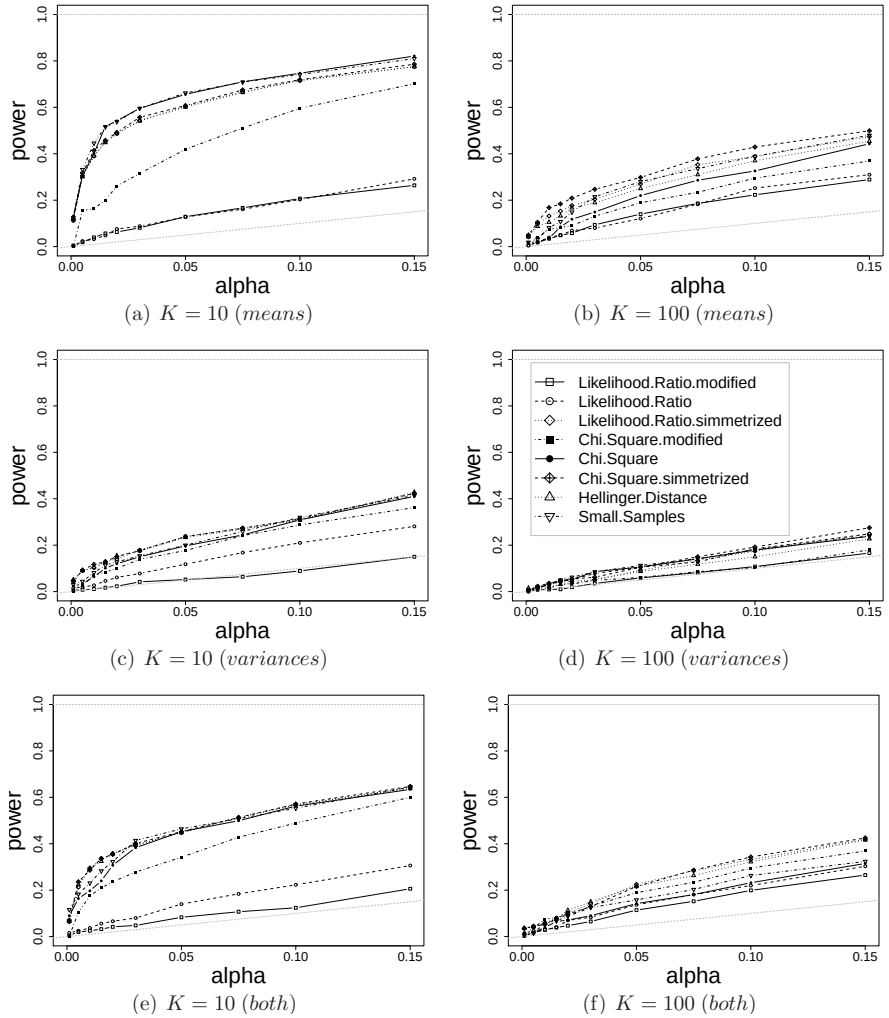


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. B.1. Goodness-of-fit tests power of MCMC smoothing compared for different k and q parameters for TS06 model

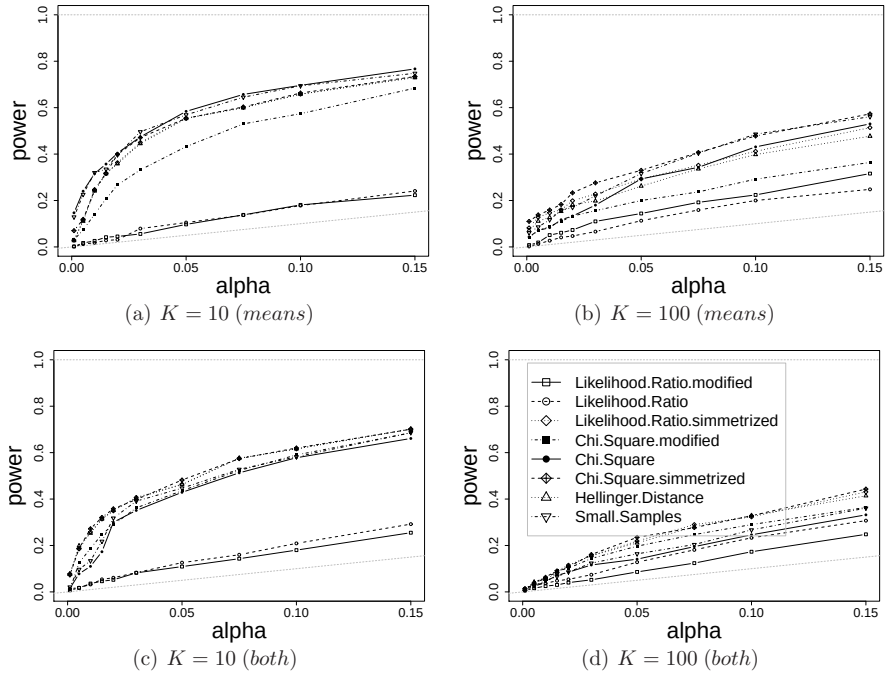


(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. B.2. Goodness-of-fit tests power of GQ for $K \in \{10, 100\}$ in TS06 model



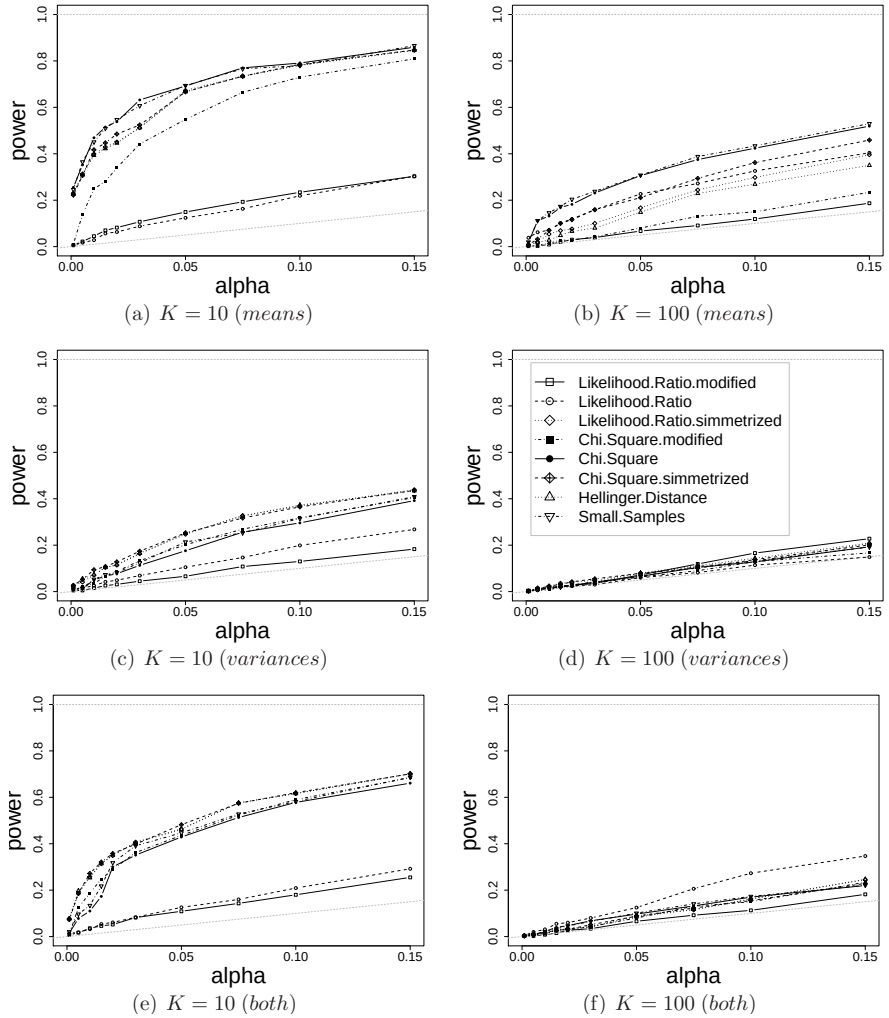
(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(both)$ (d) $K = 100$ $(both)$

Fig. B.3. Goodness-of-fit tests power of GG for $K \in \{10, 100\}$ in TS06 model

Table B.1. Goodness-of-fit tests power of Chi-Square statistic for TS06 model

α	Chi-Square statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.155	0.231	**0.622**	**0.769**	0.316	0.468	**0.585**	0.209
0.1	0.106	0.180	**0.546**	**0.689**	0.221	0.375	**0.501**	0.146
0.075	0.077	0.134	**0.496**	**0.648**	0.186	0.337	**0.445**	0.123
0.05	0.054	0.094	**0.418**	**0.550**	0.147	0.280	**0.331**	0.085
0.03	0.035	0.058	**0.360**	**0.458**	0.100	0.216	**0.260**	0.052
0.02	0.019	0.052	**0.299**	**0.379**	0.071	0.150	**0.217**	0.030
0.015	0.013	0.043	**0.260**	**0.303**	0.065	0.114	**0.203**	0.025
0.01	0.010	0.029	**0.226**	**0.245**	0.046	0.093	**0.143**	0.020
0.005	0.006	0.011	**0.172**	**0.193**	0.039	0.069	**0.106**	0.008
0.001	0.001	0.007	**0.060**	**0.122**	0.012	**0.040**	0.031	0.004



(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. B.4. Goodness-of-fit tests power of GC for $K \in \{10, 100\}$ in TS06 model

Table B.2. Goodness-of-fit tests power of Likelihood ratio symmetrized statistic for TS06 model

α	Likelihood ratio symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.429	0.402	**0.775**	**0.733**	**0.847**	0.642	0.684	0.702
0.1	0.344	0.300	**0.715**	**0.658**	**0.781**	0.566	0.593	0.616
0.075	0.279	0.246	**0.667**	**0.601**	**0.735**	0.513	0.530	0.576
0.05	0.202	0.195	**0.605**	**0.555**	**0.673**	0.447	0.476	0.464
0.03	0.151	0.139	**0.544**	**0.450**	**0.512**	0.395	0.434	0.407
0.02	0.118	0.107	**0.487**	0.363	**0.451**	0.358	**0.372**	0.349
0.015	0.104	0.091	**0.453**	0.316	**0.429**	**0.335**	0.299	0.314
0.01	0.073	0.070	**0.390**	0.247	**0.399**	0.295	**0.270**	0.258
0.005	0.057	0.033	**0.305**	0.115	**0.313**	0.213	**0.216**	0.185
0.001	0.008	0.007	**0.120**	0.029	**0.230**	0.064	0.073	**0.078**

Table B.3. Goodness-of-fit tests power of Chi-Square symmetrized statistic for TS06 model

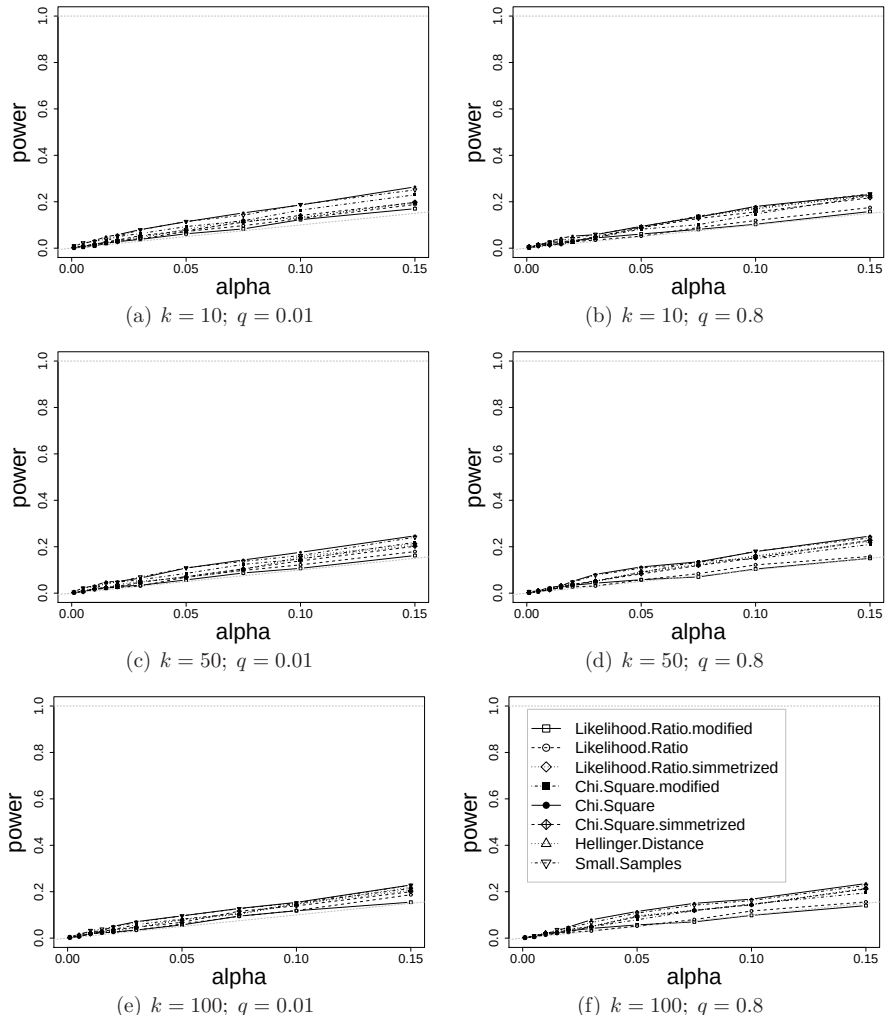
α	Chi-Square symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.436	0.435	**0.786**	**0.734**	**0.847**	0.647	0.684	0.701
0.1	0.371	0.342	**0.719**	**0.663**	**0.786**	0.571	0.595	0.620
0.075	0.311	0.285	**0.675**	**0.603**	**0.734**	0.513	0.529	0.574
0.05	0.226	0.222	**0.608**	**0.554**	**0.667**	0.452	0.476	0.482
0.03	0.171	0.175	**0.557**	**0.471**	**0.524**	0.401	0.431	0.398
0.02	0.136	0.120	**0.492**	**0.398**	**0.485**	0.354	0.369	0.358
0.015	0.120	0.099	**0.458**	0.318	**0.447**	**0.337**	0.324	0.321
0.01	0.078	0.088	**0.414**	0.245	**0.417**	**0.286**	0.280	0.272
0.005	0.065	0.038	**0.317**	0.119	**0.306**	**0.236**	0.224	0.187
0.001	0.011	0.009	**0.112**	0.070	**0.222**	0.071	**0.077**	0.073

Table B.4. Goodness-of-fit tests power of Hellinger distance statistic for TS06 model

α	Hellinger distance statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.416	0.377	**0.775**	**0.729**	**0.845**	0.638	0.682	0.701
0.1	0.308	0.288	**0.717**	**0.658**	**0.783**	0.569	0.587	0.616
0.075	0.260	0.230	**0.664**	**0.599**	**0.733**	0.513	0.531	0.576
0.05	0.204	0.174	**0.601**	**0.553**	**0.667**	0.448	0.478	0.462
0.03	0.141	0.124	**0.541**	**0.445**	**0.512**	0.393	0.433	0.404
0.02	0.110	0.090	**0.487**	0.358	**0.446**	0.356	**0.377**	0.351
0.015	0.081	0.071	**0.448**	0.315	**0.422**	**0.324**	0.294	0.314
0.01	0.072	0.062	**0.391**	0.241	**0.393**	**0.294**	0.266	0.253
0.005	0.040	0.027	**0.301**	0.111	**0.310**	**0.224**	0.212	0.198
0.001	0.007	0.007	**0.119**	0.027	**0.228**	0.065	0.072	**0.077**

Table B.5. Goodness-of-fit tests power of Chi-Square modified statistic for TS06 model

α	Chi-Square modified statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.337	0.153	**0.702**	0.684	**0.809**	0.600	0.657	**0.685**
0.1	0.246	0.101	**0.596**	0.574	**0.730**	0.489	0.567	**0.590**
0.075	0.198	0.080	0.511	**0.530**	**0.665**	0.428	0.494	**0.523**
0.05	0.145	0.049	0.419	**0.432**	**0.548**	0.342	0.427	**0.436**
0.03	0.088	0.031	0.315	0.332	**0.442**	0.278	**0.369**	**0.362**
0.02	0.067	0.022	0.261	0.267	**0.339**	0.236	**0.282**	**0.289**
0.015	0.058	0.019	0.200	0.207	**0.281**	0.210	**0.246**	**0.248**
0.01	0.036	0.017	0.164	0.141	**0.249**	0.175	**0.191**	**0.187**
0.005	0.023	0.008	**0.155**	0.073	**0.138**	0.104	**0.128**	0.125
0.001	0.009	0.003	0.001	**0.031**	**0.011**	0.001	**0.037**	0.010

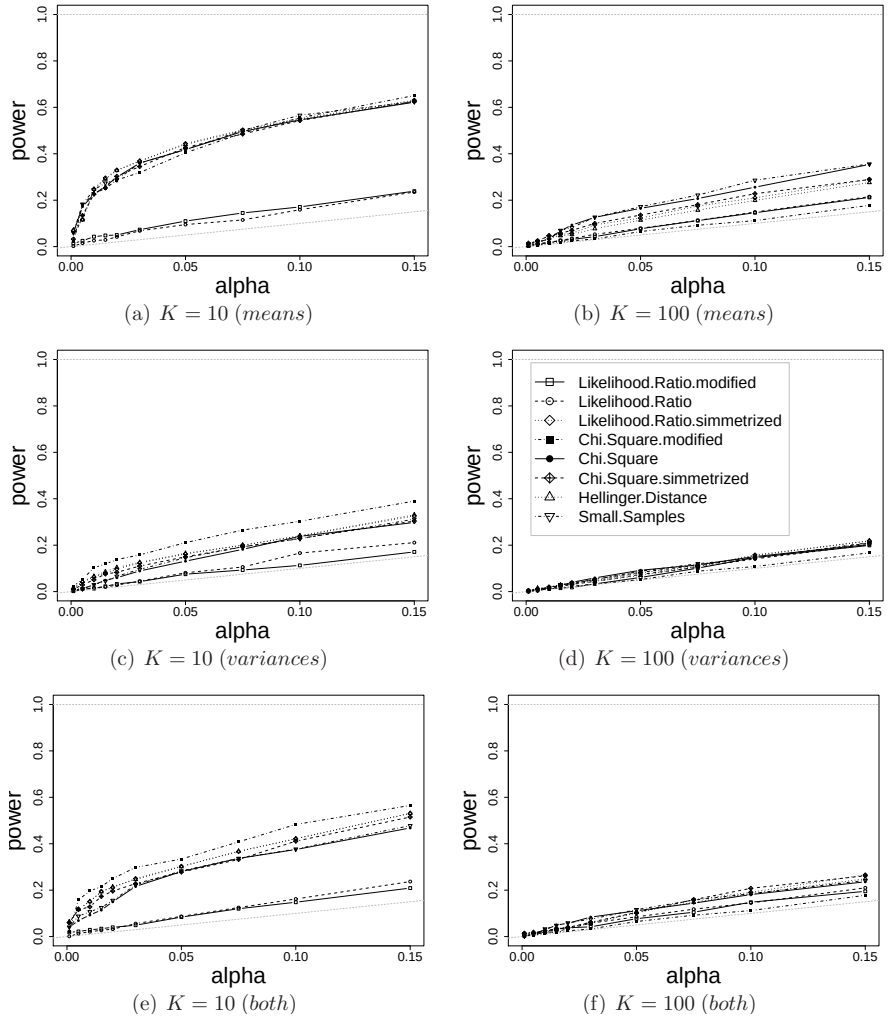


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. B.5. Goodness-of-fit tests power of MCMC smoothing compared for different k and q parameters for BS03 model

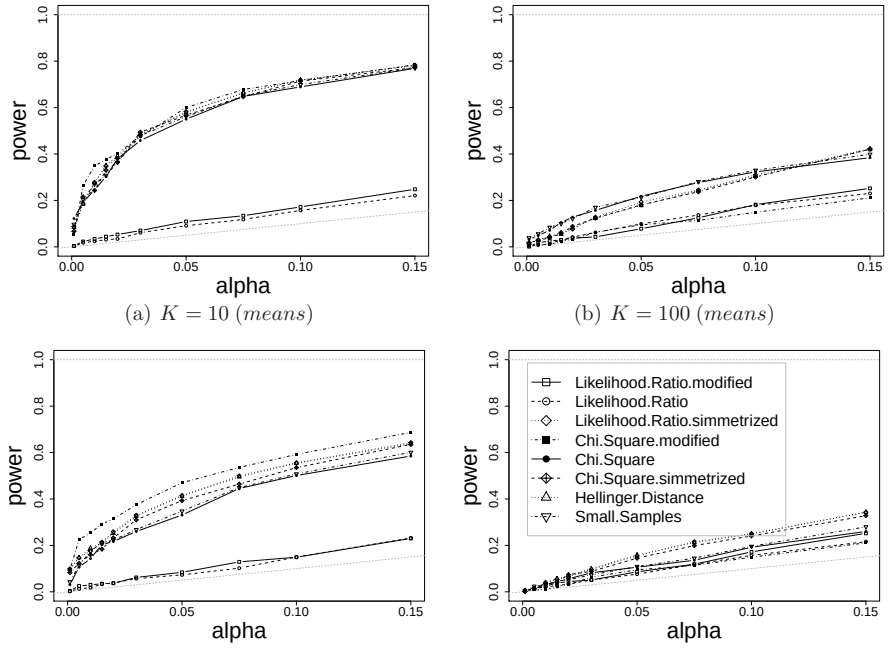


(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. B.6. Goodness-of-fit tests power of GQ for $K \in \{10, 100\}$ in BS03 model



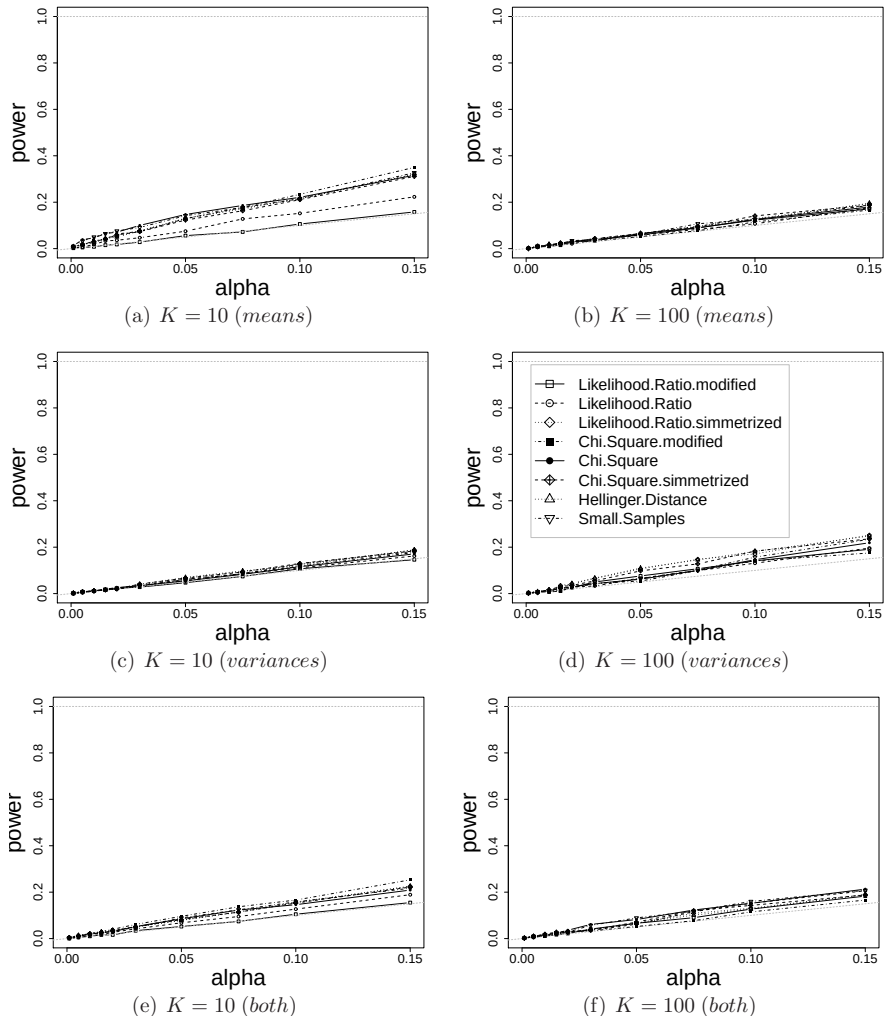
(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(both)$ (d) $K = 100$ $(both)$

Fig. B.7. Goodness-of-fit tests power of GG for $K \in \{10, 100\}$ in BS03 model

Table B.6. Goodness-of-fit tests power of Likelihood ratio symmetrized statistic for BS03 model

α	Likelihood ratio symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.205	0.215	**0.631**	**0.783**	0.320	0.531	**0.643**	0.225
0.1	0.145	0.144	**0.551**	**0.718**	0.216	0.422	**0.556**	0.154
0.075	0.110	0.121	**0.503**	**0.663**	0.175	0.366	**0.500**	0.113
0.05	0.078	0.097	**0.444**	**0.583**	0.131	0.303	**0.415**	0.079
0.03	0.042	0.050	**0.369**	**0.484**	0.077	0.250	**0.330**	0.051
0.02	0.030	0.039	**0.330**	**0.381**	0.054	0.211	**0.259**	0.036
0.015	0.023	0.025	**0.296**	**0.348**	0.043	0.195	**0.214**	0.027
0.01	0.018	0.022	**0.248**	**0.278**	0.026	0.151	**0.176**	0.021
0.005	0.010	0.007	**0.117**	**0.186**	0.018	0.119	**0.147**	0.013
0.001	0.004	0.002	**0.069**	**0.087**	0.009	0.060	**0.099**	0.004



(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. B.8. Goodness-of-fit tests power of GC for $K \in \{10, 100\}$ in BS03 model

Table B.7. Goodness-of-fit tests power of Chi-Square symmetrized statistic for BS03 model

α	Chi-Square symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.213	0.212	**0.626**	**0.772**	0.327	0.515	**0.635**	0.222
0.1	0.140	0.144	**0.543**	**0.714**	0.215	0.411	**0.535**	0.153
0.075	0.109	0.119	**0.485**	**0.647**	0.177	0.332	**0.464**	0.114
0.05	0.070	0.092	**0.427**	**0.567**	0.142	0.278	**0.393**	0.082
0.03	0.056	0.050	**0.345**	**0.494**	0.091	0.228	**0.310**	0.051
0.02	0.037	0.036	**0.303**	**0.364**	0.076	0.196	**0.232**	0.033
0.015	0.029	0.023	**0.253**	**0.330**	0.064	0.173	**0.184**	0.025
0.01	0.021	0.022	**0.226**	**0.243**	0.050	0.128	**0.160**	0.020
0.005	0.011	0.006	**0.134**	**0.215**	0.033	0.116	**0.123**	0.012
0.001	0.004	0.002	**0.032**	**0.067**	0.010	0.060	**0.083**	0.005

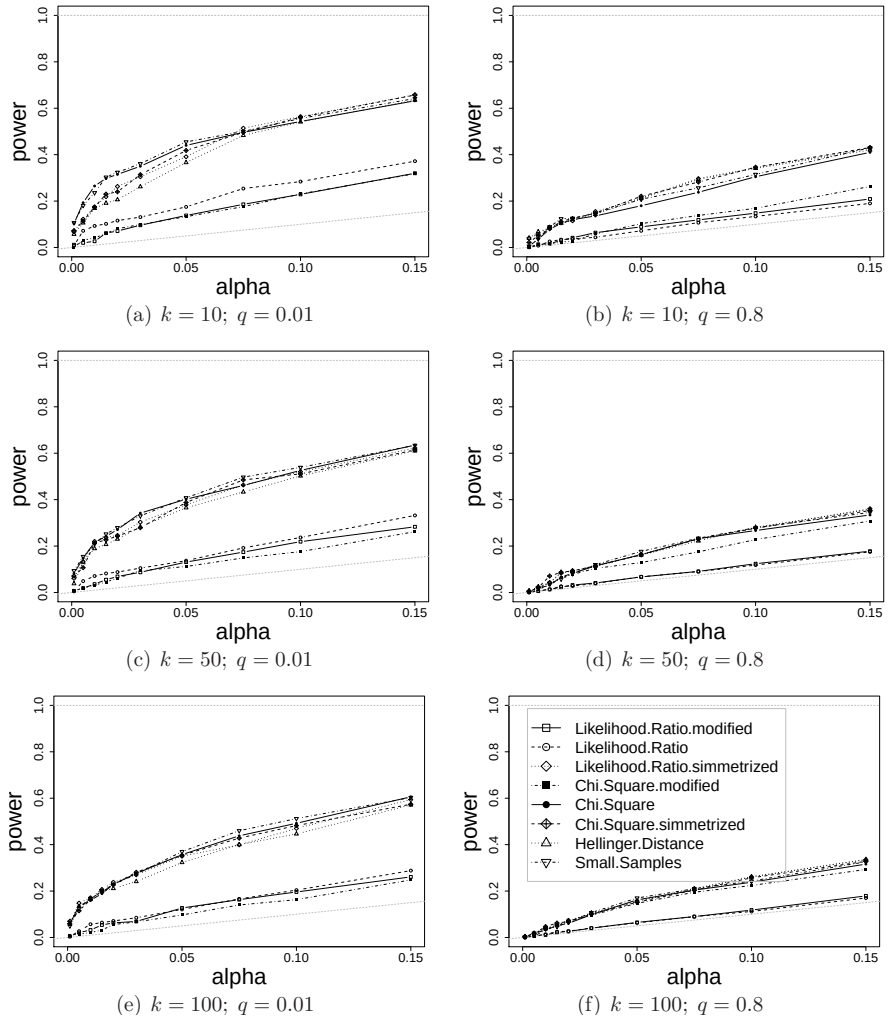
Table B.8. Goodness-of-fit tests power of Hellinger distance statistic for BS03 model

α	Hellinger distance statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.201	0.215	**0.627**	**0.780**	0.321	0.531	**0.643**	0.226
0.1	0.134	0.144	**0.547**	**0.719**	0.215	0.421	**0.556**	0.157
0.075	0.110	0.119	**0.499**	**0.663**	0.171	0.368	**0.500**	0.116
0.05	0.069	0.094	**0.440**	**0.579**	0.129	0.301	**0.415**	0.078
0.03	0.037	0.050	**0.368**	**0.482**	0.075	0.246	**0.330**	0.050
0.02	0.022	0.039	**0.328**	**0.391**	0.050	0.214	**0.259**	0.036
0.015	0.020	0.025	**0.291**	**0.350**	0.043	0.193	**0.214**	0.029
0.01	0.013	0.023	**0.247**	**0.275**	0.026	0.151	**0.176**	0.020
0.005	0.011	0.006	**0.115**	**0.184**	0.018	0.117	**0.147**	0.014
0.001	0.003	0.002	**0.071**	**0.084**	0.009	0.064	**0.099**	0.004

Table B.9. Goodness-of-fit tests power of Chi-Square modified statistic for BS03 model

α	Chi-Square modified statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.156	0.195	**0.651**	**0.784**	0.350	0.565	**0.687**	0.253
0.1	0.106	0.147	**0.555**	**0.716**	0.234	0.483	**0.592**	0.166
0.075	0.077	0.119	**0.489**	**0.678**	0.178	0.408	**0.536**	0.137
0.05	0.056	0.079	**0.404**	**0.600**	0.129	0.334	**0.470**	0.096
0.03	0.036	0.049	**0.319**	**0.474**	0.074	0.298	**0.375**	0.061
0.02	0.018	0.028	**0.287**	**0.403**	0.063	0.250	**0.314**	0.040
0.015	0.013	0.026	**0.250**	**0.376**	0.042	0.215	**0.292**	0.030
0.01	0.011	0.019	**0.229**	**0.349**	0.033	0.197	**0.254**	0.024
0.005	0.007	0.012	**0.180**	**0.265**	0.016	0.158	**0.225**	0.011
0.001	0.002	0.002	**0.023**	**0.054**	0.001	0.023	**0.095**	0.001

Appendix C. Irregular model

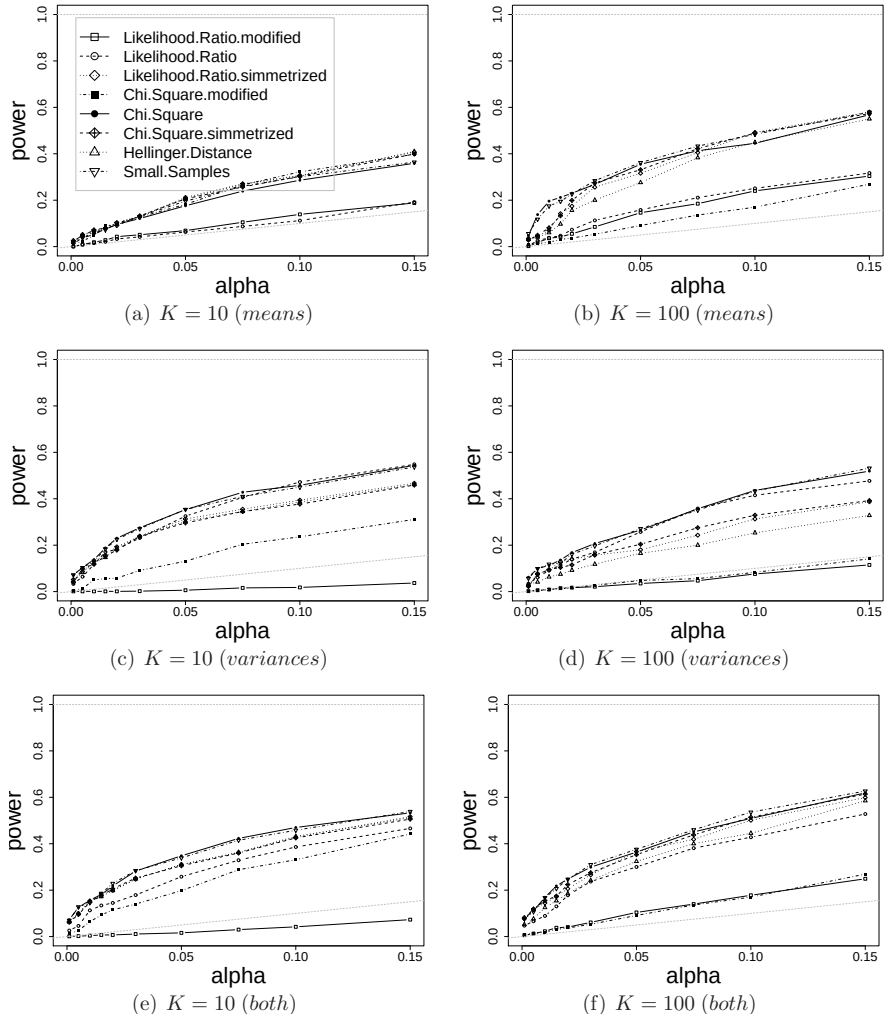


(a) $k = 10;\ q = 0.01$ (b) $k = 10;\ q = 0.8$

(c) $k = 50;\ q = 0.01$ (d) $k = 50;\ q = 0.8$

(e) $k = 100;\ q = 0.01$ (f) $k = 100;\ q = 0.8$

Fig. C.1. Goodness-of-fit tests power of MCMC smoothing compared for different k and q parameters in 2SV025 model

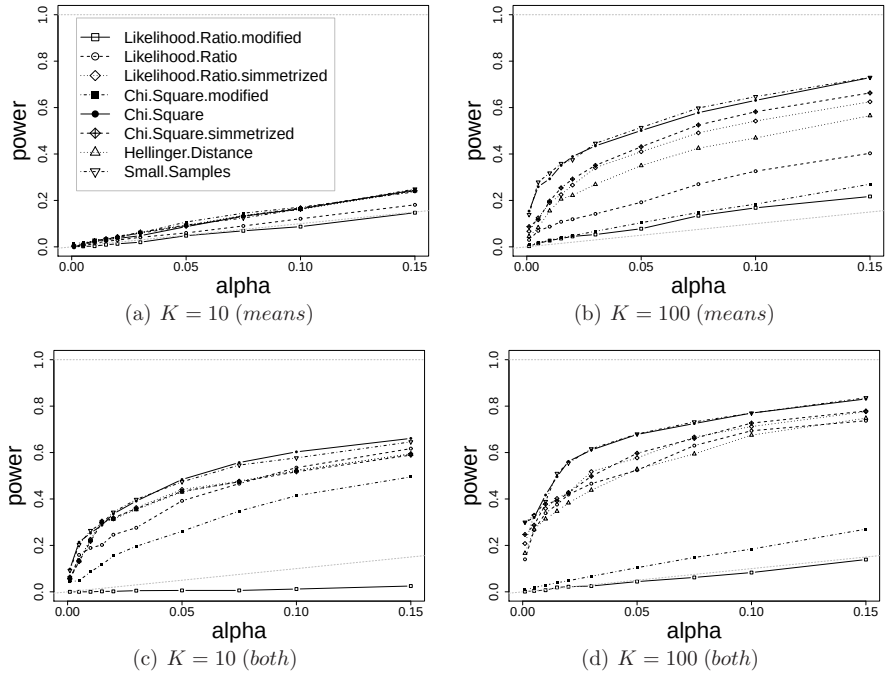


(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. C.2. Goodness-of-fit tests power of GQ for $K \in \{10, 100\}$ in 2SV025 model



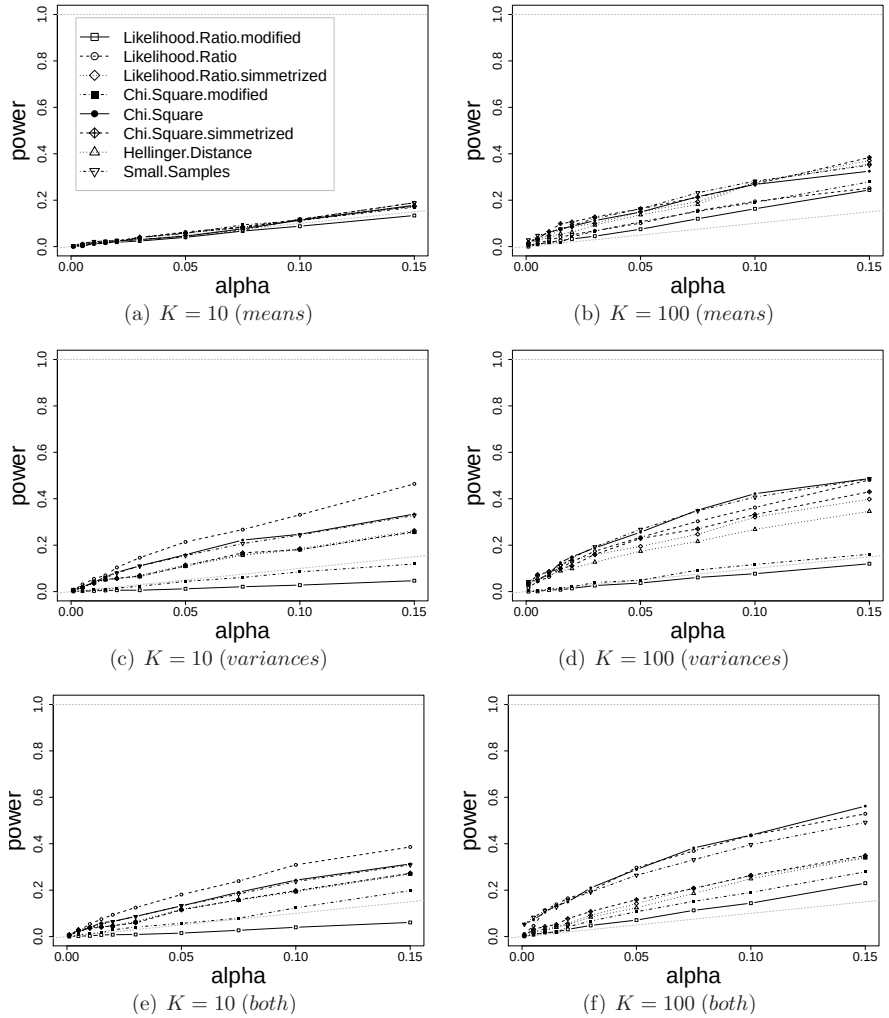
(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(both)$ (d) $K = 100$ $(both)$

Fig. C.3. Goodness-of-fit tests power of GG for $K \in \{10, 100\}$ in 2SV025 model

Table C.1. Goodness-of-fit tests power of Chi-Square statistic for 2SV025 model

α	Chi-Square statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	**0.634**	0.632	0.463	**0.720**	0.325	0.610	**0.830**	0.562
0.1	**0.555**	0.543	0.379	**0.631**	0.268	0.557	**0.767**	0.437
0.075	**0.510**	0.496	0.326	**0.571**	0.215	0.491	**0.727**	0.382
0.05	**0.444**	0.440	0.243	**0.507**	0.149	0.393	**0.676**	0.289
0.03	**0.351**	0.350	0.190	**0.438**	0.112	0.312	**0.608**	0.210
0.02	**0.325**	0.315	0.138	**0.387**	0.088	0.271	**0.562**	0.149
0.015	**0.301**	0.298	0.126	**0.360**	0.075	0.229	**0.496**	0.134
0.01	**0.283**	0.265	0.113	**0.281**	0.063	0.183	**0.417**	0.105
0.005	**0.201**	0.193	0.069	**0.252**	0.036	0.138	**0.312**	0.072
0.001	**0.111**	0.103	0.043	**0.156**	0.022	0.108	**0.299**	0.054



(a) $K = 10$ $(means)$ (b) $K = 100$ $(means)$

(c) $K = 10$ $(variances)$ (d) $K = 100$ $(variances)$

(e) $K = 10$ $(both)$ (f) $K = 100$ $(both)$

Fig. C.4. Goodness-of-fit tests power of GC for $K \in \{10, 100\}$ in 2SV025 model

Table C.2. Goodness-of-fit tests power of Small Samples statistic for 2SV025 model

α	Small Samples statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	**0.645**	0.599	0.467	**0.727**	0.351	0.626	**0.836**	0.492
0.1	**0.544**	0.511	0.408	**0.647**	0.282	0.555	**0.769**	0.396
0.075	**0.507**	0.459	0.355	**0.596**	0.232	0.516	**0.733**	0.331
0.05	**0.453**	0.368	0.268	**0.513**	0.163	0.400	**0.676**	0.264
0.03	**0.370**	0.268	0.191	**0.448**	0.121	0.311	**0.614**	0.192
0.02	**0.339**	0.225	0.167	**0.37**	0.092	0.260	**0.551**	0.158
0.015	**0.309**	0.195	0.143	**0.353**	0.076	0.221	**0.507**	0.127
0.01	**0.242**	0.163	0.119	**0.322**	0.060	0.179	**0.395**	0.112
0.005	**0.190**	0.125	0.078	**0.268**	0.048	0.144	**0.326**	0.083
0.001	**0.114**	0.050	0.030	**0.138**	0.029	0.106	**0.299**	0.053

Table C.3. Goodness-of-fit tests power of Likelihood ratio symmetrized statistic for 2SV025 model

α	Likelihood ratio symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.628	**0.658**	0.580	0.625	0.371	**0.635**	**0.776**	0.379
0.1	0.560	**0.564**	0.492	**0.542**	0.277	0.537	**0.712**	0.290
0.075	0.498	**0.514**	0.405	**0.491**	0.194	0.481	**0.667**	0.240
0.05	0.399	0.391	0.316	**0.410**	0.150	**0.408**	**0.577**	0.182
0.03	0.301	0.305	0.256	**0.341**	0.100	**0.337**	**0.518**	0.135
0.02	0.253	**0.263**	0.174	**0.266**	0.085	0.262	**0.428**	0.100
0.015	0.217	0.219	0.142	**0.225**	0.074	**0.230**	**0.402**	0.072
0.01	0.171	0.175	0.070	**0.190**	0.047	**0.202**	**0.359**	0.066
0.005	0.106	0.122	0.044	**0.125**	0.032	**0.186**	**0.330**	0.047
0.001	0.045	**0.068**	0.035	0.067	0.015	**0.094**	**0.208**	0.033

Table C.4. Goodness-of-fit tests power of Chi-Square symmetrized statistic for 2SV025 model

α	Chi-Square symmetrized statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.634	**0.657**	0.573	**0.663**	0.384	0.604	**0.779**	0.396
0.1	0.558	**0.561**	0.489	**0.582**	0.269	0.509	**0.727**	0.311
0.075	0.517	**0.496**	0.422	**0.525**	0.213	0.448	**0.661**	0.257
0.05	0.429	**0.418**	0.331	**0.431**	0.163	0.376	**0.597**	0.205
0.03	0.325	0.314	0.275	**0.351**	0.128	**0.317**	**0.498**	0.134
0.02	0.243	0.240	0.199	**0.292**	0.106	**0.258**	**0.426**	0.105
0.015	0.203	0.230	0.133	**0.254**	0.099	**0.252**	**0.392**	0.093
0.01	0.187	0.174	0.081	**0.197**	0.065	**0.188**	**0.378**	0.077
0.005	0.107	0.114	0.049	**0.117**	0.036	**0.159**	**0.287**	0.066
0.001	0.063	0.073	0.028	**0.087**	0.016	**0.111**	**0.247**	0.034

Table C.5. Goodness-of-fit tests power of Hellinger distance statistic for 2SV025 model

α	Hellinger distance statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	0.588	**0.635**	0.550	0.565	0.356	**0.629**	**0.748**	0.365
0.1	0.502	**0.541**	0.449	0.469	0.277	**0.529**	**0.675**	0.274
0.075	0.445	**0.483**	0.383	0.425	0.182	**0.469**	**0.594**	0.218
0.05	0.343	**0.366**	0.276	0.350	0.137	**0.415**	**0.528**	0.181
0.03	0.247	0.262	0.200	**0.269**	0.093	**0.308**	**0.438**	0.111
0.02	0.205	0.206	0.156	**0.222**	0.062	**0.246**	**0.383**	0.085
0.015	0.163	0.191	0.097	**0.206**	0.055	**0.239**	**0.347**	0.066
0.01	0.133	**0.169**	0.060	0.155	0.039	**0.201**	**0.314**	0.057
0.005	0.094	**0.106**	0.041	0.083	0.028	**0.127**	**0.270**	0.041
0.001	0.029	**0.057**	0.031	0.045	0.012	**0.081**	**0.166**	0.028

Table C.6. Goodness-of-fit tests power of Chi-Square modified statistic for 2SV025 model

α	Chi-Square modified statistic							
	means					both		
	*	MCMC	GQ	GG	GC	GQ	GG	GC
0.15	**0.337**	0.248	0.269	0.270	**0.279**	**0.377**	0.270	0.214
0.1	**0.262**	0.163	0.170	0.184	**0.190**	**0.254**	0.184	0.134
0.075	**0.202**	0.140	0.136	0.148	**0.152**	**0.175**	0.148	0.113
0.05	**0.124**	0.097	0.091	**0.104**	**0.107**	0.075	**0.104**	0.081
0.03	**0.091**	**0.067**	0.053	**0.067**	**0.067**	0.063	**0.067**	0.049
0.02	**0.060**	**0.056**	0.037	0.048	0.042	**0.052**	0.048	0.029
0.015	**0.054**	0.032	0.031	**0.039**	0.018	**0.048**	**0.039**	0.023
0.01	**0.045**	0.021	0.019	**0.029**	0.016	**0.045**	**0.029**	0.014
0.005	**0.028**	0.014	0.013	**0.018**	0.011	**0.018**	**0.018**	0.011
0.001	0.007	**0.010**	**0.010**	**0.008**	0.007	0.004	**0.008**	0.005

Pavel SAMUSENKO

NONPARAMETRIC CRITERIA
FOR SPARSE CONTINGENCY TABLES

Doctoral Dissertation
Physical Sciences, Mathematics (01P)

NEPARAMETRINIAI KRITERIJAI
RETŲ ĮVYKIŲ DAŽNIŲ LENTELĖMS

Daktaro disertacija
Fiziniai mokslai, matematika (01P)

2012 12 21. 11,5 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto leidykla „Technika“,
Saulėtekio al. 11, LT-10223 Vilnius, *http://leidykla.vgtu.lt*
Spausdino UAB „Ciklonas“,
J. Jasinskio g. 15, LT-01111 Vilnius