Statistica Neerlandica (2005) Vol. 59, nr. 1, pp. 82–94

Testing log-linear models with inequality constraints: a comparison of asymptotic, bootstrap, and posterior predictive *p*-values

Francisca Galindo-Garre*

Department of Clinical Epidemiology and Biostatistics, University of Amsterdam, P.O. Box 22660, 1100 DD Amsterdam, The Netherlands

Jeroen K. Vermunt

Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

An important aspect of applied research is the assessment of the goodness-of-fit of an estimated statistical model. In the analysis of contingency tables, this usually involves determining the discrepancy between observed and estimated frequencies using the likelihood-ratio statistic. In models with inequality constraints, however, the asymptotic distribution of this statistic depends on the unknown model parameters and, as a result, there no longer exists an unique *p*-value. Bootstrap *p*-values obtained by replacing the unknown parameters by their maximum likelihood estimates may also be inaccurate, especially if many of the imposed inequality constraints are violated in the available sample. We describe the various problems associated with the use of asymptotic and bootstrap *p*-values and propose the use of Bayesian posterior predictive checks as a better alternative for assessing the fit of log-linear models with inequality constraints.

Key Words and Phrases: Posterior predictive *p*-values, inequality constraints, parametric bootstrap, order-restricted inference.

1 Introduction

The variables of interest in social sciences research are often of an ordinal nature. A possible modeling strategy with such variables is to analyze them using standard categorical data techniques, such as log-linear models, implying that the available information on the order of the categories is fully ignored. A better strategy is to use models with inequality restrictions, yielding a nonparametric approach that permits defining and testing ordinal hypotheses (see e.g. DARDANONI and FORCINA, 1998;

*Francisca Galindo-Garre performed this research as a part of her PhD dissertation project at Tilburg University.
F.GalindoGarre@amc.uva.np

© VVS, 2005. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

VERMUNT, 1999, 2001). Standard log-linear models can be easily transformed into ordinal log-linear models by including inequality constraints on the log-linear parameters. Within such a framework, a positive relationship between two ordinal variables, for example, would be represented by a log-linear model that restricts all local log-odds ratios to be at least zero. Even though quite some work has already been done on the maximum likelihood (ML) estimation and the testing of categorical data models with inequality restrictions (see for example ROBERTSON, WRIGHT and DYKSTRA (1988); CROON (1990, 1991); DARDANONI and FORCINA (1998); VERMUNT (1999, 2001), further research is needed on the problem of goodness-of-fit testing with the purpose of finding a method yielding accurate p-values.

A commonly used measure for assessing the goodness-of-fit of categorical data models estimated by ML is the likelihood-ratio test statistic (G^2), sometimes referred to as the deviance statistic. Under the assumption that the target model of interest is true (and that some other regularity conditions hold), this statistic has an asymptotic chi-square (χ^2) distribution with a number of degrees of freedom equal to the number of constraints implied by the target model or, equivalently, to the difference between the number of parameters of the saturated and the target model. In models with inequality constraints, however, one of the regularity conditions, namely, that the null hypothesis has to lie within the parameter space, does not hold (CHERNOFF, 1954). It therefore turns out that in such cases G^2 does not follow a χ^2 but a chi-bar-square ($\bar{\chi}^2$) distribution, which is a mixture of central χ^2 distributions. There are, however, two problems associated with the use of this $\bar{\chi}^2$ distribution: (1) to obtain a p-value one has to specify the parameter values under the null hypothesis, which is not straightforward to do when dealing with inequality constraints, and (2) the mixture weights needed to calculate p-values are very difficult to obtain if the number of inequality constraints is larger than 5.

Because classical p-values may be difficult to obtain using asymptotic methods, various authors have suggested using the parametric bootstrap, also called plug-in method, as an alternative method for assessing the fit of ordinal models (e.g. RITOV and GILULA, 1993; VERMUNT, 1999, 2001). This procedure involves approximating the empirical distribution of the goodness-of-fit test statistic by means of Monte Carlo simulations. Since the parametric bootstrap is based on less restrictive assumptions than the asymptotic method, one might expect that it can provide a reliable approximation of the distribution of the test statistic even in those cases in which the asymptotic distribution cannot be trusted or is difficult to derive. A simulation study by GALINDO and VERMUNT (2004), however, showed that parametric bootstrap p-values do not perform well in all situations.

Rather than using ML estimation methods and classical p-values for testing, one could use Bayesian methods for estimating and assessing the fit of categorical data models with inequality constraints (e.g. see HOIJTINK and MOLENAAR, 1997; VAN ONNA, 2002). Posterior predictive checks form the Bayesian alternative to the classical test procedure. For this purpose, one can either use test statistics, which are measures that depend only on the data, or discrepancy measures, which are functions of both the data and the model parameters (Gelman, Meng and Stern, 1996). The advantage of using discrepancy measures compared with using test statistics is that the former allow the dependence on unknown parameters; that is, there is no need to know the value of the parameters under the null hypothesis, as is the case in models with inequality constraints. Posterior predictive p-values are computed by drawing new data sets from the posterior predictive distribution and subsequently comparing these replicated data with the observed data. If the target model fits the data well, the replicated data should look similar to the observed data.

The aim of the current paper is to investigate whether posterior predictive p-values are a good alternative to classical p-values when assessing the fit of order-restricted categorical data models. Section 2 describes the log-linear model with the inequality constraints of interest, as well as methods to solve the estimation problem. Section 3 introduces the various classical and Bayesian methods to obtain p-values. In Section 4, these methods are compared with one another using an empirical example and a small simulation experiment. The paper ends with a short discussion.

2 Log-linear models with inequality constraints

Suppose one wishes to test whether there exists a positive relationship between two ordinal variables cross-classified in a two-way R-by-C contingency table. One possible modeling option that can be used for this purpose is log-linear analysis. Let $\mathbf{m}$ be the vector of expected frequencies, $\mathbf{X}$ the design matrix, and $\boldsymbol{\beta}$ the vector containing the K unknown log-linear parameters. The model of interest can be formulated as follows

$$\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}. \tag{1}$$

Since we are dealing with a two-way table, the two-variable association parameters in the saturated model represent the strength of the relationship between the row and the column variable. Galindo, Vermunt and Croon (2002) showed that a coding scheme based on the differences between adjacent categories yields two-way association parameters that are directly related to local log-odd ratios. In the case of a three-by-three table, for example, the corresponding design matrix takes on the form

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where each column is related to an element of the parameter vector. Columns six to nine represent the two-way association parameters. The correspondence between

© VVS, 2005

log-odds ratios and two-way association parameters can be seen by substituting the logarithms of the expected frequencies appearing in equation (1) by the corresponding log-linear parameters. For example, it can be seen that β_9 is in fact the local log odds ratio log θ_{22},

$$\begin{aligned}\log\theta_{22} &= \log(m_4) + \log(m_9) - \log(m_8) - \log(m_6)\\ &= (\beta_1 + \beta_3 + \beta_5 + \beta_9) + (\beta_1) - (\beta_1 + \beta_5) - (\beta_1 + \beta_3) = \beta_9.\end{aligned}$$

Whereas the above standard log-linear model does not make use of information on the order of the categories of the row and column variables, it can easily be transformed into an ordinal model by imposing nonnegativity constraints on the two-way association terms; that is, in the three-by-three example, by setting $\beta_k \geq 0$, for $6 \leq k \leq 9$. This yields a weak definition of a positive relationship between the two ordinal variables, namely, that all log-odds ratios are at least zero. In the general case, we say that $K = K_1 + K_2$, where K_1 is the number of unrestricted parameters and K_2 the number of order-restricted parameters.

For parameter estimation, we use a loglikelihood function based on assuming a Poisson sampling scheme,

$$l(\mathbf{n}|\boldsymbol{\beta}) = \sum_i n_i \log m_i - \sum_i m_i,$$

where n_i and m_i represent elements of $\mathbf{n}$, the vector of observed frequencies, and $\mathbf{m}$ respectively. The parameters of the order-constrained log-linear model can either be estimated by maximum likelihood or by Bayesian estimation methods. To solve the ML estimation problem, we use an activated constraints variant of the Newton-Raphson algorithm (Gill and Murray, 1974), in which unrestricted and non-activated order-restricted parameters are updated in the usual manner, whereas parameters corresponding to activated constraints are only updated in a particular iteration if that means that they will become positive (for more details, see Galindo *et al.* (2002)).

In the Bayesian estimation of the order-restricted log-linear model, unknown parameters are treated as random variables instead of constants and the inequality constraints are dealt with as a part of the prior distribution of the parameter rather than as a part of the likelihood function (Gelfand, Smith and Lee, 1992). In our case, the distribution of the constrained parameters is assumed to be truncated at zero, implying that a restricted parameter drawn from the posterior distribution, $p(\boldsymbol{\beta}|\mathbf{n})$, will never be negative. A random walk Metropolis–Hastings (M–H) algorithm (see Gelman, Carlin, Stern and Rubin (2003, Section 11.4)) is used for drawing samples from the posterior distribution of the parameters of the order-restricted log-linear model. The employed jumping distribution has the form of a univariate normal distribution for each β_k parameter. This distribution is truncated at zero for $k > K_1$. Two different methods can be used for drawing samples from truncated normal distributions. The first method involves generating a proposal β_k^* from an unconstrained normal distribution $N(\beta_k^s, \sigma_k^2)$ for each iteration $s + 1$ until a non-negative value occurs, which simply amounts to rejecting impermissible values

© VVS, 2005

for β_k^*. In the second method, if F is the normal cumulative distribution, F^{-1} the inverse normal cumulative distribution, and U a uniform (0, 1) variate, a proposal β^* can be derived from: $\beta^* \sim F^{-1}\{F(0) + U[1 - F(0)]\}$ (Gelfand *et al.*, 1992). Once the K parameters are sampled from the corresponding proposal distributions, the candidate parameter vector is accepted with probability

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\beta}^*|\mathbf{n})J(\boldsymbol{\beta}^s|\boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^s|\mathbf{n})J(\boldsymbol{\beta}^*|\boldsymbol{\beta}^s)}\right).$$

Here, $J(\boldsymbol{\beta}^*|\boldsymbol{\beta}^s)$ is the jumping distribution which equals $\prod_k f(\beta_k^*|\beta_k^s, \sigma_k^2)$, where $f(\beta_k^*|\beta_k^s, \sigma_k^2)$ equals the normal density function $N(\beta_k^*|\beta_k^s, \sigma_k^2)$ if the parameter is unconstrained and $[N(\beta_k^*|\beta_k^s, \sigma_k^2)]/[1 - F(0|\beta_k^s, \sigma_k^2)]$ if the parameter is constrained. The same type of formula applies to $J(\boldsymbol{\beta}^s|\boldsymbol{\beta}^*)$. Several independent parallel sequences are generated and the $\sqrt{\hat{R}}$ criterion described in Gelman *et al.* (2003, Section 11.6) is employed to determine convergence.

The iterations are started with 10000 burning in samples, with σ_k^2 being the inverse of the square of the number of parameters. Then, we performed another 10 000 burning in iterations, where σ_k^2 is equated to the estimated variance from the first samples divided by the square of the number of parameters. The σ_k^2 for the subsequent iterations were equated to the estimated variance from the second set of burning in samples divided by the number of parameters. The $\sqrt{\hat{R}}$ criterion was equated to 1.001 for each parameter and determined using six independent chains. Convergence was checked at each 50 000th iteration. We retained each 50th sample to compute expected a posteriori (EAP) or posterior mean estimates of the unknown parameters, which are defined as follows:

$$E(\boldsymbol{\beta}|\mathbf{n}) = \int \boldsymbol{\beta} p(\boldsymbol{\beta}|\mathbf{n}) \mathrm{d}\boldsymbol{\beta}.$$

After obtaining parameters estimates using one of the two estimation methods described above, the goodness-of-fit test can be performed by various methods. Five of these methods are described in the next section.

3 Methods for estimating *p*-values

The null hypothesis that the order-restricted model holds is tested against the general alternative, the saturated model. Since some of the model parameters under the null hypothesis may be on the boundary of the parameter space, the standard asymptotic theory does not apply and the asymptotic distribution of the test statistic does not need to be a χ^2 distribution. In this section, several methods for obtaining p-values are discussed. First, we introduce the test statistic and the discrepancy measure that we will use in the description of the methods for computation of the p-values. Subsequently, we derive the asymptotic distribution of the test statistic. At the end of this section, two non-asymptotic methods for computing p-values are described.

3.1 Test statistics and discrepancy measures

The likelihood-ratio test statistic or deviance is defined as

$$G^2 = 2\sum_i n_i \log\left(\frac{n_i}{m_i}\right), \tag{2}$$

where each m_i represents an expected frequency estimated by ML. In order to compare the posterior predictive checks described below with the classical approach, G^2 is also computed for samples simulated from the posterior distribution. RUBIN and STERN (1994) used this test statistic for monitoring a latent class model.

In addition to the use of test statistics, the Bayesian approach allows the use of summary measures that are functions of both the unknown parameters and the data (GELMAN *et al.*, 1996). Since we are interested in checking the goodness-of-fit of an ordinal categorical data model, a natural summary measure is the deviance, defined as

$$T(\mathbf{n}, \boldsymbol{\beta}) = 2\sum_i n_i \log\left(\frac{n_i}{m_i(\boldsymbol{\beta})}\right),$$

which is the discrepancy between the observed data and the frequencies sampled from the posterior distribution, $m_i(\boldsymbol{\beta})$.

3.2 Asymptotic p-values

Next we derive the asymptotic distribution of the likelihood ratio statistic. Suppose that the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate normal with 0 mean and variance-covariance matrix $\mathbf{I}$, where $\hat{\boldsymbol{\beta}}$ is the ML estimator of $\boldsymbol{\beta}$ and $\mathbf{I}$ can be approximated by the Fisher information matrix. Let Θ be the parameter space, which is a cone or linear space. Then, Theorem 2.1 in SHAPIRO (1985) shows that the asymptotic distribution of G^2 is the same as the distribution of a Wald statistic measuring the differences between unrestricted and restricted parameter estimates

$$\min_{\boldsymbol{\beta}\in\Theta}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $\mathbf{H}$ is a function of $\mathbf{I}$. This asymptotic distribution is $\bar{\chi}^2$, which is a mixture of χ^2 distributions with weights $w_\ell(\mathbf{H},\Theta)$, namely,

$$P[\bar{\chi}^2 \geq c] = \sum_{\ell=0}^{K_2} w_\ell(\mathbf{H}, \Theta) P[\chi^2_\ell \geq c]. \tag{3}$$

Here χ^2_ℓ denotes a chi-square random variable with ℓ degrees of freedom and $P[\chi^2_0 \geq c] = 0$. Each $w_\ell(\mathbf{H},\Theta)$ represents the probability that exactly ℓ constraints are activated in a particular sample, which depends on the matrix $\mathbf{H}$ and on Θ. For our model, the asymptotic distribution of G^2 turns out to be a mixture of $K_2 = (R - 1)(C - 1)$ chi-square distributions.

There are two kinds of problems in the application of the asymptotic method described in this section: (1) in order to find a p-value one has to specify the values of the unknown parameters under H_0, which is non-trivial when dealing with inequality constraints, and (2) the analytic computation of the weights of the $\bar{\chi}^2$ distribution is impossible if the number of constraints is larger than 5. Dardanoni and Forcina (1998, p. 1117) proposed a procedure to obtain fairly accurate estimates using Monte Carlo simulation methods. Their procedure involves drawing a reasonable number of parameter vectors from a normal distribution with mean equal to the hypothesized parameter values and a covariance matrix equal to the estimated information matrix under H_0, and subsequently projecting these parameter vectors into the restricted parameters space.

As far as the specification of the values of the unknown parameters under H_0 is concerned, there are two options. The first option is to replace the hypothesized parameter values by their ML estimates under the order-restricted model. The weights of the $\bar{\chi}^2$ are then approximated by drawing parameter vectors from $N(\hat{\boldsymbol{\beta}}, \mathbf{H})$. This procedure is referred to as a *local test* by Dardanoni and Forcina (1998). Although this local test may underestimate the number of inequalities that hold as equalities in the population, which will produce too small p values, it is expected that the difference between the nominal and the actual p-values will be small if the sample size is large relative to the number of cells in the contingency table. In the second option, it is assumed that all inequality constraints hold as equalities, which defines the least favorable value of the parameters under the model of interest (see e.g. Bartolucci and Forcina, 2000; Dardanoni and Forcina, 1998). The corresponding weights are computed by drawing parameter vectors from $N(0,\mathbf{H})$. We refer to this procedure as a *global test*. This global test tends to produce somewhat too large p-values, which means that the order-restricted model is accepted more often than expected given a certain significance level.

3.3 Plug-in p-values

The parametric bootstrap is a generally accepted method for estimating the distribution of G^2 when either the standard approximation does not apply or the accuracy of such an approximation is suspect. This procedure has been used by various authors to test the fit of models with inequality restrictions. For example, Ritov and Gilula (1993) proposed such a procedure in ML correspondence analysis with ordered category scores, and Galindo and Vermunt (2004) applied parametric bootstrap to test the goodness-of-fit of ordered row-column association models.

The parametric bootstrap replaces the unknown parameters by their ML estimates. The distribution of $T = G^2$ is approximated as follows. Firstly, R independent replicate samples $\mathbf{n}_1^*, \ldots, \mathbf{n}_r^*, \ldots, \mathbf{n}_R^*$ are drawn from the expected frequencies under the target model. Subsequently, the model of interest is estimated for each replicate sample $\mathbf{n}_r^*$ and the test statistic t_r^* is computed by equation (2). The bootstrap p-value is defined as follows:

$$p = \frac{\sum_{r=1}^{R} I(t_r^* > T)}{R},$$

where I is an indicator function taking the value one if the inequality holds and zero otherwise. In other words, the plug-in or bootstrap p-value is defined as the proportion of replication samples with a larger value than the original value of T.

Let $\hat{\boldsymbol{\beta}}^*$ be the vector of bootstrap parameter estimates. The asymptotic validity of the bootstrap requires that, with probability one, the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ equals the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. However, Andrews (2000) showed with a simple counter example that this condition does not hold when inequality constraints are imposed on the model parameters, implying that the parametric bootstrap may produce inaccurate p-values. Galindo and Vermunt (2004) showed in a simulation study that the parametric bootstrap may produce p-values that are slightly higher than expected given a certain nominal level, especially with weak relationship between variables combined with large samples.

3.4 Posterior predictive p-values

Posterior predictive checks are the Bayesian alternative to the classical statistical tests (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996; Berkhof, Mechelen and Hoijtink, 2000). A posterior predictive p-value is defined as the probability that a statistic $T(\mathbf{n}^{rep})$, which is solely a function of the replicated observations $\mathbf{n}^{rep}$, is larger than or equal to the observed value of $T(\mathbf{n})$ given that the model M_0 is true,

$$P_B = P[T(\mathbf{n}^{rep}) \geq T(\mathbf{n})|M_0, \mathbf{n}].$$

To assess the fit of the order-restricted model, $L = 6000$ parameter vectors (1000 for each of the 6 chains) are drawn from the posterior distribution using the M–H algorithm described in the previous section. For each parameter sample $\boldsymbol{\beta}^l$, a data set $\mathbf{n}^{rep,l}$ with the same size as the original data set is generated from the multinomial distribution defined by the expected frequencies under the model. Next, $T(\mathbf{n}) = G^2$ is computed by equation (2). Note that ML estimates for the order-restricted model must be obtained for each simulated sample making this method computationally intensive. The value of each $T(\mathbf{n}^{rep,l}) = G^2_{rep}$ is compared with the G^2 value for the observed data set. For each replicate sample, also the discrepancy measure $T(\mathbf{n}^{rep,l}, \boldsymbol{\beta}^l)$ is compared with $T(\mathbf{n}, \boldsymbol{\beta}^l)$, the discrepancy between the observed data and the parameter estimates under the model.

In the next section, we compare the two asymptotic p-values and the plug-in p-values to the posterior predictive p-values computed from the test statistic and the discrepancy measure described here by mean of a empirical example and a small simulation study.

4 An empirical example and a simulation experiment

4.1 Analysis of a four-by-four table

The order-restricted log-linear model will be illustrated with the analysis of a two-way contingency table taken from AGRESTI's textbook *Categorical Data Analysis* (AGRESTI, 2002, Table 9.3). The two variables of interest are attitude toward 'Teenage birth control' and attitude toward 'Premarital sex'. Both variables have four levels. The data are summarized in Table 1. The research question of interest is as to whether subjects having more favorable attitudes about teen birth control also tend to have more tolerant attitudes about premarital sex.

Table 2 reports the ML estimates of the two-way association parameter according to the order-restricted and the saturated model, as well as the EAP estimates for the order-restricted model. The latter were obtained using the M–H algorithm and assuming uniform priors for the unconstrained parameters and uniform densities in the admissible areas of the parameters space for the constrained parameters. In this example, there are $K_1 = (4-1)+(4-1)+1 = 7$ unconstrained main effect parameters and $K_2 = (4-1)(4-1) = 9$ two-way association parameter that are constrained to be zero in the independence model, constrained to be non-negative in the order-restricted model, and unconstrained in the saturated model.

From Table 2 it can be seen that the number of parameters of the ordinal model in which the non-negative constraints hold as equalities does not always correspond with the number of estimates with a negative value in the saturated model. Note that β_{10}, β_{13} and β_{14} are equated to zero in the ordinal model while only $\hat{\beta}_{10}$ and $\hat{\beta}_{14}$ have a negative value in the saturated model. Though the value of $\hat{\beta}_{13}$ under the saturated model is rather large, it is restricted to be zero under the ordinal model. Table 2 also shows that EAP estimates for restricted parameters are all positive.

Table 1. Two-way cross-tabulation of the opinion about premarital sex and the opinion about teenage birth control.

	Teenage birth control			
Premarital sex	Strongly disagree	Disagree	Agree	Strongly agree
Always wrong	81	68	60	38
Almost always wrong	24	26	29	14
Wrong only sometimes	18	41	74	42
Not wrong at all	36	57	161	157

Table 2. Estimates of the two-way association parameters obtained with the data of Table 1.

	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
ML Saturated	0.25	0.23	**–0.27**	0.74	0.48	0.16	**–0.36**	0.45	0.54
ML Order-constrained	0.25	0.14	**0.00**	0.52	0.61	**0.00**	**0.00**	0.33	0.53
EAP Order-constrained	0.20	0.32	0.12	0.44	0.46	0.14	0.17	0.22	0.42

Bold values are the parameters associated with the violated order restrictions.

© VVS, 2005

The independence model does not fit the data ($G^2 = 127.6$, $p = 0.000$), indicating that there is a significant association between the variables. The order-restricted log-linear model performs much better than the independence model ($G^2 = 1.584$). A likelihood-ratio test comparing these two nested models will probably be significant. The question of main interest in the context of this paper is which method yields the most reliable p-values when testing the goodness-of-fit of the latter model.

To assess the goodness-of-fit of the order-restricted model using the asymptotic theory, we have to assume an unique value for the population parameters. The p-value obtained when assuming that all inequalities hold as equalities in the population (the global test) equals 0.909 and the critical value for a nominal level $\alpha = 0.05$ is 12.28. According to the local test, the p-value equals 0.378 and the critical value for $\alpha = 0.05$ is 5.82. The difference between the p-values is caused by the fact that in the global test, larger weights are given to the χ^2 distributions with large number of degrees of freedom (see equation (3)) while in the local test larger weights are given to the χ^2 distribution with three or less degrees of freedom. The plug-in p-value equals 0.503 and is thus larger than the p-value obtained with the local test and smaller than the one from the global test. The obtained posterior predictive p-values are 0.355 for the test statistic $T(\mathbf{n}) = G^2$ and 0.461 for the discrepancy measure $T(\mathbf{n}, \boldsymbol{\beta})$, where it should be noted that the discrepancy measure is the most natural measure for assessing the fit of a model within a full Bayesian analysis. The fact that the p-value associated with the discrepancy measure is not very extreme indicates that the model describes the data well.

4.2 A small simulation experiment

In order to compare the performance of the asymptotic, plug-in, and posterior predictive p-values, 1000 data set where generated from a two-way cross table whose parameters were chosen to be equal to the order-restricted ML estimates of the empirical example. We took these parameter values because the two-way association terms were quite small, a situation corresponding to the most problematic case in the simulation study performed by Galindo and Vermunt (2004). For each sample, the order-restricted model was estimated and the five p-values described above were computed. The proportion of samples in which the order-restricted log-linear model was rejected at a significance level of 0.05 were 0.036 for the global test, 0.288 for the local test, 0.197 for the plug-in method, 0.320 for the posterior predictive check based on the test statistic, and 0.005 for the posterior predictive check based on the discrepancy measure. As can be seen, none of the methods yields a rejection proportion that is in agreement with the nominal α level. The global test and the posterior predictive check with a discrepancy measure are too conservative while the other methods are too liberal.

According to Bayarri and Berger (2000), in the perfect case, p-values should be uniformly distributed random variables. However, in most problems exact uniformity of p-values cannot be attained. From Figure 1 it can be seen that all

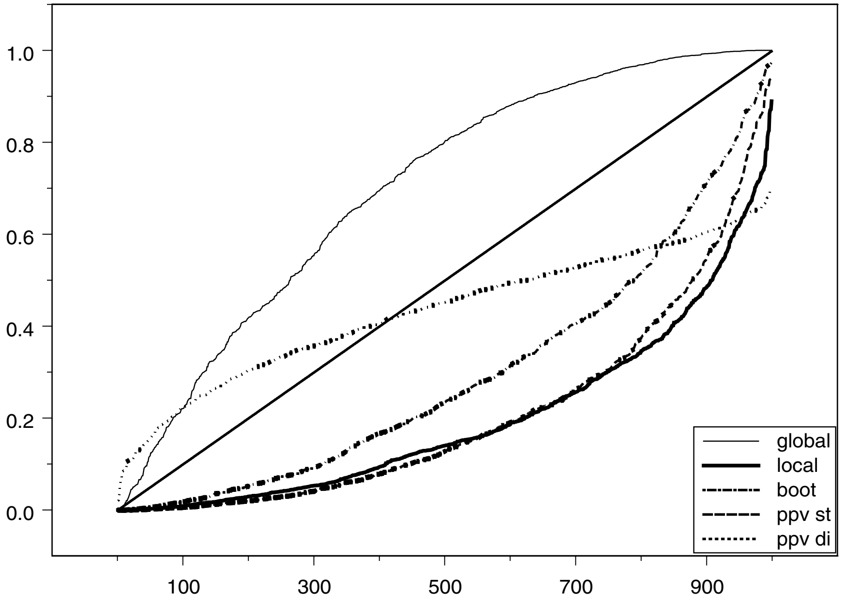

Fig. 1. Rejection rates at all alpha levels for the five methods.

five methods produce p-values that clearly deviate from the uniform distribution, and that, contrary to our expectation, the p-values obtained with the plug-in and the global test methods are closer to the uniform distribution than the posterior predictive p-values. These results are, however, in agreement with the results of Bayarri and Berger (2000).

Comparison of the local test to the posterior predictive p-values based on the statistic shows that these two methods produce similar results. The posterior predictive p-values has however the advantage that it is not based on asymptotic assumptions and that the parameters do not need to be fixed to a particular value (Gelman *et al.*, 2003).

5 Discussion

We compared classical and Bayesian methods for assessing the goodness-of-fit of order-restricted log-linear models. The main strengths and weaknesses of the various procedures were discussed. The main problem of the asymptotic methods is that the number of parameters that are close to the boundary in the population cannot be known, and that, depending on the assumption about this number, the test may be too conservative (if all the constraints are supposed activated) or too liberal (if the local test is used). Another disadvantage of the asymptotic methods is that in most situations the weights of the $\bar{\chi}^2$ distribution have to be approximated using computationally quite intensive simulation methods.

A disadvantage shared by the local test and the parametric bootstrap is that they both depend on the reliability of the ML estimates. In our small simulation study, we encountered that parameter estimates obtained by ML are biased, specially the order-constrained parameters. Whereas the number of parameters on the boundary in the postulated population was only two, the proportion of samples with three or more parameter estimates on the boundary was 0.749. This may be an explanation for the bad performance of the parametric bootstrap in the simulation study. However, the simulation study also shows that the posterior predictive checks do not produce more accurate p-values than the parametric bootstrap. Not only are the local test and the bootstrap procedure too liberal, but the posterior predictive p-value corresponding to the statistic turns out to be too liberal.

A more extended simulation study should be performed to assess whether our results can be generalized. An alternative solution to the testing problem may be to use plug-in p-values in which the unknown parameters are replaced by an estimate other than the maximum likelihood estimate, for example, the posterior mean or median.

References

AGRESTI, A. (2002), *Categorical data analysis*, New York: Wiley.

ANDREWS, D. W. K. (2000), Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space, *Econometrica* **68**, 399–405.

BARTOLUCCI, F. and A. FORCINA (2000), A likelihood ratio test for MTP$_2$ within binary variables, *The Annals of Statistics* **28**, 1206–1218.

BAYARRI, M. J. and O. BERGER (2000), p-values for composite null models, *Journal of the Americal Statistical Association* **95**, 1127–1142.

BERKHOF, J., I. VAN MECHELEN and H. HOIJTINK (2000), Posterior predictive checks: principles and discussion, *Computational Statistics* **15**, 337–354.

CHERNOFF, H. (1954), On the distribution of the likelihood ratio, *Annals of Mathematical Statistics* **25**, 573–578.

DARDANONI, V. and A. FORCINA (1998), A unified approach to likelihood inference on stochastic ordering in a nonparametric context, *Journal of the American Statistical Association* **93**, 1112–1123.

GALINDO, F. and J. K. VERMUNT (2004), The order-restricted association model: Two estimation algorithms and issues in testing, *Psychometrika*, in press.

GALINDO, F., J. K. VERMUNT and M. A. CROON (2002), Likelihood-ratio tests for order-restricted log-linear models: a comparison of asymptotic and bootstrap methods, *Metodología de las Ciencias del Comportamiento* **4**, 325–337.

GELFAND, A. E., A. F. SMITH and T. M. LEE (1992), Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling, *Journal of the American Statistical Association* **87**, 523–532.

GELMAN, A., X. MENG and H. STERN (1996), Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica* **6**, 733–807.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN (2003), *Bayesian data analysis*. Chapman & Hall, London.

GILL, P. E. and W. MURRAY (1974), *Numerical methods for constrained optimization*. Academic Press Inc., London.

© VVS, 2005

Hoijtink, H. and I. W. Molenaar (1997), A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks, *Psychometrika* **62**, 171–190.

Meng, X. (1994), Posterior predictive *p*-values, *The Annals of Statistics* **22**, 1142–1160.

Ritov, Y. and Z. Gilula (1993), Analysis of contingency tables by correspondence models subject to order constraints, *Journal of the American Statistical Association* **88**, 1380–1387.

Rubin, D. B. (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician, *The Annals of Statistics* **12**, 1151–1172.

Rubin, D. B. and H. S. Stern (1994), Testing in latent class models using a posterior predictive check distribution, in: A. von Eye and C. Clogg (eds), *Latent variable analysis: applications for developmental research*, Sage, California, 420–430.

Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrika* **72**, 133–144.

van Onna, M. (2002), Bayesian estimation and model selection in ordered latent class models for polytomous IRT models, *Psychometrika* **67**, 519–538.

Vermunt, J. K. (1999), Nonparametric models for ordinal data, *Sociological Methodology* **29**, 187–223.

Vermunt, J. K. (2001), The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models, *Applied Psychological Measurement* **25**, 283–294.

Received: June 2004. Revised: December 2004.

© VVS, 2005