Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 07–02

PENALIZED SMOOTHING OF SPARSE TABLES

PIERRE JACOB AND PAULO EDUARDO OLIVEIRA

ABSTRACT: In models using categorical data one may use some adjacency relations to justify the use of smoothing to improve upon simple histogram approximations of the probabilities. This is particularly convenient when in presence of a sparse number of observations. Moreover, in many models, the prior knowledge of a marginal distribution is available. We propose two families of polynomial smoothers that incorporate this marginal information into the estimates. Besides, one of the family, the penalized polynomial smoothers, corrects the well known drawback of the polynomial smoothers of producing negative approximations. A simulation study show a good performance of the proposed estimators with respect to usual error criteria. Our estimators, and particularly the penalized family, perform especially well for sparse situations.

KEYWORDS: polynomial smoothing, penalized smoothing, sparse observations.
AMS SUBJECT CLASSIFICATION (2000): 62H12, 62H17.

1. Introduction

Models using categorical data usually assume that there does not exist any relation between adjacent points of the distribution. This is not the case for continuous distributions, where many estimation procedures are based on the fact that observations that fall near the approximation site, 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 estimators or the local polynomial estimator are, in fact, clever ways to use this idea to improve your estimates. In many situations where categorical models are used, adjacency of points does mean some kind of contiguity on the information described. This is often the case when using some scale to categorize the observations. In such situations it becomes natural to use adjacent observations to construct the estimates. This idea has been used in the literature to smooth over discrete

Received January 11, 2007.
First author supported by a grant from Centro de Matemática da Universidade de Coimbra, Fundação para a Ciência e Tecnologia (FCT).
Second author supported by Centro de Matemática da Universidade de Coimbra, Fundação para a Ciência e Tecnologia (FCT) and POCTI.

distributions. This smoothing becomes even more interesting when facing situations where we have few observations when compared with the number of points of the underlying distribution, that is, we have a sparse number of observations. In such cases, the use of the classical point frequency estimator seems inadequate: there would be many points of the distribution support without any observation or with only one observation, thus we would come out with an approximation for the distribution with many zeros and almost uniform on the remaining points. Such an approximation seems quite unintuitive. Convenient smoothing over adjacent points does contribute to improve this handicap of the histogram. With this in mind Simonoff [8] and Hall and Titterington [6] studied estimators that correspond to smooth the histogram with an uniform like distribution, although this was justified in quite different way, while Burman [3] studied a discrete version of the kernel estimator. More recently Simonoff [9, 10], Dong and Simonoff [4] or Aerts, Augustyns and Janssen [1, 2] studied discrete versions of the local polynomial estimator.

Smoothing over sparse observations has been used in the literature concentrating mainly on the asymptotic properties of the estimators. To keep the sparseness of the problem, it was common to assume that quotient between the number of observations and the number of points in the support converge to some finite limit. This assumes that, as we collect more data, we are able to somehow refine the support of the distribution, that is, in presence of a larger number of observations we are able to differentiate observations with increasing precision. With this approach in mind, the asymptotics of the mean sum of squared errors was studied, for example, by Simonoff [8, 9], Hall and Titterington [6], Burman [3], Dong and Simonoff [4] or Aerts, Augustyns and Janssen [1]. Other criteria, trying to adapt sparse situations, was introduced by Simonoff [8] who studied the properties of the smoother he introduced. Later, in Aerts, Augustyns and Janssen [2], the properties of the local polynomial with respect to the sparse error criterion were studied. We note that the assumption of a larger support when the number of observations increasing is quite natural if we assume that there exists some density, whose discretization generates the discrete support.

Our first interest in this kind of problems arose when analyzing data from an anthropological study. The sample size was small, especially when compared with the size of the support. Moreover, the inclusion of new units in the sample was quite expensive, both in time and financially, so there was increased interest on extracting as much information as possible from the (few) available observations. The asymptotic properties were not very helpful in this situation, although the above mentioned references proved the smoothers behave quite effectively. Moreover, the methods that have been shown to have the best asymptotic results, the polynomial smoothers, quite often produce negative estimations for some of the probabilities, and this is obviously unacceptable for the practitioner. This behaviour occurs naturally in regions where the smoothing window find almost no observation at all.

A particular aspect of the anthropological study, that was not at all addressed by the discussed methods, was the fact that a marginal distribution was known. So, we are interested in smoothing over a sparse table and produce an approximation for the two dimensional distribution that does not produce negative values and agrees with given values for one the margins. To our best knowledge, this problem was not addressed in the literature.

Our estimates are obtained as a solution of a minimization problem and have explicit formulations. As we were mainly interested in their finite sample properties we undertook some simulation work. These show a general advantage on the behaviour of the estimators we are defining. We considered the usual error criterion, mean sum of squared errors and the sup-norm. This better behaviour attenuates when the number of observations increases with respect to the support size. We simulated up to a mean of five observations per point of the support. This is already considered a nonsparse situation. The estimators proposed behave still well for each cases.

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 cube $[0,1] \times [0,1]$, considering equally sized squared cells with midpoints $(x_i, y_j) = \left(\frac{i-1/2}{K}, \frac{j-1/2}{L}\right)$, $i = 1, \ldots, K$, $j = 1, \ldots, L$. Then, we might think of $\mathbf{P}$ as the result a discretization of a continuous underlying probability distribution with a density function f on $[0,1]\times[0,1]$: for each $i=1,\ldots,K$, $j=1,\ldots,L$,

$$P_{i,j}=\int_{C_{i,j}} f(x,y)\,d(x,y).$$

The special feature of this paper is that we assume $\mathbf{P}$ to be partially known. More precisely, we assume that the marginal distribution

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

is known.

Given an estimator $\mathbf{P}^*=\left(P^*_{i,j}\right)$, the first error criterion studied was the mean sum of squared errors:

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

We will also compare the performance of the estimators with respect to the sup-norm:

$$\mathrm{NSUP}(\mathbf{P}^*)=\sup_{1\le i\le K,\,1\le j\le L}\left|P^*_{i,j}-P_{i,j}\right|. \tag{2}$$

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 $\mathbf{N}$ by reflecting its 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 cube $[-1,2]\times[-1,2]$, the cells being equally sized squares 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 $K\times L$ square block of the enlarged matrix.

The enlargement of $\mathbf{P}$ is easily described. Let the matrices $\overline{\mathbf{P}}_*$, $\overline{\mathbf{P}}^*$ and $\overline{\mathbf{P}}^*_*$ 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,

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

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

In order 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 symmetrical densities with support included in $[-1/2,\ 1/2]$. Given $h_1,\ h_2>0$, define

$$\mathcal{K}_{\mathbf{H}}(u,v)=\frac{1}{h_1h_2}\mathcal{K}_1\left(\frac{u}{h_1}\right)\mathcal{K}_2\left(\frac{v}{h_2}\right),$$

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

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

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 the matrix $\mathbf{P}$, with the components listed in the lexicographic order.

3. The estimators

In this section we describe the estimators we propose. As mentioned earlier, they will be constructed as solutions of optimization problems. The first family of estimators, denoted CPS, for constrained polynomial smoother, correspond to constrained, because of the marginal distribution being given, polynomial estimators. These estimators will appear as an additive correction of the usual local polynomial estimators. This family of estimators, like the classical local polynomial estimators, may produce negative approximations for some cells. Thus, we propose to optimize another error function that will construct estimators that will be always nonnegative. This family of estimators appears when modifying the error function by penalizing the error in a relative way with respect to true probabilities. This penalizing idea was inspired by the chi-squared tests, where errors are measured relative to their expected values. These estimators will be denoted CPPS, for constrained penalized polynomial smoother.

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

$$H_{i,j} = \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)^t \mathbf{K}_{i,j} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right), \tag{3}$$

where $\beta_{i,j} = (\beta_{0,i,j}, \dots, \beta_{5,i,j})^t$. 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 polynomial smoothers of different degree 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

$$\widehat{\beta}_{i,j} = \left(\mathbf{X}_{i,j}^t \mathbf{K}_{i,j} \mathbf{X}_{i,j}\right)^{-1} \mathbf{X}_{i,j}^t \mathbf{K}_{i,j} \overrightarrow{\mathbf{P}}, \tag{4}$$

which is exactly the usual polynomial estimator of $\beta_{i,j}$. Now, the situation differs near the border. In fact, it is well known that the usual local polynomial estimate gives some automatic correction to border effects at the cost of a somewhat intricate expression for the regression coefficients. The fact that we use the replication device described earlier, that may seem somewhat painful to describe, gives in return the advantage of correcting border effects without strictly modify the general expression of the estimate near the boundary of the table $\mathbf{C}$. 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, $i = 1, \dots, K$. 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:

$$\begin{gathered} \text{minimize} \sum_{l=1}^{L} H_{i,l} \\ \text{subject to } \textstyle\sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i. \end{gathered} \tag{5}$$

If $\widehat{\beta}^c_{i,j}$, $j = 1, \dots, L$, are the minimizers of this problem, the constrained polynomial smoother of degree 2 is $\mathrm{CPS}_{i,j}(2) = \widehat{\beta}^c_{0,i,j}$, the constant term of $\widehat{\beta}^c_{i,j}$. For constrained polynomial smoothers with different degrees, the solution appears by modifying the matrices $\mathbf{X}_{i,j}$ appropriately. In Section 5 we show that

$$\mathrm{CPS}_{i,j}(p) = \mathrm{PS}_{i,j}(p) + \frac{1}{L}\left(\Pi_i - \sum_{l=1}^{L} \mathrm{PS}_{i,l}(p)\right). \tag{6}$$

An explicit formula for $\mathrm{CPS}_{i,j}(2)$ is given in Section 6. Expressions for $\mathrm{CPS}_{i,j}(0)$ and $\mathrm{CPS}_{i,j}(1)$ are given in Section 10.

It is well known that the polynomial smoother $\widehat{\beta}_{i,l}$ tend to produce negative values for its constant term. This means that $\mathrm{PS}_{i,j}(p)$ might be negative, except for the case $p = 0$, when this estimator is the classical Nadaraya-Watson estimator. Unfortunately, the expression above does not always correct this drawback for $\widehat{\beta}^c_{i,j}$.

We now propose a new estimator. For this purpose, define, for each $i = 1, \dots, K$, and $j = 1, \dots, L$, the penalized target functions

$$H^*_{i,j} = \frac{1}{\beta_{0,i,j}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right), \tag{7}$$

where

$$\mathbf{W}_{i,j} = \frac{1}{\sum_{s,t} \mathcal{K}_H(x_s - x_i, y_t - y_j)} \mathbf{K}_{i,j}.$$

This is a penalized error measure, so we will call the estimator derived a penalized polynomial smoother of degree 2. For different degrees, just change the matrices $\mathbf{X}_{i,j}$ conveniently. The idea of considering errors that are relative to the probability we are trying to estimate will contribute to keep the nonnegativity and to some overestimation of very small probabilities. The minimization of (7), for each (i, j), does not necessarily produce a probability distribution, less one that respects the knowledge of the given marginal. Thus, our constrained penalized polynomial smoother, is defined by the solution of the optimization problem:

$$\begin{gathered} \text{minimize } H^*_i = \sum_{l=1}^{L} H^*_{i,l} \\ \text{subject to } \textstyle\sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i. \end{gathered} \tag{8}$$

The constrained penalized polynomial smoother of degree 2, denoted $\mathrm{CPPS}_{i,j}(2)$ is the first coordinate of the minimizer of the above problem. For different degrees of the smoother, we modify the matrices $\mathbf{X}_{i,j}$ accordingly.

To present a matricial expression for $\mathrm{CPPS}_{i,j}$, define the matrices $\widetilde{\mathbf{X}}_{i,j}$ by replacing in $\mathbf{X}_{i,j}$ the first column by zeros. Then, the matrix $\widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,j} \widetilde{\mathbf{X}}_{i,j}$ is of the form

$$\left[\begin{array}{cc} 0 & \cdots\ 0 \\ \vdots & \mathbf{M} \\ 0 & \end{array}\right].$$

Whenever the block $\mathbf{M}$ is nonsingular we shall say the matrix has a generalized inverse and write

$$\left(\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l\right)^{-1} = \left[\begin{array}{cc} 0 & \cdots\ 0 \\ \vdots & \mathbf{M}^{-1} \\ 0 & \end{array}\right].$$

With this definition, put

$$\mathbf{A}_{i,j} = \mathbf{W}_{i,j}^t \widetilde{\mathbf{X}}_{i,j} \left(\widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,j} \widetilde{\mathbf{X}}_{i,j}\right)^{-1} \widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,j}. \tag{9}$$

We show in Section 8 that

$$\mathrm{CPPS}_{i,j}(p) = \Pi_i \frac{\left|\overrightarrow{\mathbf{P}}^t (\mathbf{W}_{i,j} - \mathbf{A}_{i,j}) \overrightarrow{\mathbf{P}}\right|^{1/2}}{\sum_{l=1}^L \left|\overrightarrow{\mathbf{P}}^t (\mathbf{W}_{i,l} - \mathbf{A}_{i,l}) \overrightarrow{\mathbf{P}}\right|^{1/2}}. \tag{10}$$

Notice that the matrix $\mathbf{A}_{i,j}$ depends on the degree of the polynomial smoother. At a first glance one can observe that $\mathrm{CPPS}_{i,j}$ gives positive estimations for each $P_{i,j}$. Moreover, $\mathrm{CPPS}_{i,j}$ may be interpreted as a multiplicative correction of the estimator obtained when minimizing $H_{i,j}^*$ without constraint, in contrast to $\mathrm{CPS}_{i,j}$ which appears as an additive correction of the unconstrained estimator $\mathrm{PS}_{i,j}$. Alternatively, $\mathrm{CPPS}_{i,j}$ can be viewed as a kind of conditional estimator through the obvious formula

$$P_{i,j} = \Pi_i \frac{P_{i,j}}{\sum_{l=1}^L P_{i,l}}.$$

The expression for $\mathrm{CPPS}_{i,j}$ given in (10) uses absolute values. In fact, in general, the sign of $\overrightarrow{\mathbf{P}}^t (\mathbf{W}_{i,j} - \mathbf{A}_{i,j}) \overrightarrow{\mathbf{P}}$ depends on the properties of the kernel. In Section 9 below, it will be shown that, although the previous expression is random, its sign is the same as the sign of the nonrandom expression

$$\mathbf{e}^t \left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right) \mathbf{e} = 1 - \left(\frac{\tau_2^4 \sigma_1^4 + \tau_1^4 \sigma_2^4 + 2\sigma_1^4 \sigma_2^4}{\tau_1^4 \tau_2^4 - \sigma_1^4 \sigma_2^4}\right),$$

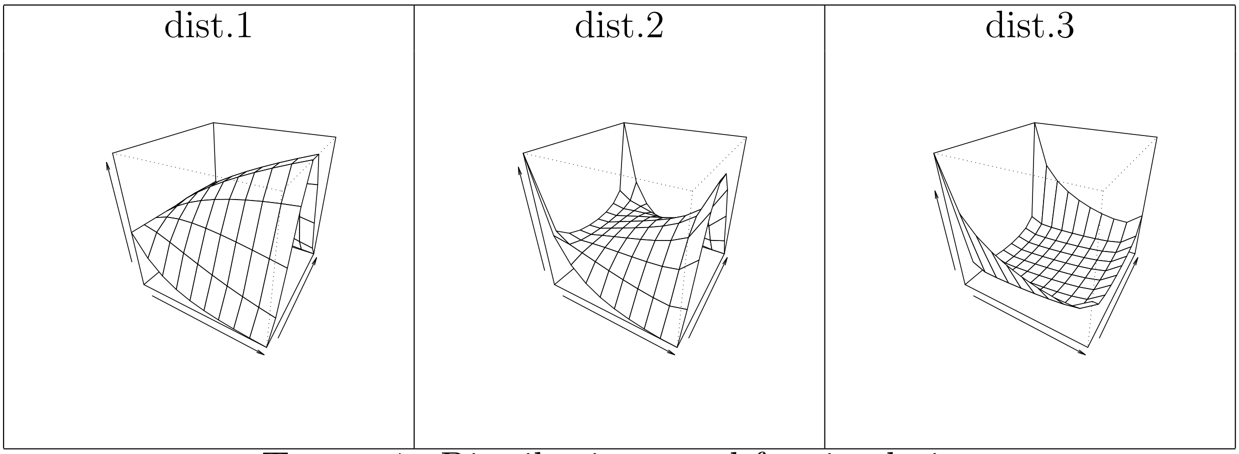


Table 1. Distributions used for simulation.

whenever it is different from zero. This might be, in fact, any real number, and it is independent of the cell (i, j). It is interesting to note that $\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e} = 0$ if we choose the kernels $\mathcal{K}_1$ and $\mathcal{K}_2$ such that their moments verify $\tau_1^4 = 3\sigma_1^4$ and $\tau_2^4 = 3\sigma_2^4$. This holds, for example, if the kernels are gaussian densities. As it will become apparent from the computational details of Sections 8 and 9, it seems convenient that $\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e}$ should be positive. This can be achieved by a suitable choice of the kernel, thus the sign becoming independent of the data set. A simple solution to have this control is to choose kernels that are leptokurtic.

4. 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. We considered three distributions obtained as mixtures of discretized Beta with different parameters. These type of distributions are used in the literature for smoothers over a discrete one dimensional support (see, for example, Aerts, Augustyns and Janssen [1], Simonoff [8, 9] or Dong and Simonoff [4]).

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

	dist.1		dist.2		dist.3	
	$n = N$	$n = 5N$	$n = N$	$n = 5N$	$n = N$	$n = 5N$
$K = 10$, $L = 10$						
CPS(0)	0.003811	0.003860	0.004047	0.000768	0.000815	0.000916
CPS(2)	0.006631	0.006619	0.006859	0.001343	0.001358	0.001414
CPPS(0)	0.001970	0.002063	0.002333	0.000687	0.000685	0.000828
CPPS(1)	0.002109	0.002201	0.002409	0.000691	0.0007012	0.000793
CPPS(2)	0.003671	0.003764	0.003892	0.001318	0.001347	0.001365
$K = 30$, $L = 10$						
CPS(0)	0.001225	0.001251	0.001324	0.000251	0.000256	0.000296
CPS(2)	0.002139	0.002173	0.002249	0.000441	0.000439	0.000461
CPPS(0)	0.000632	0.000640	0.000748	0.000222	0.000216	0.000268
CPPS(1)	0.000676	0.000686	0.000771	0.000223	0.000221	0.000256
CPPS(2)	0.001186	0.001211	0.001257	0.000430	0.000440	0.000446
$K = 30$, $L = 30$						
CPS(0)	0.000454	0.000459	0.000486	0.000091	0.000093	0.000110
CPS(2)	0.000795	0.000797	0.000821	0.000167	0.000165	0.000173
CPPS(0)	0.000245	0.000246	0.000306	0.000075	0.000077	0.000106
CPPS(1)	0.000259	0.000261	0.000311	0.000077	0.000078	0.000099
CPPS(2)	0.000469	0.000463	0.000481	0.000163	0.000163	0.000162

TABLE 2. Simulated values for the MSSE.

discretized over 10×10, 30×10 and 30×30 tables. We performed simulations with the number of observations equal to the number of cells, still considered a sparse situation, and with the number of observations equal to five times de number of cells, already a non sparse situation. We performed some other simulations with fewer observations over the table, that is, in a situation clearly more sparse than those reported. The results were in accordance with the ones that are to be described, so we decided not to include them here. We considered two symmetric and leptokurtic kernels, one in each direction. All the numerical results were obtained by running 500 Monte Carlo samples in each of the considered situations.

In Table 2 we show the simulated values for the MSSE of the estimators. We verify that the constrained penalized polynomial smoothers of degrees 0 and 1, CPPS(0) and CPPS(1) perform better that all the others in every sparse situation, the next best performance is always for the constrained polynomial smoothers of degree 0, CPS(0). The constrained smoothers of degree 2 are consistently the worst ones, although the constrained penalized smoother CPPS(2) exhibits some advantage over the constrained polynomial smoother CPS(2). For distribution 3, this advantage vanishes, due to a very flat distribution. For nonsparse situations, the constrained penalized smoothers continue to exhibit a better performance for small tables. They seem to loose some of their advantage when the table size increases. The relative performance between the estimators is essentially the same as for sparse observations. This loss of performance seems linked to the support of the smoothing kernel. In order to compare results we kept the smoothing kernel over the simulation process. It is clear that the size of the support in the 30×30 tables means that we are looking for smoothing over a quite small neighborhood, while the same kernel looks at a significant neighborhood of the 10×10 table.

Tables 3 and 4 report the results obtained for the simulation of sup-norm. We show the empirical distributions obtained, as they give an easier to read general impression of the performance of each estimator. In Table 3 we show the results for sparse observations, while Table 4 shows results for nonsparse observations. In general, for sparse observations, the constrained penalized smoothers of degrees 0, CPPS(0), shows the better performance, while the constrained polynomial smoother of degree 2, CPS(2), shows the worst performance. In general, the constrained penalized smoothers of degrees 0 and 1, CPPS(0) and CPPS(1), show an equivalent performance, with some advantage to CPPS(0).

As what regards nonsparse observations, from Table 4 we can see that the constrained polynomial and the constrained penalized polynomial of degrees 2, CPS(2) and CPPS(2) show equivalent performance. The constrained penalized smoothers of degrees 0 and 1, CPPS(0) and CPPS(1), and the constrained polynomial smoother of degree 1, CPS(1), also show similar behaviour, although with some advantage to the penalized smoothers. The behaviour is distinct for distribution 3, a rather flat one with strong peaks at the four corners of the table, and a 30×30 table. Here the smoothers of degree 0 show the worst performance, and it is the constrained penalized smoother of degree 2, CPPS(2), that exhibits the best behaviour.

According to these comments, we could indicate a rule of thumb for the choice of the estimator one should use: in general, prefer CPPS(0) or CPPS(1), consider CPS(0) if you decide to smooth with a kernel that has a small support when compared to the size of the table. If you have the information that the distribution you are estimating is almost constant with a few peaks and your data is no longer very sparse, you should also consider CPPS(2).

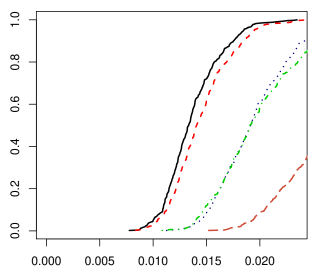
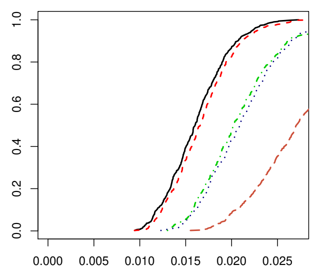
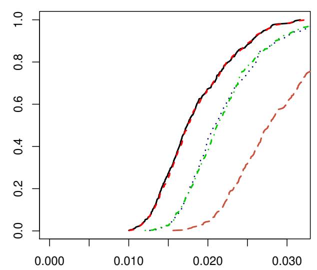
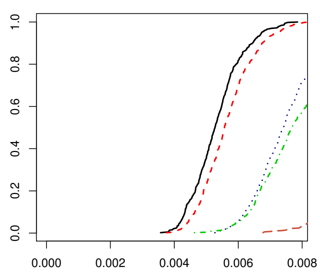
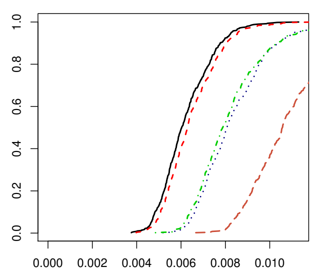
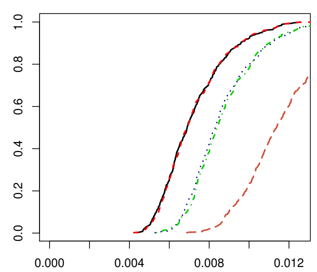
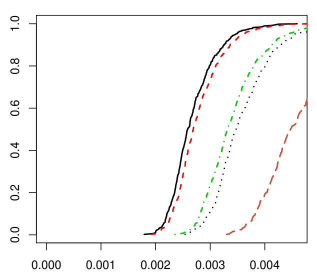
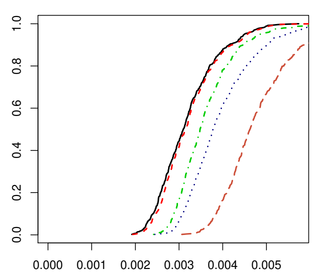
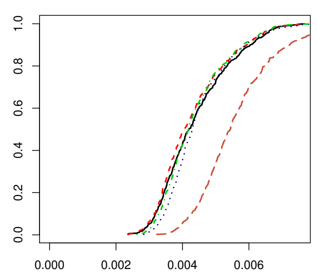
	dist.1	dist.2	dist.3
10×10			
30×10			
30×30			

TABLE 3. Empirical distribution of sup-norm for sparse observations. Smoothers: CPPS(0) (solid), CPPS(1) (dashed), CPPS(2) (dotted), CPS(0) (dotdashed), CPS(2) (longdashed).

5. Derivation of the CPS estimators

In order solve the optimization problem (5), introduce the Lagrange function

$$H_i = \sum_{l=1}^{L} H_{i,l} + 2\nu \left(\sum_{l=1}^{L} \mathbf{h}^t \beta_{i,l} - \Pi_i \right), \tag{11}$$

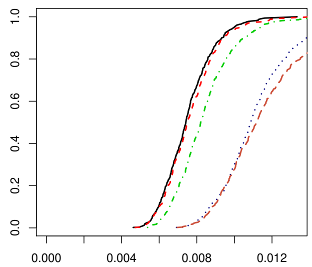
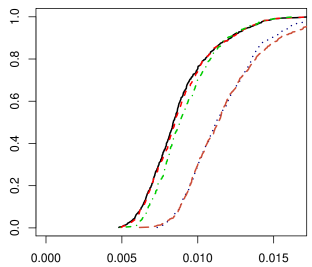
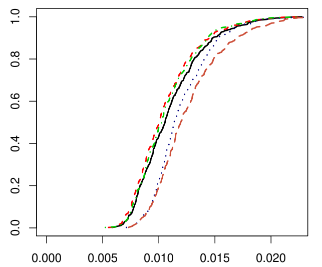
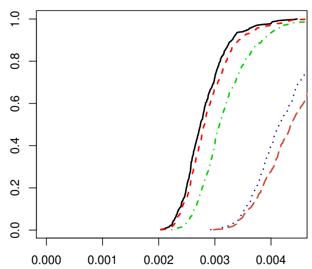
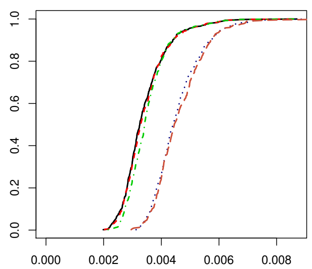
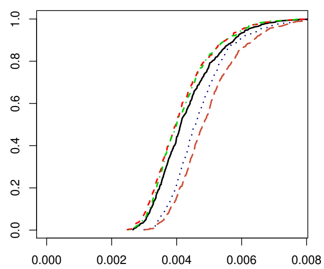
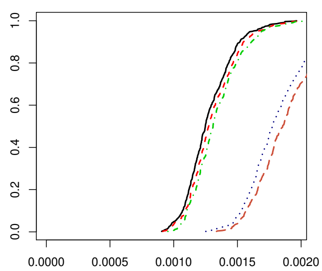
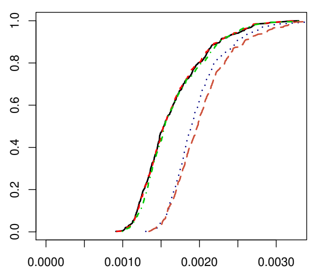
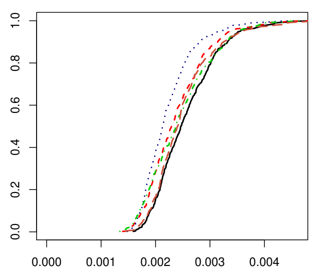
	dist.1	dist.2	dist.3
10×10			
30×10			
30×30			

TABLE 4. Empirical distribution of sup-norm for nonsparse observations. Smoothers: CPPS(0) (solid), CPPS(1) (dashed), CPPS(2) (dotted), CPS(0) (dotdashed), CPS(2) (longdashed).

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

$$\frac{\partial H_i}{\partial \beta_{i,l}} = -2\mathbf{X}_{i,l}^t \mathbf{K}_{i,l} \overrightarrow{\mathbf{P}} + 2\mathbf{X}_{i,l}^t \mathbf{K}_{i,l} \mathbf{X}_{i,l} \beta_{i,l} + 2\nu \mathbf{h} = 0, \qquad l = 1, \ldots, L. \quad (12)$$

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

$$\begin{bmatrix} \mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\mathbf{X}_{i,l} & \mathbf{h} \\ \mathbf{h}^t & 0 \end{bmatrix} \times \begin{bmatrix} \beta_{i,l} \\ \nu \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\overrightarrow{\mathbf{P}} \\ \beta_{0,i,l} \end{bmatrix}. \tag{13}$$

Multiplying on the left by the matrix

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

it follows

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

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

$$\begin{aligned} \nu &= \frac{-\Pi_i + \sum_{l=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\mathbf{X}_{i,l}\right)^{-1}\mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\overrightarrow{\mathbf{P}}}{\sum_{l=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\mathbf{X}_{i,l}\right)^{-1}\mathbf{h}} \\ &= \frac{-\Pi_i + \sum_{l=1}^L \widehat{\beta}_{0,i,l}}{\sum_{l=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\mathbf{X}_{i,l}\right)^{-1}\mathbf{h}}. \end{aligned} \tag{14}$$

Now from (12) we derive

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

Multiplying on the left by $\mathbf{h}^t$ and introducing the expression obtained for ν, it follows

$$\widehat{\beta}_{0,i,j}^c = \widehat{\beta}_{0,i,j} + \frac{\mathbf{h}^t\left(\mathbf{X}_{i,j}^t\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1}\mathbf{h}}{\sum_{l=1}^L \mathbf{h}^t\left(\mathbf{X}_{i,l}^t\mathbf{K}_{i,l}\mathbf{X}_{i,l}\right)^{-1}\mathbf{h}}\left(\Pi_i - \sum_{l=1}^L \widehat{\beta}_{0,i,l}\right). \tag{15}$$

Finally, observe that, under our construction of $\mathbf{X}_{i,l}$ and $\mathbf{W}_{i,l}$, the matrix $\mathbf{h}^t\left(\mathbf{X}_{i,l}^t\mathbf{W}_{i,l}\mathbf{X}_{i,l}\right)^{-1}\mathbf{h}$ does not depend on the index l, and use the notation for the estimators, to find

$$\mathrm{CPS}_{i,j}(p) = \mathrm{PS}_{i,j}(p) + \frac{1}{L}\left(\Pi_i - \sum_{l=1}^{L} \mathrm{PS}_{i,l}(p)\right), \tag{16}$$

that is, expression (6).

6. An explicit expression for $\mathrm{CPS}_{i,j}(2)$

In order to give an explicit formula for the estimator $\mathrm{CPS}_{i,j}(2)$, given by (16), we will find an expression for the polynomial smoother

$$\mathrm{PS}_{i,j}(2) = \mathbf{h}^t \widehat{\beta}^c_{i,j} = \mathbf{h}^t \left(\mathbf{X}^t_{i,j}\mathbf{K}_{i,j}\mathbf{X}_{i,j}\right)^{-1} \mathbf{X}^t_{i,j}\mathbf{K}_{i,j}\overrightarrow{\mathbf{P}}.$$

We start by noting that, with replication device we used, all the cells of he 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 replace, in the expression above, the matrices $\mathbf{K}_{i,j}$ by $\mathbf{W}_{i,j}$. That is, we have the representation

$$\mathrm{PS}_{i,j}(2) = \mathbf{h}^t \left(\mathbf{X}^t_{i,j}\mathbf{W}_{i,j}\mathbf{X}_{i,j}\right)^{-1} \mathbf{X}^t_{i,j}\mathbf{W}_{i,j}\overrightarrow{\mathbf{P}} = \sum_{s,t} R_{s,t}\overline{P}_{s,t},$$

where the coefficients $R_{s,t}$ are to be determined. Moreover, recalling that $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$, it is convenient to introduce the system of product weights

$$p_1(s-i)p_2(t-j) = \frac{\mathcal{K}_{\mathbf{H}}(x_s - x_i, y_t - y_j)}{\sum_{u,v}\mathcal{K}_{\mathbf{H}}(x_s - x_u, y_t - y_v)},$$

and the sums

$$S_{\alpha,\beta} = \sum_{s,t}(x_s - x_i)^\alpha (y_t - y_j)^\beta p_1(s-i)p_2(t-j).$$

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 one of the coefficients α or β is odd. Define now the second and fourth moments of the weight functions p_1 and p_2:

$$\sigma_1^2 = \textstyle\sum_z z^2 p_1(z), \qquad \sigma_2^2 = \sum_z z^2 p_2(z),$$

$$\tau_1^4 = \textstyle\sum_z z^4 p_1(z), \qquad \tau_2^4 = \sum_z z^4 p_2(z).$$

Then, it is 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

$$\mathbf{h}^t \left(\mathbf{X}_{i,j}^t \mathbf{W}_{i,j} \mathbf{X}_{i,j}\right)^{-1} = \begin{bmatrix} U & 0 & 0 & V & 0 & W \end{bmatrix}$$

where

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

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

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

Now, it is easy to verify that

$$R_{s,t} = p_1(s-i)p_2(t-j)\left[U + V\left(\frac{s-i}{K}\right)^2 + W\left(\frac{t-j}{L}\right)^2\right],$$

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 our estimator $\mathrm{CPPS}_{i,j}(2)$.

7. Derivation of the CPPS estimators

In order to solve the minimization problem (8), introduce the Lagrange function

$$H_i^* = \sum_{l=1}^{L} H_{i,l}^* + \lambda \left(\sum_{l=1}^{L} \mathbf{h}^t \beta_{i,l} - \Pi_i \right) \tag{17}$$

where $\mathbf{h} = (1,0,0,0,0,0)^t$, as before, and λ stands for the Lagrange multiplier. The first order conditions $\frac{\partial H_i}{\partial \beta_{i,l}} = 0$, $l = 1, \ldots, L$, give rise, for each l, to the system of six equations

$$\begin{aligned} &\lambda \beta_{0,i,l} \mathbf{h} = \\ &= -2\mathbf{X}_{i,l}^t \mathbf{W}_{i,l} \left(\overrightarrow{\mathbf{P}} - X_{i,l}\beta_{i,l} \right) - \frac{\mathbf{h}}{\beta_{0,i,l}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,l}\beta_{i,l} \right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,l}\beta_{i,l} \right) \end{aligned} \tag{18}$$

We will now analyze this system for fixed l. The first equation is non linear, so it deserves a special treatment. We start by solving the linear part of this system of equations. For this purpose define $\widetilde{\beta}_{i,l} = \beta_{i,l} - \beta_{0,i,l}\mathbf{h}$, and recall that $\widetilde{\mathbf{X}}_{i,j}$ is the matrix obtained by replacing the first column of $\mathbf{X}_{i,j}$ by a column of zeros. Then

$$\mathbf{X}_{i,l}\beta_{i,l} = \widetilde{\mathbf{X}}_{i,j}\widetilde{\beta}_{i,l} + \beta_{0,i,l}\mathbf{e}, \tag{19}$$

where $\mathbf{e}^t = (1, \ldots, 1)$, and the linear part of (18) reduces to

$$\widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,l} \left(\overrightarrow{\mathbf{P}} - \beta_{0,i,l}\mathbf{e} \right) = \widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,l} \widetilde{\mathbf{X}}_{i,l} \widetilde{\beta}_{i,l}.$$

The matrix $\widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,l} \widetilde{\mathbf{X}}_{i,l}$ has a null first line and column, so it is not invertible. Nevertheless, using the generalized invertibility, as described in Section 5, we can write

$$\widetilde{\beta}_{i,l} = \left(\widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,l} \widetilde{\mathbf{X}}_{i,l} \right)^{-1} \widetilde{\mathbf{X}}_{i,j}^t \mathbf{W}_{i,l} \left(\overrightarrow{\mathbf{P}} - \beta_{0,i,l}\mathbf{e} \right). \tag{20}$$

Return to the first (non linear) equation which can be written, after a multiplication of (18) by $\beta_{0,i,l}\mathbf{h}^t$:

$$\left(\overrightarrow{\mathbf{P}} - \widetilde{\mathbf{X}}_{i,j}\widetilde{\beta}_{i,l} + \beta_{0,i,l}\mathbf{e} \right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \widetilde{\mathbf{X}}_{i,j}\widetilde{\beta}_{i,l} - \beta_{0,i,l}\mathbf{e} \right) = \lambda \beta_{0,i,l}^2.$$

Expanding and noting that $\mathbf{e}^t \mathbf{W}_{i,j} \mathbf{e} = 1$, this equation reduces to

$$\left(\overrightarrow{\mathbf{P}} - \widetilde{\mathbf{X}}_{i,j}\widetilde{\beta}_{i,l} \right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \widetilde{\mathbf{X}}_{i,j}\widetilde{\beta}_{i,l} \right) - \beta_{0,i,l}^2 = \lambda \beta_{0,i,l}^2. \tag{21}$$

Now using (20) and a straightforward calculus we find

$$\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \overrightarrow{\mathbf{P}} - \beta_{0,i,l}^2 \mathbf{e}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \mathbf{e} = \lambda \beta_{0,i,l}^2. \tag{22}$$

where $\mathbf{A}_{i,j}$ is defined by (9). Suppose for the moment that $\mathbf{e}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \mathbf{e}$ does not depend upon the index l and denote it by R_i, for shortness. This will be proved later in Section 9. Taking into account that $\beta_{0,i,l}$ should be positive, it follows that

$$\left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \overrightarrow{\mathbf{P}}\right|^{1/2} = \beta_{0,i,l} \left|R_i + \lambda\right|^{1/2},$$

thus, using the constraint, we find

$$\sum_{l=1}^{L} \left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \overrightarrow{\mathbf{P}}\right|^{1/2} = \Pi_i \left|R_i + \lambda\right|^{1/2}.$$

Finally from these two expressions, we derive the following estimate for $P_{i,j}$:

$$\mathrm{CPPS}_{i,j}(2) = \Pi_i \frac{\left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right) \overrightarrow{\mathbf{P}}\right|^{1/2}}{\sum_{l=1}^{L} \left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \overrightarrow{\mathbf{P}}\right|^{1/2}}.$$

It remains to verify that we have indeed identified a minimum, as the objective function is not quadratic. First remark that, as the constraint is linear in $\beta_{i,l}$, the term corresponding to the Lagrange multiplier does not appear in the Hessian matrix of H_i^*. This implies that $\nabla^2 H_i^*$ is a block diagonal matrix with blocks defined by the Hessian matrices of each $H_{i,l}^*$, $l = 1, \ldots, L$. It is now easy to check that, for $\mathbf{u} = (\mathbf{u}_1 | \cdots | \mathbf{u}_L)$ a general $L \times 6$ vector, we have

$$\mathbf{u}^t \nabla^2 H_i^* \mathbf{u} = \sum_{l=1}^{L} \mathbf{u}_l^t \nabla^2 H_i^* \mathbf{u}_l = 2 \sum_{l=1}^{L} \frac{1}{\beta_{0,i,l}^2} \mathbf{Z}_l^t \mathbf{Z}_l \geq 0,$$

where $\mathbf{Z}_l = \frac{u_{0,l}}{\beta_{0,i,l}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,l}\beta_{i,l}\right)^t \mathbf{W}_{i,l}^{1/2} + \mathbf{u}^t \mathbf{X}_{i,l}^t \mathbf{W}_{i,l}^{1/2}$. As the objective function and the constraint are convex the Karush-Kuhn-Tucker conditions ensure we have, in fact, a minimum.

8. An explicit expression for $\mathrm{CPPS}_{i,j}(2)$

According to (10), we need to compute first matrix

$$\mathbf{A}_{i,j} = \mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}\left(\widetilde{\mathbf{X}}_{i,j}^t\mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}\right)^{-1}\widetilde{\mathbf{X}}_{i,j}^t\mathbf{W}_{i,j}.$$

The matrix $\widetilde{\mathbf{X}}_{i,j}^t\mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}$ is obtained from $\mathbf{X}_{i,j}^t\mathbf{W}_{i,j}\mathbf{X}_{i,j}$ by replacing the first line and column of by zeros. Thus, using the notation introduced in Section 6:

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

It follows that $\widetilde{\mathbf{X}}_{i,j}^t\mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}$ is invertible, in the generalized sense introduced before, and

$$\left(\widetilde{\mathbf{X}}_{i,j}^t\mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}\right)^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & S_{2,0}^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & S_{0,2}^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & S_{0,4}\Delta^{-1} & 0 & S_{2,2}\Delta^{-1} \\ 0 & 0 & 0 & 0 & S_{2,2}^{-1} & 0 \\ 0 & 0 & 0 & S_{2,2}\Delta^{-1} & 0 & S_{4,0}\Delta^{-1} \end{bmatrix}.$$

where $\Delta = S_{4,0}S_{0,4} - S_{2,2}^2$.

Next, the typical line of $\mathbf{W}_{i,j}\widetilde{\mathbf{X}}_{i,j}$ is of the form

$$p_1(s-i)p_2(t-j)\left[\begin{array}{cccccc} 0 & \frac{s-i}{K} & \frac{t-j}{L} & \left(\frac{s-i}{K}\right)^2 & \frac{s-i}{K}\frac{t-j}{L} & \left(\frac{t-j}{L}\right)^2 \end{array}\right].$$

Thus, the general entry of $\mathbf{A}_{i,j}$, that is, the entry on line (s,t) and column (s',t') is

$$\begin{aligned}
\mathbf{A}_{i,j}\Big((s,t),(s',t')\Big) &= p_1(s-i)p_2(t-j)p_1(s'-i)p_2(t'-j) \\
\times\Bigg[&\frac{(s-i)(s'-i)}{K^2}\,S_{2,0}^{-1} + \frac{(t-j)(t'-j)}{L^2}\,S_{0,2}^{-1} \\
&+\frac{(s-i)^2(s'-i)^2}{K^4}\,S_{0,4}\Delta^{-1} + \frac{(t-j)^2(t'-j)^2}{L^4}\,S_{4,0}\Delta^{-1} \\
&+\frac{(s-i)(s'-i)}{K^2}\frac{(t-j)(t'-j)}{L^2}\,S_{2,2}^{-1} \\
&+\frac{(s-i)^2(t'-j)^2}{K^2L^2}\,S_{2,2}\Delta^{-1} + \frac{(s'-i)^2(t-j)^2}{K^2L^2}\,S_{2,2}\Delta^{-1}\Bigg]
\end{aligned}$$

So, finally,

$$\begin{aligned}
\overrightarrow{\mathbf{P}}^t\mathbf{A}_{i,j}\overrightarrow{\mathbf{P}} &= \sum_{(s,t),(s',t')} \overline{P}_{s,t}\overline{P}_{s',t'}\mathbf{A}_{i,j}\Big((s,t),(s',t')\Big) \\
&= \frac{1}{\sigma_1^2}\left(\sum_{(s,t)}(s-i)p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2 \\
&+\frac{1}{\sigma_2^2}\left(\sum_{(s,t)}(t-j)p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2 \\
&+\frac{\tau_2^4}{\tau_1^4\tau_2^4-\sigma_1^4\sigma_2^4}\left(\sum_{(s,t)}(s-i)^2p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2 \\
&+\frac{\tau_1^4}{\tau_1^4\tau_2^4-\sigma_1^4\sigma_2^4}\left(\sum_{(s,t)}(t-j)^2p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2 \\
&+\frac{1}{\sigma_1^2\sigma_2^2}\left(\sum_{(s,t)}(s-i)(t-j)p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2 \\
&+\frac{2\sigma_1^2\sigma_2^2}{\tau_1^4\tau_2^4-\sigma_1^4\sigma_2^4}\left(\sum_{(s,t)}(s-i)^2p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)\times \\
&\qquad\times\left(\sum_{(s,t)}(t-j)^2p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right).
\end{aligned}$$

To complete the expression for $\mathrm{CPPS}_{i,j}(2)$, the remaining term is

$$\overrightarrow{\mathbf{P}}^t \mathbf{W}_{i,j} \overrightarrow{\mathbf{P}} = \sum_{(s,t)} p_1(s-i) p_2(t-j) \overline{P}^2_{s,t}. \tag{23}$$

9. The term $\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e}$

In this section we characterize the term $\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e}$. Replacing $\overline{P}_{s,t}$ by 1 in the calculations of the previous section gives,

$$\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e} = 1 - \left(\frac{\tau_2^4\sigma_1^4 + \tau_1^4\sigma_2^4 + 2\sigma_1^4\sigma_2^4}{\delta}\right).$$

where $\delta = \tau_1^4\tau_2^4 - \sigma_1^4\sigma_2^4$. Thus, this expression depends only on the moments of the marginal kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, and its sign depends on these moments. It should be noticed that $\mathbf{e}^t\left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right)\mathbf{e}$ may even be null. This happens, for instance, for weight distributions which satisfy the "gaussian property" $\tau_1^4 - 3\sigma_1^4 = 0$, $\tau_2^4 - 3\sigma_2^4 = 0$.

It is worth noticing that the first order conditions for $H_i^* = \sum_{l=1}^L H_{i,l}^*$, with no constraint, are obtained by setting $\lambda = 0$ in the formulas (17), (18), (21) and (22). In particular, (22) reduces to

$$\overrightarrow{\mathbf{P}}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\overrightarrow{\mathbf{P}} = \beta^2_{0,i,l}\mathbf{e}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\mathbf{e}.$$

It follows then that $\overrightarrow{\mathbf{P}}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\overrightarrow{\mathbf{P}}$ has a constant sign given by the sign of $\mathbf{e}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\mathbf{e}$.

In case the $\mathbf{e}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\mathbf{e} = \mathbf{0}$ there is no solution to the first order equations for $H_i^* = \sum_{l=1}^L H_{i,l}^*$. This means that the unconstrained minimization problem of minimizing has no finite positive solution for $\beta_{0,i,l}$. In such a case, we are unable to precise the sign of $\overrightarrow{\mathbf{P}}^t\left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right)\overrightarrow{\mathbf{P}}$. However the proof developed in Section 7, for the constrained problem, runs even in this case.

10. The smoothers of degree 0 and 1

The previous sections present a detailed description of the smoothers based upon polynomials of degree $p = 2$. Similar estimators are easily produced for degrees $p < 2$, by suppressing the non relevant columns in the matrix $\mathbf{X}_{i,j}$.

It is easy to check that for $p = 0, 1$, due to the symmetry of the marginal kernels,

$$\mathrm{PS}_{i,j}(p) = \sum_{s,t} R_{s,t}\overline{P}_{s,t}$$

with

$$R_{s,t} = p_1(s-i)p_2(t-j).$$

Thus $\mathrm{PS}_{i,j}(p)$ reduces to a smoother based upon the weights of the matrix $\mathbf{W}_{i,j}$. For the constrained polynomial smoother, and we still have

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

For the penalized polynomial smoother, we still have, for $p=1$,

$$\mathrm{CPPS}_{i,j}(1) = \Pi_i \frac{\left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,j} - \mathbf{A}_{i,j}\right) \overrightarrow{\mathbf{P}}\right|^{1/2}}{\sum_{l=1}^{L} \left|\overrightarrow{\mathbf{P}}^t \left(\mathbf{W}_{i,l} - \mathbf{A}_{i,l}\right) \overrightarrow{\mathbf{P}}\right|^{1/2}},$$

with $\overrightarrow{\mathbf{P}}^t \mathbf{W}_{i,j} \overrightarrow{\mathbf{P}}$ given by (23), while $\overrightarrow{\mathbf{P}}^t \mathbf{A}_{i,l} \overrightarrow{\mathbf{P}}$ reduces to

$$\frac{1}{\sigma_1^2}\left(\sum_{(s,t)} p_1(s-i)p_2(t-j)(s-i)\overline{P}_{s,t}\right)^2 + \frac{1}{\sigma_2^2}\left(\sum_{(s,t)} (t-j)p_1(s-i)p_2(t-j)\overline{P}_{s,t}\right)^2.$$

Finally, for $p=0$, a direct computation gives, as the matrices $\mathbf{A}_{i,j}$ become null,

$$\mathrm{CPPS}_{i,j}(0) = \Pi_i \frac{\left|\overrightarrow{\mathbf{P}}^t \mathbf{W}_{i,j} \overrightarrow{\mathbf{P}}\right|^{1/2}}{\sum_{l=1}^{L} \left|\overrightarrow{\mathbf{P}}^t \mathbf{W}_{i,l} \overrightarrow{\mathbf{P}}\right|^{1/2}}.$$

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. (1997), 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] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.

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

[7] Ruppert, D. and Wand, M.P. (1994), Multivariate locally weighted least squares regression, *Ann. Statist.* 22, 1346–1370.

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

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

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

Pierre Jacob
I3M, cc 051, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Paulo Eduardo Oliveira
Dep. Matemática, Univ. Coimbra, Apartado 3008, 3001 - 454 Coimbra, Portugal

E-mail address: paulo@mat.uc.pt
URL: http://www.mat.uc.pt