Smoothing ordered sparse contingency tables and the χ^2 test

Klaus Abberger, University of Konstanz, Germany

Abstract
To estimate cell probabilities for ordered sparse contingency tables several smoothing techniques have been investigated. It has been recognized that nonparametric smoothing methods provide estimators of cell probabilities that have better performance than the pure frequency estimators. With the help of simulation examples it is shown in this paper that these smoothing techniques may help to get test which are more powerful than χ^2 test with raw data. But the distribution of the χ^2 statistics after smoothing is unknown. This distribution can also be estimated by simulation methods.

Keywords: nonparametric estimation, local polynomial smoothers, local likelihood, sparse contingency tables, χ^2 test, independence test

1 Introduction

There is a vast literature on nonparametric regression smoothers for continuous dependent and independent variables. Many different methods for estimation regression curves have been proposed, including kernel, local polynomial, spline and wavelet estimators. In this paper smoothing is applied to the estimation of probabilities in categorical data. In contrast to the situation of continuous data, where the benefits of smoothing (in form of scatterplot smoothers, for example) are obvious, the applicability of smoothing methods to discrete data is less clear.

For a d-dimensional contingency table with k_j ordered cells in the j-th dimension ($j = 1, 2, ..., d$) cell probabilities are usually estimated by frequency estimators. Tables which have small-to-moderate cell counts are called sparse tables. Such sparse tables occur when $k = \prod_{j=1}^{d} k_j$ (the total number of cells) and n (the total number of observations) are both large. For sparse tables it is recognized that nonparametric smoothing techniques provide estimators for the cell probabilities with better performance than frequency estimators (see Aerts et al. (1997) for discussion).

Knowing the advantages of smoothing frequencies we are interested in the consequences of smoothing on statistical inference, in particular in the behaviour of the χ^2 test of independence for two dimensional sparse contingency tables. In the next section two smoothing methods for categorical data recently discussed in the literature are presented. Section 3 contains the main part of this paper and shows power simulations for the χ^2 test of independence.

2 Smoothing methods for ordinal contingency tables

In this section two nonparametric estimators for ordinal contingency tables are presented. For a more comprehensive treatise on smoothing methods for discrete data see Simonoff and Tutz (2000).

Using weighted least-squares polynomial fitting is a possibility to smooth contingency tables. This is a well known method for smoothing scatterplots (Fan and Gijbels, 1996). For example, a local linear estimator $\hat{\pi}_{ij}$ for the probability of falling in the (i,j)th cell of an $R \times C$ two-dimensional table is $\hat{\beta}_0$, where $\hat{\beta}$ is the minimizer of

$$\sum_{k=1}^{R}\sum_{l=1}^{C}\left[p_{kl} - \beta_0 - \beta_1\left(\frac{i}{R} - \frac{k}{R}\right) - \beta_2\left(\frac{j}{C} - \frac{l}{C}\right)\right]^2 K_{h_R,h_C}(i,j,k,l,R,C), \quad (1)$$

with p_{kl} the relative frequencies and $K_{h_R,h_C}(\cdot)$ is a two dimensional kernel function with h_R and h_C the smoothing parameters for either rows and columns. A common technique for generating K_d is using the product of univariate kernels:

$$K_d(u) = \prod_{j=1}^{d} K_1(u_j). \quad (2)$$

A difficulty with local polynomial probability estimates is that while an arbitrary regression function can take on positive or negative values, a probability vector cannot take on negative values. The problem is that the estimator is based on the minimization of a local least squares criterion, which is appropriate for regression data, but not for categorical data.

To overcome these difficulties Simonoff (1998) introduced an estimator which is based on local likelihood, rather than local least squares. The local linear likelihood estimator for a two-dimensional table is $exp(\hat{\beta}_0)$, where $\hat{\beta}_0$ is the constant term of the minimizer of

$$\sum_{k=1}^{R}\sum_{l=1}^{c} \quad \left\{ n_{lk}\left[\beta_0+\beta_1\left(\frac{i}{R}-\frac{k}{R}\right)+\beta_2\left(\frac{j}{C}-\frac{l}{C}\right)\right] \right. \tag{3}$$
$$\left. -exp\left[\beta_0+\beta_1\left(\frac{i}{R}-\frac{k}{R}\right)+\beta_2\left(\frac{j}{C}-\frac{l}{C}\right)\right]\right\} K_{h_R,h_C}(i,j,k,l,R,C).$$

Thus it is guaranteed that the estimates will be nonnegative. For a detailed motivation and discussion of this estimator see Simonoff and Tutz (2000).

Although we prefer the likelihood method proposed by Simonoff the simulations in the next section are calculated with the LOESS procedure which grounds on local polynomial estimation. LOESS is used because of its fast implementation in S-Plus. For the simulation studies this is very important since for power simulations a huge amount of repetitions are required.

3 Power simulations for the χ^2 test

Being aware of the advantages of smoothing frequencies to estimate probabilities in sparse ordered contingency tables the purpose of this simulation study is to examine the effect of smoothing on the usual χ^2 test of independence. Does the improved estimates yield more powerful tests?

In the simulations examples the following data pattern is chosen. The dimension of the table is 5×5 and the total number of observations is always $n = 100$. For easy control of the dependency structure the underlying random process is bivariate normal with varying correlations. In the independence situation the correlation coefficient is set to zero. The 100 observations are generated from this bivariate standard normal. The resulting sample is standardized by the span so that the observed values lie between -1 and 1. This bivariate data set is then categorized. For the first dimension we have 5 categories. The observation falls in category I, if $-1 \leq x_i < -0.3$, in category II, if $-0.3 \leq x_i < -0.05$, in category III, if $-0.05 \leq x_i < 0.05$, in category VI, if $0.05 \leq x_i < 0.3$, and in category V, if $0.3 \leq x_i \leq 1$. The same categorization is applied to the second dimension. This procedure yields independent 5×5 contingency table. A "typical " data set is shown in Table 1.

It is possible to use a χ^2 test to test the independence of this data. Since the counts are small and even zero some times smoothing the table may be of advantage. As mentioned in the previous section for smoothing the LOESS procedure is used. The polynomial degree is fixed as one so that we arrive at local linear smoothing. We chose the in S-Plus implemented default smoothing parameter

	I	II	III	IV	V	
I	1	1	1	4	0	
II	4	10	10	18	2	
III	1	10	3	6	0	
IV	1	6	4	10	2	
V	1	1	2	2	0	
						100

Table 1: Example of cell counts a of categorized random sample from an uncorrelated bivariate normal distribution

which is $span = 2/3$, with $span$ the percentage of the total number of points used in the smoothing. Both the estimation method and the choice of smoothing parameter can be further improved and calibrated. But as we will see below even this straightforward but very fast smoothing method leads to appealing results.

The above described data generating algorithm is replicated 10,000 times to get impressions about the χ^2 statistic.

Figure 1 shows the estimated densities of χ^2 statistics once for the raw data and twice for the smoothed data. For the χ^2 statistic of the raw data there is nothing exceptional. Testing for independence with $\alpha = 0.05$ and $4 \cdot 4 = 16$ degrees of freedom leads to a simulation based estimate of $\hat{\alpha}_{da} = 0.0538$. So 538 of the 10,000 tests are significant. The fixed α is kept very well, although the usual rule of thumb that all cell counts should have a minimum size of five is violated.

Also shown in Figure 1 is the estimated density of χ^2 statistics after smoothing. Unsurprisingly, the usual χ^2 behaviour is destroyed. The χ^2 statistic after smoothing is not χ^2 distributed. Especially the scale is completely changed and quite different from the scale of the usual χ^2 statistic. So the standard χ^2 tables are not applicable to the smoothed χ^2.

This problem will be discussed further at the end of this section. For the power simulations the critical value can be estimated from the simulated density in Figure 1. Since the simulations are done under the null hypothesis of independence the $1-\alpha$ quantile of this density can be used as an estimate of the critical value. For $\alpha = 0.05$ the estimated critical value is 4.163184 in comparison with 26.3, which is the critical value of the χ^2 distribution with 16 degrees of freedom.

After fixing the critical values for both procedures, the correlation coefficient of the data generating bivariate normal process can be varied to study the power of the two procedures. 10,000 repetitions for the correlation coefficients

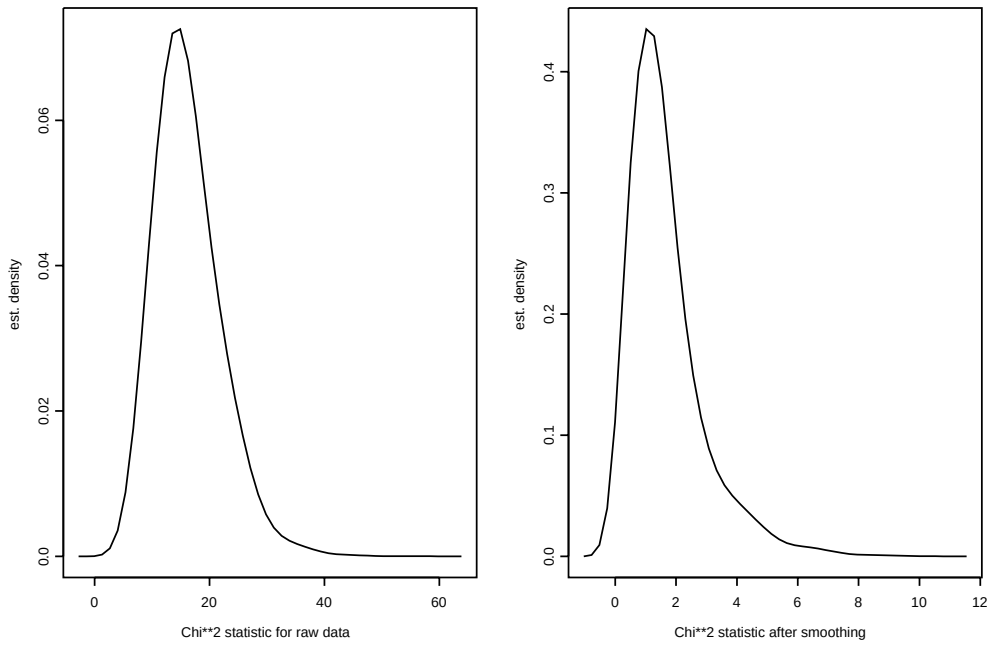

Figure 1: Monte Carlo estimated densities of χ^2 statistics for raw and smoothed data

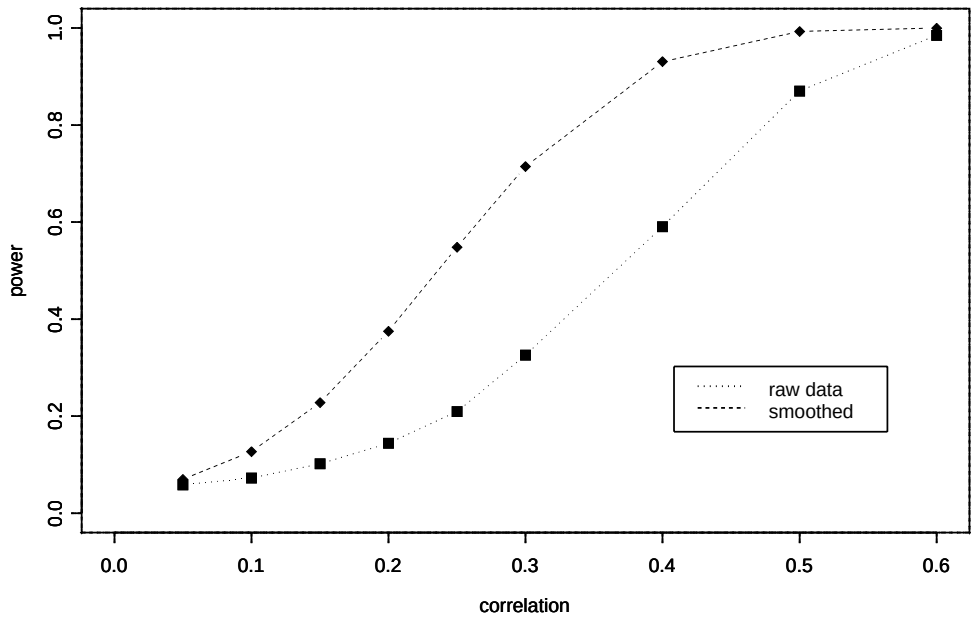

Figure 2: Monte Carlo estimated power of the χ^2 test for raw and smoothed data

$r = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6$ are calculated and the fraction of significant test is used as an estimate of the power.

Figure 2 shows the results of these calculations. The figure illustrates the benefits of smoothing very clear, because the power function after smoothing the frequencies is much steeper than the power function of the usual χ^2 test. Thus smoothing leads to a considerable improvement of the common χ^2 test relating to the power of the procedure.

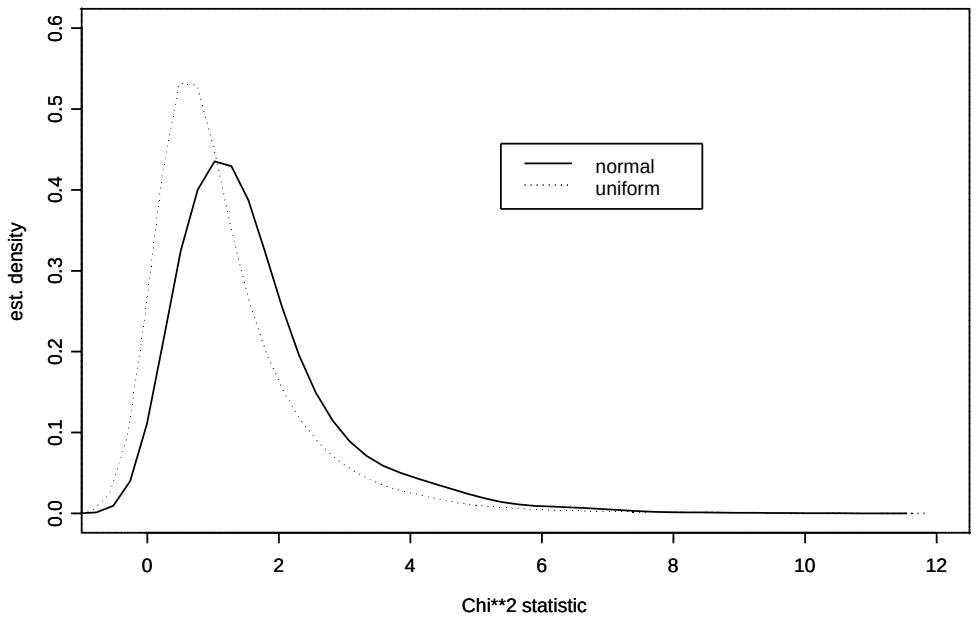

Figure 3: Monte Carlo estimated densities of χ^2 statistics for smoothed data with independent normal and independent uniform data generating process

The price we have to pay for this improvement is the impossibility of making use of the χ^2 distribution table. Instead we have to use more complicated methods.

Figure 3 shows again the density of the χ^2 statistic after smoothing independent bivariate normal date already included in Figure 1. In addition Figure 3 shows the simulation based estimate of the density of χ^2 statistics for smoothed categorized data generated by two independent uniform distributions. The two densities do not coincide. Thus the density of the χ^2 statistic and therefore the

critical value depends on the marginal distributions of the table. The critical value also depends on the kind of smoothing especially on the chosen bandwidth. Therefore the suitable critical value depends on the specific problem at hand.

A Monte Carlo based estimation method for this critical value of a specific table consists of the following steps: 1. Take the marginal distributions as fixed. 2. Chose smoothing method and smoothing parameter. 3. Draw bivariate observations from two independent uniform distributions. 4. Discretisize the data according to the relative marginal frequencies from step 1. 5. Calculate the χ^2 statistic. Now repeat the steps 1-5 many times to achieve an estimate of the specific distribution of the statistic under the null hypothesis and chose the $(1-\alpha)$ quantile of this distribution as critical value.

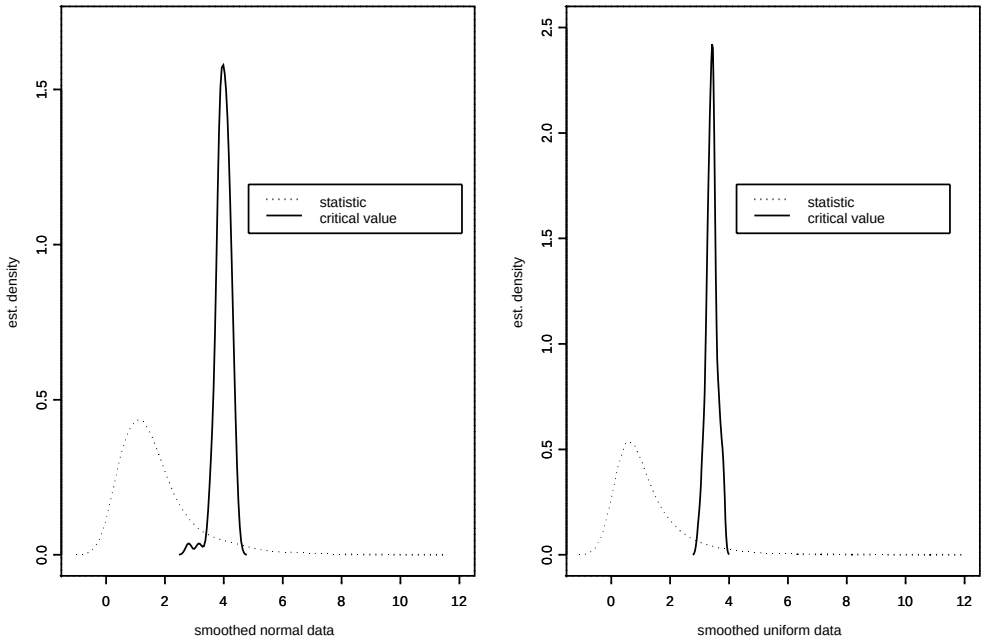

Figure 4: Estimated densities of Monte Carlo based estimates of critical values

Figure 4 illustrates the results of an simulation experiment based on the above described algorithm. The two data generating processes independent normal and independent uniform which are already used in Figure 3 are used again. For both processes we first draw one sample of size 100 which is used in step 1. Then the steps 3-5 are repeated 1,000 times each to generate a density and an estimate of the critical value. The whole procedure is then repeated 100 times to get an

estimated density of critical values. These new densities are presented in Figure 4 together with the densities of χ^2 statistics from Figure 3. From these calculations one can conclude that the above described algorithm yields quite accurate estimates of the critical value.

To sum up the various simulations in this section we can state first that smoothing ordered sparse contingency tables may lead to more powerful χ^2 test than testing without smoothing. The price we have to pay for this improvement is an uncertainty about the test distribution and furthermore about the suitable critical value. The critical value may be determined with simulation methods. Therefore an algorithm is proposed which seems to give suitable results. Improvements of the whole procedure are especially possible by the estimation method and the choice of smoothing parameter.

4 References

Aerts M., Augustyns I., Janssen P. (1997): Smoothing Sparse Multinomial Data Using Local Polynomial Fitting, *Nonparametric Statistics* , 8, 127-147.

Aerts M., Augustyns I., Janssen P. (1997): Local Polynomial Estimation of Contingency Table Cell Probabilities, *Statistics* , 30, 127-148.

Aerts M., Augustyns I., Janssen P. (1997): Sparse Contingency and Smoothing for Multinomial Data, *Statistics and Probability Letters* , 33, 41-48.

Cleveland W.S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots, *J. Amer. Statist. Assoc.*, 74, 829-836.

Fan J., Gijbels I. (1996): *Local Polynomial Modeling and its Applications*, Chapman and Hall, London.

Simonoff J.S. (1995): Smoothing Categorical Data, *L. Statist. Plann. Inf.*, 47, 41-69-156.

Simonoff J.S. (1998): Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation, *International Statistical Review* , 66, 137-156.

Simonoff J.S., Tutz G. (2000): Smoothing Methods for Discrete Data, *in: Smoothing and Regression: Approaches, Computation, and Application (Ed.: Schimek M. G.* , 193-228, Wiley, New York.