Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 10–11

LOCAL SMOOTHING WITH GIVEN MARGINALS

PIERRE JACOB AND PAULO EDUARDO OLIVEIRA

Abstract: In models using categorical data one may use adjacency relations to justify smoothing to improve upon simple histogram approximations of the probabilities. This is particularly convenient for sparsely observed or rather peaked distributions. Moreover, in a few models, prior knowledge of a marginal distribution is available. We adapt local polynomial estimators to include this partial information about the underlying distribution and give explicit representations for the proposed estimators. An application to a set of anthropological data is included.

Keywords: local polynomial smoothing, marginal distributions.
AMS Subject Classification (2010): 62H12, 62H17, 62G07.

1. Introduction

Models using categorical data usually assume that there is no relation between adjacent cells. This is not the case for continuous distributions, where many estimation procedures are based on the fact that observations falling near the approximation site do give some information about the function we are trying to estimate, whether this is a density or a regression function. This information by proximity is at the base of the modifications that have been proposed throughout the years to the histogram. The classical kernel or local polynomial estimators are, in fact, clever ways to use this idea to improve upon rough estimates. In many situations where categorical models are used, adjacency of cells does mean some kind of contiguity. This is often the case when using some scale to categorize collected data, thus becoming natural to use adjacency to construct estimates. This idea has been used to smooth over discrete distributions, with increased interest when few observations are available when compared with the number of cells of the underlying distribution or when the observations tend to concentrate too much in a few cells of the support, indicating that the underlying distribution is quite peaked. In such cases, the use of the classical cell frequency estimator seems inadequate: there would be many cells of the distribution support without any or very few observations, thus reflecting into an approximation for the distribution with many zeros. Convenient smoothing over adjacent cells does contribute to

Received April 7, 2010.

improve this. For one dimensional distributions Simonof [11], Hall and Titterington [7] smoothed the histogram with an uniform like distribution, while Burman [3] discretized the kernel estimator. More recently Simonof [12, 13], Dong and Simonof [4] or Aerts, Augustyns and Janssen [1, 2] studied discrete versions of local polynomial estimators for higher dimensional data. Faddy and Jones [5] proposed a semi-parametric smoothing method based on iterating Markov chain transformations. The above references considered estimation for one dimensional discrete distributions, although the methods are easily extended to higher dimensions. Another approach, used in Jacob and Oliveira [9], used the local polynomial approach but with respect to a relativized L^2-error, showing good performance for one dimensional data. The extension of these methods to higher dimensional data introduces a few difficulties because of the geometry of the space, as studied in Hall, Seifert and Turlach [6]. The relativized local polynomials that looked promising for one dimensional distributions, specially in presence of sparse data, but their extension to higher dimensional data showed to perform poorly compared with the discretized local polynomials. This seems linked to some geometric features that do not appear in one dimension.

Our first interest in this kind of problems arose when analyzing data from an anthropological study. The sample size was small when compared with the size of the support and observations tended to concentrate in a more or less defined area, although the full support of the distribution was known to include cells not observed at all. Moreover, a particular aspect of the anthropological study was that a marginal distribution was known. So, we were interested in smoothing over a contingency table and construct an approximate distribution that has a given marginal distribution.

Our estimates are obtained as solutions of a minimization problem and do have explicit formulæ, even in high dimension. For simplicity, we will explain our approach assuming the dimension is $d = 2$, but this is easily extended to higher dimensional problems, although the description of a few of the matrices considered might become somewhat cumbersome.

2. The framework

Consider $N = K \times L$ cells $C_{i,j}$, $i = 1 \ldots, K$, $j = 1, \ldots L$, arranged in a table $\mathbf{C} = (C_{i,j})$, and denote $\mathbf{P} = (P_{i,j})$ the probability distribution on $\mathbf{C}$. The observation counts over each cell are described by $\mathbf{N} = (N_{i,j})$, or equivalently, by the empirical probability distribution $\overline{\mathbf{P}} = (\overline{P}_{i,j} = N_{i,j}/n)$, where $n = \sum_{i,j} N_{i,j}$, on $\mathbf{C}$. Rearranging the rows in order to have a N-dimensional vector, $\mathbf{N}$ is multinomially distributed.

The table $\mathbf{C}$ might be identified with the unit square $[0,1]\times[0,1]$, considering equally sized rectangular cells with midpoints $(x_i, y_j) = \left(\frac{i-1/2}{K}, \frac{j-1/2}{L}\right)$, $i = 1, \ldots, K$, $j = 1, \ldots, L$. The special feature of this paper is that we assume $\mathbf{P}$ to be partially known. More precisely, we assume that the marginal distribution is known:

$$\Pi_i = \sum_{j=1}^{L} P_{i,j}, \qquad i = 1, ..., K.$$

In order to avoid computational difficulties with border and edge effects, we consider a replication of the given table $\mathbf{C}$, and likewise for the distribution $\mathbf{P}$ and observation counts $\mathbf{N}$. We enlarge $\mathbf{C}$, $\mathbf{P}$ and $\mathbf{N}$ using the well-known replication device (see, for example, Schuster [10]), by reflecting their cells with respect to each one of the four borders and edges. For the cell table $\mathbf{C}$, this enlarged table is identified with the square $[-1,2]\times[-1,2]$, the cells being equally sized rectangles with midpoints $(x_s, y_t) = \left(\frac{s-1/2}{K}, \frac{t-1/2}{L}\right)$, $s = 1-K, \ldots, 2K$, $t = 1-L, \ldots, 2L$. In this way, we have $9N$ cells, arranged in a $(3K)\times(3L)$ matrix. The original table $\mathbf{C}$ corresponds to the central $K\times L$ block of the enlarged matrix.

The enlargement of $\mathbf{P}$ is easily described. Let the matrices $\mathbf{P}_*$, $\mathbf{P}^*$ and $\mathbf{P}_*^*$ have (i,j) entries equal to $P_{K+1-i,j}$, $P_{i,L+1-j}$ and $P_{K+1-i,L+1-j}$, respectively. The enlarged $\mathbf{P}$ matrix is then

$$\begin{bmatrix} \mathbf{P}_*^* & \mathbf{P}_* & \mathbf{P}_*^* \\ \mathbf{P}^* & \mathbf{P} & \mathbf{P}^* \\ \mathbf{P}_*^* & \mathbf{P}_* & \mathbf{P}_*^* \end{bmatrix}.$$

For the enlargement of $\mathbf{C}$ and $\mathbf{N}$ we have similar descriptions. For these enlarged matrices the rows are indexed from $1-K$ to $2K$, while the columns are indexed from $1-L$ to $2L$.

We will have in mind the use of local polynomials of order at most 2. To define the functions to be optimized for the construction of the estimators, consider the indexes (s,t) of the enlarged matrices ordered lexicographically. In fact, any order of these indexes is acceptable, but we will refer to the lexicographic one. For each cell (i,j) of the original central table, define the

$(9N)\times 6$ matrix $\mathbf{X}_{i,j}$ whose (s,t)-line is

$$\begin{bmatrix}1 & (x_s-x_i) & (y_t-y_j) & (x_s-x_i)^2 & (x_s-x_i)(y_t-y_j) & (y_t-y_j)^2\end{bmatrix}.$$

For the smoothing, let $\mathcal{K}_1$ and $\mathcal{K}_2$ be bounded and symmetric densities with support included in $[-1/2,\,1/2]$. Given $h_1,\,h_2>0$, define $\mathcal{K}_{\mathbf{H}}(u,v)=h_1^{-1}h_2^{-1}\,\mathcal{K}_1(u/h_1)\mathcal{K}_2(v/h_2)$, where $\mathbf{H}=(h_1,h_2)$. For each (i,j) in the original table, that is, for $i=1,\ldots,K$ and $j=1,\ldots,L$, consider the $(9N)\times(9N)$ weight matrix

$$\begin{aligned}\mathbf{K}_{i,j}=\operatorname{diag}\big[&\mathcal{K}_{\mathbf{H}}(x_{1-L}-x_i,y_{1-K}-y_j),\ldots,\\ &\mathcal{K}_{\mathbf{H}}(x_s-x_i,y_t-y_j),\ldots,\mathcal{K}_{\mathbf{H}}(x_{2L}-x_i,y_{2K}-y_j)\big].\end{aligned}$$

Notice that, due to the symmetry of the weight functions and the replication device introduced, the local polynomial of odd degree $2p+1$ coincides with the local polynomial of even degree $2p$, so we will be interested in the local polynomials of degrees 0 or 2. We could go for higher order degrees but then the explicit expressions that we find would become too complex to be useful.

Finally, to introduce the notation to be used below, write

$$\overrightarrow{\mathbf{P}}=\left(\overline{P}_{1-K,1-L},\ldots,\overline{P}_{s,t},\ldots,\overline{P}_{2K,2L}\right)^t,$$

the vector of the empirical distribution $\overline{P}_{s,t}$, over the enlargement of $\mathbf{P}$, with the components listed in the lexicographic order.

3. The estimators

In this section we describe the estimators. As mentioned earlier, they will be constructed as solutions of an optimization problem, and are denoted CPS, for constrained local polynomial smoother, because of the marginal distribution being given. These estimators will appear as an additive correction of the usual local polynomial estimators.

For each cell $C_{i,j}$, the classical local polynomial smoother of degree 2, that we will denote by $\mathrm{PS}_{i,j}(2)$, appears as the solution of the minimization of

$$H_{i,j}=\left(\overrightarrow{\mathbf{P}}-\mathbf{X}_{i,j}\beta_{i,j}\right)^t\mathbf{K}_{i,j}\left(\overrightarrow{\mathbf{P}}-\mathbf{X}_{i,j}\beta_{i,j}\right), \tag{1}$$

where $\beta_{i,j}=(\beta_{0,i,j},\ldots,\beta_{5,i,j})^t$. If $\widehat{\beta}_{i,j}$ the minimizer of $H_{i,j}$, then $\mathrm{PS}_{i,j}(2)=\widehat{\beta}_{0,i,j}$, the constant term of $\widehat{\beta}_{i,j}$. The local polynomial smoothers of different degree p, $\mathrm{PS}_{i,j}(p)$, appear as solution of the minimization of $H_{i,j}$ with obvious changes of the matrices $\mathbf{X}_{i,j}$.

Whenever the cell $C_{i,j}$ is such that $\mathcal{K}_{\mathbf{H}}(x-x_i, y-y_j) = 0$, for each $(x,y) \notin [0,1]\times[0,1]$, the minimizer of $H_{i,j}$ is

$$\widehat{\beta}_{i,j} = \left(\mathbf{X}_{i,j}^t \mathbf{K}_{i,j} \mathbf{X}_{i,j}\right)^{-1} \mathbf{X}_{i,j}^t \mathbf{K}_{i,j} \overrightarrow{\mathbf{P}}, \tag{2}$$

which is exactly the usual local polynomial estimator of $\beta_{i,j}$. Now, the situation differs near the border and edges. In fact, it is well known that the usual local polynomial estimate gives some automatic correction to border and edge effects at the cost of a somewhat intricate expression for the regression coefficients. The fact that we use the replication device introduced earlier, that may seem somewhat painful to describe, gives in return the advantage of correcting border and edge effects without strictly modify the general expression of the estimate near the boundary of the table **C**. Evidently, it amounts to an automatic revision of the weights around each border or edge cell.

The above mentioned construction does not take into account the knowledge of the marginal distribution Π_i, $i = 1, \ldots, K$. In order to use this knowledge of the marginal distribution, we introduce a new estimate of $P_{i,j}$ as the solution of the optimization problem:

$$\begin{aligned} &\text{minimize} \sum_{\ell=1}^{L} H_{i,\ell} \\ &\text{subject to } \textstyle\sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1, \ldots, K. \end{aligned} \tag{3}$$

If $\widehat{\beta}_{i,j}^c$, $j = 1, \ldots, L$, are the minimizers of this problem, the constrained local polynomial smoother of degree 2 is $\mathrm{CPS}_{i,j}(2) = \widehat{\beta}_{0,i,j}^c$, the constant term of $\widehat{\beta}_{i,j}^c$. For constrained local smoothers with different degrees, the solution appears by modifying the matrices $\mathbf{X}_{i,j}$ appropriately. For each degree p for the local polynomial, we show in Appendix A, that

$$\mathrm{CPS}_{i,j}(p) = \mathrm{PS}_{i,j}(p) + \frac{1}{L}\left(\Pi_i - \sum_{\ell=1}^{L} \mathrm{PS}_{i,\ell}(p)\right). \tag{4}$$

More explicit representations for $\mathrm{CPS}_{i,j}(p)$, $p = 0, 2$ are given in Appendix B. The corresponding expressions for d-dimensional approximations are given in Appendix C.

In Jacob and Oliveira [9] a family of relative local smoothers was introduced. This family of estimators was constructed having in mind sparsely observed distributions and showed a good performance, especially for peaked distributions. We considered the higher dimensional extension of this family, obtained by solving the optimization problems:

$$\begin{aligned} &\text{minimize } H_i^* = \sum_{\ell=1}^{L} \frac{1}{\beta_{0,i,\ell}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right)^t \mathbf{W}_{i,\ell} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right), \\ &\text{subject to } \sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1, \ldots, K, \end{aligned} \tag{5}$$

where $\mathbf{W}_{i,\ell}$ is the weight matrix $\mathbf{K}_{i,\ell}$ normalized so the weights sum up to 1. It is possible to find matricial and even explicit expressions of the estimators obtained, proceeding as in [9]. However, these estimators showed a poor performance with respect to the constrained local smoothers. This seems due to a different behaviour of a term of the form $\mathbf{e}^t \left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right) \mathbf{e}$, where $\mathbf{e} = (1, \ldots, 1)$ and $\mathbf{A}_{i,j}$ is defined in [9] with obvious adaptations for higher dimensional distributions. In fact, in dimension 1, this term is always nonnegative while for higher dimensions it might become negative (see Jacob and Oliveira [8] for computational details).

4. Bandwidth selection

As usual with smoothing methods, a crucial step is the bandwidth choice. We discuss this in more detail for the local smoother of degree 0. We looked at this from two perspectives. Firstly, assume the $P_{i,j}$ are obtained as the result of a discretization of a continuous underlying probability distribution with density function f on $[0,1] \times [0,1]$: $P_{i,j} = \int_{C_{i,j}} f(x,y)\ d(x,y)$, $i = 1, \ldots, K$, $j = 1, \ldots, L$. Expanding f in a Taylor polynomial of order 2, using the lower left point of the (i,j) cell as reference point, gives an approximation to the bias of $\mathrm{PS}_{i,j}(0)$ as, taking into account the symmetry of the weight functions:

$$\frac{1}{2K^3L} \frac{\partial^2 f}{\partial x^2} \sum_z z^2 w_1(z) + \frac{1}{2KL^3} \frac{\partial^2 f}{\partial y^2} \sum_z z^2 w_2(z).$$

Proceeding likewise with the variance and summing up with respect to (i,j), if we assume that the underlying density f has axial symmetry with respect to the mid point of the square $[0,1] \times [0,1]$, we find the approximation to the mean square error:

$$\frac{1}{n}\sum_z w_1^2(z)\sum_z w_2^2(z) + \frac{1}{4K^2L^2}\left(\frac{A\sigma_1^2}{K^2}+\frac{B\sigma_2^2}{L^2}\right)^2$$

where $\sigma_1^2 = \sum_z z^2 w_1(z)$, $\sigma_2^2 = \sum_z z^2 w_2(z)$, are the second order moments of the weight functions w_1 and w_2, respectively, $A = \sum_{i,j} \frac{\partial^2 f}{\partial x^2}(x_i, y_j)$ and $B = \sum_{i,j} \frac{\partial^2 f}{\partial y^2} f(x_i, y_j)$. The above expression clearly depends on the weight functions and, in particular, in their supports. An approach to the choice of the bandwidth would be to take some reference weight function and choose the discretization that minimizes this approximation to the MSE. It seems reasonable to take as reference density a product of one dimensional densities. We took, in accordance with the underlying distributions used for simulation in the literature (see, for example, [11, 12, 4, 1]), product Beta densities with equal parameters, so the axial symmetry is satisfied. The minimization of the above expression always leads to small supports on the weight functions with, typically 3 or 5 points, regardless of the size of the support and spread of the observations. This seems appropriate if the sample is large, with a good coverage of the all cells in the table.

For discrete distributions it often happens that the support is known and large, while the observations tend to be few and, especially, we are far from having observations at every cell. Using the weight functions that follow from the previous approach would leave many cells with a zero approximation for its probability. We propose, to avoid this problem, the following algorithm:

(1) the support of the product weight function $w_1(\cdot)w_2(\cdot)$ is a rectangle with odd sides, that will be centered at the point where estimation is being made;

(2) for each point (x_i, y_j) identify the smallest, with respect to area of the rectangle, odd-sided rectangle centered at (x_i, y_j) that will find some observation on the table; let $u_{1,i,j}^*$ be the size of this rectangle in the horizontal direction and $u_{2,i,j}^*$ the size in the vertical direction; define $u_1^* = \max_{i,j} u_{1,i,j}^*$, $u_2^* = \max_{i,j} u_{2,i,j}^*$;

(3) take $u_1 \geq u_1^*$, $u_2 \geq u_2^*$; define w_1 as the discretization of the Epanechnikov kernel over u_1 cells and w_2 as the discretization of the Epanechnikov kernel over u_2 cells.

The idea behind this algorithm is to be sure that the smoothing really does give some mass to each cell. The support sizes for the weight functions u_1

and u_2 should be slightly larger than u_1^* and u_2^*. The amount of enlargement used depends on the type of error we want to prevail and also the degree of the local polynomial we apply. Moreover, this enlargement helps to avoid the problem of negative approximations being produced by the smoothing and correction due to the marginal distribution. From a practical point of view, based on the simulated results, taking u_1 and u_2 slightly but strictly larger than u_1^* and u_2^* can capture reasonably the almost optimal choice of the bandwidths with respect to mean square error. A typical choice, used in our simulations giving prevalence to the mean square error, takes $u_1 = u_1^* + 4$ and $u_2 = u_2^* + 4$. If we are interested in other types of error measure considered in the literature for sparsely observed data, different enlargements should be adopted. Again, the simulations performed suggest smaller enlargements for the support of the weight functions than the ones used for the mean square error.

5. Simulation results

In this section we compare the performance of the different constrained smoothers with respect to the mean sum of squared errors MSSE and to the sup-norm NSUP:

$$\mathrm{MSSE}(\mathbf{P}^*) = \mathrm{E}\left(\sum_{i=1}^{K}\sum_{j=1}^{L}\left(P_{i,j}^* - P_{i,j}\right)^2\right),$$

$$\mathrm{NSUP}(\mathbf{P}^*) = \sup_{1\leq i\leq K,\, 1\leq j\leq L}\left|P_{i,j}^* - P_{i,j}\right|,$$

where $\mathbf{P}^* = \left(P_{i,j}^*\right)$ is the estimate for $\mathbf{P} = (P_{i,j})$. We considered three distributions obtained as mixtures of discretized Beta's with different parameters. These type of distributions are used in the literature for smoothers over a discrete one dimensional support (see, for example [11, 12, 4, 1]). In Table 1 we graph the three distributions used for simulation. The given marginal, considered as know for estimation purpose, was the uniform for distribution 1, and Beta(.8,.8) for the other distributions. The conditional distribution over the second coordinate given the first coordinate, was of the form Beta(a, a) with a ranging from -1 to 2 for distribution 1, from .5 to 3 for distribution 2, and from .3 to 5.2 for distribution 3. These distributions were discretized over a 30×30 table. We performed simulations with the number of observations being 450, half of the number of cells in the support of the distribution,

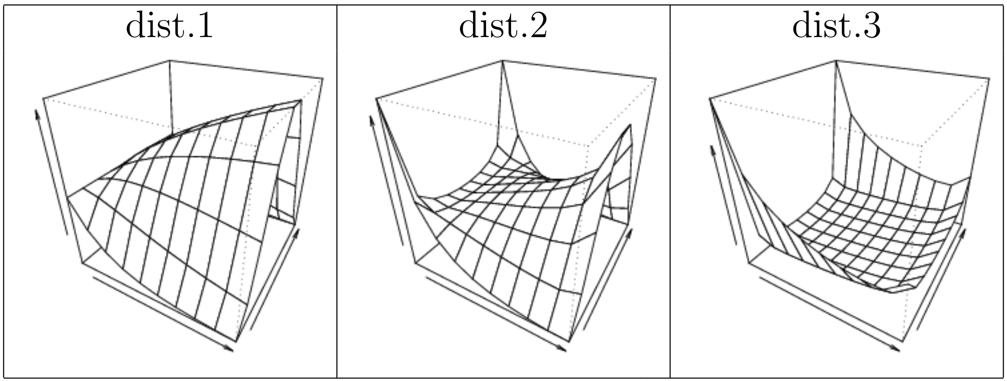

TABLE 1. Distributions used for simulation.

		dist.1		dist.2		dist.3	
		$\times 10^{-5}$		$\times 10^{-5}$		$\times 10^{-4}$	
bandwidth		optimal	data driven	optimal	data driven	optimal	data driven
CPS(0)	450	3.063	5.668	7.536	9.262	3.032	3.321
	900	1.990	3.035	6.191	5.971	8.776	14.148
CPS(2)	450	45.930	294.969	44.866	159.471	10.818	22.280
	900	27.882	171.420	25.615	152.593	8.851	44.996

TABLE 2. Simulated values for the MSSE.

and equal to 900. All the numerical results were obtained by running 500 Monte Carlo samples in each of the considered situations. The simulated values for optimal choice of the bandwidth and applying the algorithm described above are reported in Tables 2 and 3. The data driven choices for the bandwidth are close to the almost optimal simulated values and, even, sometimes better (the almost optimal was computed forcing $u_1 = u_2$, which seemed adequate with respect to support of the distribution but is apparently not taking into account any asymmetry on the data distribution). Our algorithm seems not to perform so well for the local smoother of degree 2 with few observations, but this is probably due to the good behaviour for this smoother being only asymptotical. Besides, the local smoother of degree 2 is more prone to produce negative approximations, thus requiring more data available.

6. An application

The data of the anthropological study consisted in a series of measurements of the bone condition on people that had been killed in consequence of a criminal action. One had a sample of size 164 observations over a 19 × 4 table,

		dist.1		dist.2		dist.3	
		$\times 10^{-5}$		$\times 10^{-5}$		$\times 10^{-4}$	
bandwidth		optimal	data driven	optimal	data driven	optimal	data driven
CPS(0)	450	6.039	8.805	3.228	3.561	4.926	6.581
	900	5.125	6.540	2.909	3.400	9.288	9.450
CPS(2)	450	26.836	48.394	3.710	5.336	7.807	9.278
	900	20.191	55.180	2.864	6.735	6.688	10.129

TABLE 3. Simulated values for the NSUP.

a somewhat sparse situation. The 19 lines correspond to age intervals and the four columns to a scale classification depending on the bone condition of a corpse. For forensic purposes, the bone condition of a corpse is observable, and you want to estimate the distribution of the age at the moment of death given this observed bone condition. It is reasonable to assume that the bone condition does not have an influence on the fact that someone is criminally attacked, so we may infer the joint distribution of age and bone condition for the entire population from this very specially selected sample. Using our algorithms, we construct estimates for the joint distribution over the complete table, with the marginal corresponding to the age distribution of the population given. The observations available indicate a choice of $u_1^* = 7$ and $u_2^* = 3$. We take $u_1 = u_1^* = 7$, as this the largest possible value given the size of the support. On the other direction, we take $u_2 = 5$, enlarging somewhat u_2^*. The estimates are given in Table 4. The local smoother of degree 2 produces a few negative estimates for the probabilities. Enlarging a little further does not have an effect on this problem. One way around this problem could be projecting over the subspace of distributions with zero probability on the cells where we find negative estimates and correcting the marginal afterwards. This has the drawback of assigning zero probability to a few cells of the support, but it is clearly better than having negative estimates for the probabilities. Of course, this is an effect of a sparsely observed distribution. Notice the estimates produced for the first row of the support all coincide. All these values are constructed from the two observations that fell into cells (4,1) and (4,2).

In Table 5 we show a graphical representation of the local polynomials of degrees 0 and 2, together with the frequency estimator. In all three representations the vertical scale ranges from 0 to 0.08. Projecting the local polynomial of degree 2 to avoid negative estimates keeps the overall picture.

Observed				known								
cell counts				marginal	CPS(0)				CPS(2)			
0	0	0	0	0.052094	0.013023	0.013023	0.013023	0.013023	0.013023	0.013023	0.013023	0.013023
0	0	0	0	0.051904	0.013132	0.013043	0.012908	0.012820	0.013027	0.012928	0.013024	0.012925
0	0	0	0	0.055966	0.015703	0.014824	0.013344	0.012096	0.016487	0.013393	0.011993	0.014094
1	1	0	0	0.066500	0.021679	0.019034	0.014680	0.011107	0.030394	0.018765	0.009099	0.008243
12	3	1	0	0.076370	0.027660	0.022976	0.015605	0.010129	0.046536	0.026488	0.008070	-0.004723
10	8	1	0	0.078665	0.029655	0.023988	0.015341	0.009680	0.051176	0.028598	0.011009	-0.012117
4	13	0	1	0.073527	0.026769	0.021593	0.014373	0.010793	0.036593	0.025705	0.020188	-0.008957
2	9	0	4	0.074428	0.023314	0.019820	0.015883	0.015411	0.021712	0.024665	0.029848	-0.001797
0	9	6	0	0.070347	0.017908	0.016944	0.016750	0.018745	0.011152	0.020364	0.031362	0.007468
0	5	3	5	0.066254	0.013431	0.014749	0.017340	0.020734	0.001990	0.013878	0.031179	0.019208
0	4	4	9	0.062042	0.009693	0.013088	0.017602	0.021658	-0.000922	0.006827	0.028640	0.027497
0	6	0	6	0.055180	0.005894	0.010978	0.016943	0.021365	-0.000864	0.000045	0.023749	0.032252
0	2	3	10	0.053197	0.005025	0.010493	0.016715	0.020963	-0.002411	-0.002480	0.022064	0.036028
0	2	3	8	0.051966	0.005603	0.010565	0.016134	0.019663	-0.002018	-0.001923	0.019790	0.036117
0	0	0	4	0.043835	0.005346	0.009181	0.013405	0.015903	0.003888	0.000725	0.013268	0.025953
0	0	0	1	0.033610	0.005017	0.007418	0.009908	0.011267	0.005479	0.001741	0.008962	0.017429
0	0	0	3	0.019477	0.002904	0.004347	0.005760	0.006466	0.003067	0.000329	0.005056	0.011024
0	0	0	0	0.010469	0.001193	0.002239	0.003263	0.003775	0.001451	-0.000634	0.002690	0.006964
0	0	0	1	0.004169	0.000027	0.000772	0.001503	0.001868	0.000408	-0.001225	0.001004	0.003982

Table 4. Constrained local polynomial of degrees 0 and 2.

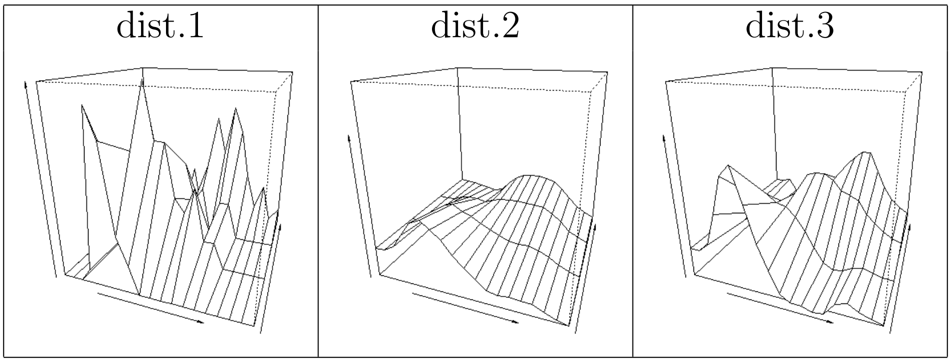

Table 5. Graphical representation of the data and the local polynomials of degrees 0 and 2.

Although the local polynomial of degree 0 produces a smoother approximation, the local polynomial of degree 2 seems to be able to capture a few more details on the link between the two margins.

Appendix A. Derivation of the CPS estimators

In order solve the optimization problem (3), introduce the Lagrange function

$$H_i = \sum_{\ell=1}^{L} H_{i,\ell} + 2\nu \left(\sum_{\ell=1}^{L} \mathbf{h}^t \beta_{i,\ell} - \Pi_i \right), \tag{6}$$

where $\mathbf{h}=(1,0,0,0,0,0)^t$ and 2ν stands for the Lagrange multiplier. The first order conditions are

$$\frac{\partial H_i}{\partial \beta_{i,\ell}} = -2\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\overrightarrow{\mathbf{P}} + 2\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\beta_{i,\ell} + 2\nu\mathbf{h} = 0, \qquad \ell = 1,\ldots,L. \tag{7}$$

From this system of equations and $\mathbf{h}^t\beta_{i,\ell} = \beta_{0,i,\ell}$ we obtain, for each $\ell = 1,\ldots,L$, the following matricial expression

$$\begin{bmatrix} \mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell} & \mathbf{h} \\ \mathbf{h}^t & 0 \end{bmatrix} \times \begin{bmatrix} \beta_{i,\ell} \\ \nu \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\overrightarrow{\mathbf{P}} \\ \beta_{0,i,\ell} \end{bmatrix}. \tag{8}$$

Multiplying on the left by the matrix

$$\begin{bmatrix} \mathrm{Id} & 0 \\ -\mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1} & 1 \end{bmatrix},$$

it follows

$$\begin{bmatrix} \mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell} & \mathbf{h} \\ 0 & -\mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{h} \end{bmatrix} \times \begin{bmatrix} \beta_{i,\ell} \\ \nu \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\overrightarrow{\mathbf{P}} \\ \beta_{0,i,\ell} - \mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\overrightarrow{\mathbf{P}} \end{bmatrix}.$$

Summing the last line of this equation over $\ell = 1,\ldots,L$, gives

$$\begin{aligned} \nu &= \frac{-\Pi_i + \sum_{\ell=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\overrightarrow{\mathbf{P}}}{\sum_{\ell=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{h}} \\ &= \frac{-\Pi_i + \sum_{\ell=1}^L \widehat{\beta}_{0,i,\ell}}{\sum_{\ell=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{h}}. \end{aligned} \tag{9}$$

Now from (7) we derive

$$\widehat{\beta}_{i,j}^c = \left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1}\left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\overrightarrow{\mathbf{P}} - \nu\mathbf{h}\right) = \widehat{\beta}_{i,j} - \nu\left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1}\mathbf{h}.$$

Multiplying on the left by $\mathbf{h}^t$ and using (9), it follows

$$\widehat{\beta}_{0,i,j}^c = \widehat{\beta}_{0,i,j} + \frac{\mathbf{h}^t\left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1}\mathbf{h}}{\sum_{\ell=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{h}}\left(\Pi_i - \sum_{\ell=1}^L \widehat{\beta}_{0,i,\ell}\right). \tag{10}$$

Finally, observe that, under our construction of $\mathbf{X}_{i,\ell}$ and $\mathbf{K}_{i,\ell}$, the matrix $\mathbf{h}^t\left(\mathbf{X}_{i,\ell}^t\mathbf{K}_{i,\ell}\mathbf{X}_{i,\ell}\right)^{-1}\mathbf{h}$ does not depend on the index ℓ, so (4) follows.

Appendix B. An explicit expression for $\mathrm{CPS}_{i,j}(p)$, $p \leq 2$ (dimension 2)

We start by the case $p = 2$. To give an explicit representation for $\mathrm{CPS}_{i,j}(2)$ given by (4), we will find an expression for the local polynomial smoother

$$\mathrm{PS}_{i,j}(2) = \mathbf{h}^t\widehat{\beta}_{i,j}^c = \mathbf{h}^t\left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1}\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\overrightarrow{\mathbf{P}} = \sum_{s,t} R_{s,t}\overline{P}_{s,t}, \qquad (11)$$

where the coefficients $R_{s,t}$ are to be determined. We start by noting that, with replication device we used, all the cells of the original table $\mathbf{C}$ are interior, so $\sum_{s,t}\mathcal{K}_H(x_s - x_i, y_t - y_j)$ does not depend upon (i,j), thus we may, in the expression above, replace the weights defined by the entries of $\mathbf{K}_{i,j}$, by normalized weights. Denote the matrix of normalized weights by $\mathbf{W}_{i,j}$. Recalling that $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$, it is convenient to introduce the system of product normalized weights

$$\begin{aligned} w_1(s-i)w_2(t-j) &= \frac{\mathcal{K}_{\mathbf{H}}(x_s - x_i, y_t - y_j)}{\sum_{u,v}\mathcal{K}_{\mathbf{H}}(x_s - x_u, y_t - y_v)} \\ &= \frac{\mathcal{K}_1(x_s - x_i)\mathcal{K}_2(y_t - y_j)}{\sum_u \mathcal{K}_1(x_s - x_u)\sum_v \mathcal{K}_2(y_t - y_v)}, \end{aligned}$$

and the sums

$$S_{\alpha,\beta} = \sum_{s,t}(x_s - x_i)^\alpha (y_t - y_j)^\beta w_1(s-i)w_2(t-j), \qquad \text{for } \alpha, \beta \geq 0 \text{ and } \alpha + \beta \leq 4.$$

The symmetry of $\mathcal{K}_1$ and $\mathcal{K}_2$ entails the symmetry of p_1 and p_2, hence $S_{\alpha,\beta} = 0$ if at least one of the coefficients α or β is odd. We have already defined the second order moments of the weight functions: $\sigma_1^2 = \sum_z z^2 w_1(z)$ and $\sigma_2^2 = \sum_z z^2 w_2(z)$. Define now the fourth moments of these weight functions:

$$\tau_1^4 = \sum_z z^4 w_1(z), \qquad \tau_2^4 = \sum_z z^4 w_2(z).$$

Notice that the choice of bandwidths h_1 and h_2 translates into choosing the number of points in the support of w_1 and w_2, respectively.

It is now easy to check that $S_{0,0}=1$, $S_{2,0}=\sigma_1^2/K^2$, $S_{0,2}=\sigma_2^2/L^2$, $S_{2,2}=\sigma_1^2\sigma_2^2/K^2L^2$, $S_{4,0}=\tau_1^4/K^4$, and $S_{0,4}=\tau_2^4/L^4$. These sums may be used to describe the matrix

$$\mathbf{X}_{i,j}^t\mathbf{W}_{i,j}\mathbf{X}_{i,j}=\begin{bmatrix} S_{0,0} & S_{1,0} & S_{0,1} & S_{2,0} & S_{1,1} & S_{0,2} \\ S_{1,0} & S_{2,0} & S_{1,1} & S_{3,0} & S_{2,1} & S_{1,2} \\ S_{0,1} & S_{1,1} & S_{0,2} & S_{2,1} & S_{1,2} & S_{0,3} \\ S_{2,0} & S_{3,0} & S_{2,1} & S_{4,0} & S_{3,1} & S_{2,2} \\ S_{1,1} & S_{2,1} & S_{1,2} & S_{3,1} & S_{2,2} & S_{1,3} \\ S_{0,2} & S_{1,2} & S_{0,3} & S_{2,2} & S_{1,3} & S_{0,4} \end{bmatrix}.$$

As $\left(\mathbf{X}_{i,j}^t\mathbf{W}_{i,j}\mathbf{X}_{i,j}\right)^{-1}$ is left multiplied by $\mathbf{h}^t$, we only need the first line of this matrix. A simple calculation shows that $\mathbf{h}^t\left(\mathbf{X}_{i,j}^t\mathbf{W}_{i,j}\mathbf{X}_{i,j}\right)^{-1}=\Big(U,0,0,V,0,W\Big)$, where

$$U=\frac{S_{4,0}S_{0,4}-S_{2,2}^2}{(S_{4,0}-S_{2,0}^2)(S_{0,4}-S_{0,2}^2)}, \qquad V=\frac{S_{2,2}S_{0,2}-S_{2,0}S_{0,4}}{(S_{4,0}-S_{2,0}^2)(S_{0,4}-S_{0,2}^2)},$$

$$W=\frac{S_{2,2}S_{2,0}-S_{0,2}S_{0,4}}{(S_{4,0}-S_{2,0}^2)(S_{0,4}-S_{0,2}^2)}.$$

Now, it is easy to verify that

$$R_{s,t}=w_1(s-i)w_2(t-j)\left[U+V\left(\frac{s-i}{K}\right)^2+W\left(\frac{t-j}{L}\right)^2\right],$$

so we have explicit expressions for the coefficients appearing in the linear combination defining $\mathrm{CPS}_{i,j}(2)$.

Note that, due to Schwarz's inequality, $U>0$ and $V<0$, $W<0$. Moreover, $\sum_{s,t}R_{s,t}=U+S_{2,0}V+S_{0,2}W=1$, thus, the weights $R_{s,t}$ may be viewed as a bidimensional kernel of order 4. This means that $\mathrm{PS}_{i,j}(2)$, as well as $\mathrm{CPS}_{i,j}(2)$, may produce negative estimates of $P_{i,j}$. This a drawback that we can avoid by using weight functions with more points in their support.

We now consider the constrained local smoother of degree $p=0$. The previous calculations are easily adapted to this case by suppressing the non relevant columns in the matrix $\mathbf{X}_{i,j}$. It is easy to check that for $p=0$, due to the symmetry of the marginal weight functions,

$$\mathrm{PS}_{i,j}(0)=\sum_{s,t}R_{s,t}\overline{P}_{s,t}, \qquad \text{with} \qquad R_{s,t}=w_1(s-i)w_2(t-j).$$

Thus $\mathrm{PS}_{i,j}(0)$ reduces to a smoother based upon the normalized weights. For the constrained polynomial smoother just remember (4) to obtain an explicit representation.

Appendix C. An explicit expression for $\mathrm{CPS}_{i,j}(p)$, $p \leq 2$ (dimension $d \geq 2$)

We will refer to the matricial representation (11), which is independent of the dimension of distribution. Some extra notation is required. Consider a d-dimensional distribution with N_k cells in the k^{th} direction, thus with $N = N_1 \times \cdots \times N_d$ cells. The support of the distribution may be identified with $[0,1]^d$ considering d-dimensional rectangles with midpoints $(x_{i_1}, \ldots, x_{i_d}) = (\frac{i_1-1/2}{N_1}, \ldots, \frac{i_d-1/2}{N_d})$. The general cell is represented by $c = (i_1, \ldots, i_d)$, where $i_k = 1, \ldots, N_k$. The local smoother of degree p is represented as

$$\mathrm{PS}_c(p) = \sum_{c'} R_{c'} \overline{P}_{c'}.$$

Take d symmetric weight functions $w_1, \ldots, w_d$ to construct the product weights similarly as done for the case $d = 2$. The matrix $\mathbf{X}_{i,j}^t \mathbf{W}_{i,j} \mathbf{X}_{i,j}$ has rows and columns indexed by $\alpha = (\alpha_1, \ldots, \alpha_d)$, for $\alpha_k \geq 0$ and $\sum_k \alpha_k \leq 2$. To describe this matrix we define, for each cell c, the sums

$$S_{(\alpha_1,\ldots,\alpha_d)} = \sum_{c'} \frac{(x_{j_1} - x_{i_1})^{\alpha_1}}{N_1^{\alpha_1}} \times \cdots \times \frac{(x_{j_d} - x_{i_d})^{\alpha_d}}{N_d^{\alpha_d}},$$

where $\alpha_1, \ldots, \alpha_d \geq 0$, $\alpha_1 + \cdots + \alpha_d \leq 4$ and $c' = (j_1, \ldots, j_d)$. Due to the symmetry of the weight functions $S_{(\alpha_1,\ldots,\alpha_d)} = 0$ whenever one of the α_k is odd. On the sequel, denote $\mathbf{e_k} = (0, \ldots, 0, 2, 0, \ldots, 0)$, the 2 being on the k^{th} coordinate. The matrix $\mathbf{X}_{i,j}^t \mathbf{W}_{i,j} \mathbf{X}_{i,j}$ has entries defined $S_{(\alpha_1+\beta_1,\ldots,\alpha_d+\beta_d)}$, where $\alpha_k, \beta_k \geq 0$ and $\sum_k \alpha_k, \sum_k \beta_k \leq 2$. Notice further that $S_{(\alpha_1+\beta_1,\ldots,\alpha_d+\beta_d)} = S_{(\alpha_1,\ldots,\alpha_d)} S_{(\beta_1,\ldots,\beta_d)}$. Assume that at least one of the coordinates of $\alpha = (\alpha_1, \ldots, \alpha_d)$ is equal to 1. Then, on the row corresponding to α, the matrix $\mathbf{X}_{i,j}^t \mathbf{W}_{i,j} \mathbf{X}_{i,j}$ only has a nonzero entry on the main diagonal. This allows to identify the nonzero entries on the first row of $\left(\mathbf{X}_{i,j}^t \mathbf{W}_{i,j} \mathbf{X}_{i,j}\right)^{-1}$ as the ones corresponding to the cells $\beta = (\beta_1, \ldots, \beta_d) = (0, \ldots, 0), \mathbf{e}_1, \ldots, \mathbf{e}_d$.

To describe the matrix and the coefficients introduce, for each $k = 1, \ldots, d$, $\gamma_k = \frac{S_{2\mathbf{e}_k}}{S_{\mathbf{e}_k}^2}$, the kurtosis associated to the k^{th} kernel. The explicit calculation of the determinants and inverse matrices may be completely described using these kurtosis coefficients. A careful computation shows that

$$R_{c'}=w_1(j_1-i_1)\cdots w_d(j_d-i_d)\left(A_0+\sum_{k=1}^{d}A_k\frac{(j_k-i_k)^2}{N_k^2}\right),$$

where

$$A_0=\frac{1}{\Delta}\begin{vmatrix} S_{\mathbf{e}_1+\mathbf{e}_1} & \cdots & S_{\mathbf{e}_1+\mathbf{e}_d} \\ \cdots & \cdots & \cdots \\ S_{\mathbf{e}_d+\mathbf{e}_1} & \cdots & S_{\mathbf{e}_d+\mathbf{e}_d} \end{vmatrix},$$

$$A_k=\frac{(-1)^{d+k}}{\Delta}\begin{vmatrix} S_{\mathbf{e}_1} & S_{\mathbf{e}_1+\mathbf{e}_1} & \cdots & S_{\mathbf{e}_1+\mathbf{e}_{k-1}} & S_{\mathbf{e}_1+\mathbf{e}_{k+1}} & \cdots & S_{\mathbf{e}_1+\mathbf{e}_d} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ S_{\mathbf{e}_d} & S_{\mathbf{e}_d+\mathbf{e}_1} & \cdots & S_{\mathbf{e}_d+\mathbf{e}_{k-1}} & S_{\mathbf{e}_d+\mathbf{e}_{k+1}} & \cdots & S_{\mathbf{e}_d+\mathbf{e}_d} \end{vmatrix},$$

$$\Delta=\begin{vmatrix} 1 & S_{\mathbf{e}_1} & S_{\mathbf{e}_2} & \cdots & S_{\mathbf{e}_d} \\ S_{\mathbf{e}_1} & S_{\mathbf{e}_1+\mathbf{e}_1} & S_{\mathbf{e}_1+\mathbf{e}_2} & \cdots & S_{\mathbf{e_1}+\mathbf{e}_d} \\ S_{\mathbf{e}_2} & S_{\mathbf{e}_2+\mathbf{e}_1} & S_{\mathbf{e}_2+\mathbf{e}_2} & \cdots & S_{\mathbf{e_2}+\mathbf{e}_d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ S_{\mathbf{e}_d} & S_{\mathbf{e}_d+\mathbf{e}_1} & S_{\mathbf{e}_d+\mathbf{e}_2} & \cdots & S_{\mathbf{e_d}+\mathbf{e}_d} \end{vmatrix}=S_{\mathbf{e}_1}^2\cdots S_{\mathbf{e}_d}^2\prod_{k=1}^{d}(a_k-1).$$

The explicit evaluation of the determinants gives

$$A_k=\frac{(-1)^d}{S_{\mathbf{e}_k}(a_k-1)},\quad k=1,\ldots,d,$$

and, denoting by Δ_0 the numerator defining A_0, $\Delta_0=S_{\mathbf{e}_1}^2\cdots S_{\mathbf{e}_d}^2\Psi_0$, where

$$\Psi_0=\begin{cases} \Big(a_1(a_1+a_2-2)-(1-a_1)^2\Big) \\ \quad\times\displaystyle\prod_{\substack{i\,\text{odd}\\ 1\leq i<d}}\Big((a_{i-1}+a_i-2)(a_i+a_{i+1}-2)-(1-a_i)^2\Big) & \text{if } d \text{ even} \\[2ex] \Big(a_1(a_1+a_2-2)-(1-a_1)^2\Big) \\ \quad\times\displaystyle\prod_{\substack{i\,\text{odd}\\ 1\leq i<d}}\Big((a_{i-1}+a_i-2)(a_i+a_{i+1}-2)-(1-a_i)^2\Big) & \text{if } d \text{ odd.} \\ \quad\times(a_{d-1}+a_d-2) \end{cases}$$

References

[1] Aerts, M., Augustyns, I. and Janssen, P. (1997), Local polynomial estimation of contingency table cell probabilities, *Statistics* 30, 127–148.

[2] Aerts, M., Augustyns, I. and Janssen, P. (1998), Sparse consistency and smoothing for multinomial data, *Statist. Probab. Letters* 33, 41–48.

[3] Burman, P. (1987), Smoothing sparse contingency tables, *Sankhya, Ser. A* 49, 24–36.

[4] Dong J. and Simonof, J.S. (1995), A geometric combination estimator for d-dimensional ordinal contingency tables, *Ann. Statist.* 23, 1143–1153.

[5] Faddy, M. and Jones, M. (1998), Semiparametric Smoothing for Discrete Data, *Biometrika* 85, 131–138.

[6] Hall, P., Seifert, B. and Turlach, B., (2001), On adaptation to sparse design in bivariate local linear regression, *J. Korean Math. Soc.* 30, 231–246.

[7] Hall, P. and Titterington, D.M. (1987), On smoothing sparse multinomial data, *Austral. J. Statist.* 29, 19–37.

[8] Jacob, P., and Oliveira, P.E. (2007), Penalized smoothing of sparse tables, Preprint 07-02, Dep. Matemática, Universidade de Coimbra.

[9] Jacob, P. and Oliveira, P.E., (2010), Relative smoothing of discrete distributions with sparse observations, *J. Stat. Comput. Simul.*, iFirst http://dx.doi.org/10.1080/00949650903218861.

[10] Schuster, E., (1985), Incorporating support constraints into nonparametric estimators of densities, *Comm. Statist. A—Theory Methods* 14, 1123–1136.

[11] Simonof, J.S. (1983), A penalty function approach to smoothing large sparse contingency tables, *Ann. Statist.* 11, 208–218.

[12] Simonof, J.S. (1995), Smoothing categorical data, *J. Statist. Plann. Inference* 47, 41–69.

[13] Simonof, J.S. (1996), *Smoothing methods in statistics*, Springer-Verlag, New York., 1996.

Pierre Jacob
I3M, Dep. Mathématiques, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, FRANCE

E-mail address: pmjacob@orange.fr

Paulo Eduardo Oliveira
CMUC, Department of Mathematics, University of Coimbra, PORTUGAL

E-mail address: paulo@mat.uc.pt