Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 06–28

PENALIZED SMOOTHING OF DISCRETE DISTRIBUTIONS WITH SPARSE OBSERVATIONS

PIERRE JACOB AND PAULO EDUARDO OLIVEIRA

ABSTRACT: It happens quite often that we are faced with a sparse number of observations over a finite number of cells and we are interested in the estimation of the cell probabilities. The simple histogram produces approximations with the zero value for too many cells. Some polynomial smoothers have been proposed to circumvent this problem which show good properties in the analysis of such sparse situations but have the drawback of producing negative values. We propose a penalized polynomial smoothing for this problem. The estimators that are proposed in this paper are always positive and a simulation study show a very good behaviour with respect to the natural error criterias: mean squared sum of errors, sparse sup and the sup-norm. Our estimator perform specially well for sparse observations. Nevertheless, when the number of observations increases the proposed estimators still show good performance.

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

1. Introduction

For discrete distributions the idea of smoothing the frequency estimates using information concerning the observations in adjacent points may seem strange when we think on models for categorical data. However, it is not rare that the categorizations used take into account some contiguity properties on the underlying scales. In such models this use of adjacent information becomes somewhat natural. This procedure seems even more advantageous when we have large supports and a reduced number of observations, that a sparse table, from which to construct the approximations. The use of the classical frequency cell estimator would provide a zero approximation for too many points of the support which is, in many models, quite unintuitive. Smoothing conveniently over adjacent or contiguous points of the support does contribute to improve the handicap on the histogram.

Received June 7, 2006.
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.

Existing literature on smoothing over discrete distributions concentrate mainly on asymptotic properties of the proposed estimators, even when considering sparse tables. Results on the asymptotic behaviour of the mean sum of squares of the errors are studied, for example, in Simonof [8], Burman [3], Hall and Titterington [6], Simonof [9], Dong and Simonof [4] or Aerts, Augustyns and Janssen [1]. Another error criteria, trying to adapt to sparse situations, was introduced by Simonof (see [10] for example) who proved the first asymptotic results for kernel estimators. The asymptotics of the local polynomials smoothers were studied by Aerts, Augustyns and Janssen [2], establishing sufficient conditions this sparse consistency. The framework for asymptotic results always supposed that the support increases with the number of observations in such a manner that their quotient is convergent to some positive constant.

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. Moreover, the methods that have been shown to have the best asymptotic results (the polynomial smoothers) quite often produce negative estimations for the probabilities, and this is obviously unacceptable for the practitioner. The correct treatment of our original problem takes place in the domain of sparse tables. However, we found methods of estimation that show interesting behaviour also in the case of one dimensional discrete distributions.

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 criteria, mean sum of squared errors and the sparse consistency criteria introduced by Simonof, and also the sup-norm. For the underlying distributions usually considered in the literature, the empirical behaviour of the new estimators defined in this paper was almost always clearly better, for sparse observations, than the polynomial smoothers or the Nadaraya-Watson estimator. Besides, our estimators always produce nonnegative approximations for the probabilities. Our estimators tend to behave in a somewhat less favorable way when the underlying distributions may be too close to zero. But in such cases one can not expect reasonable approximations when dealing with sparse observations.

Let us now define in more detail our framework. Consider k cells $C_1, \ldots, C_k$ and the vector $\mathbf{P} = (P_1, \ldots, P_k)^t$ of the cell probabilities. The observation counts over each cell are described by a multinomial vector $\mathbf{N} = (N_1, \ldots, N_k)$ of size n. The straightforward estimator of $\mathbf{P}$ is the cell frequency vector

$$\overline{\mathbf{P}} = (\overline{P}_1, \ldots, \overline{P}_k) = \left(\frac{N_1}{n}, \ldots, \frac{N_k}{n}\right). \tag{1}$$

Better estimates may be produced by smoothing. An extra justification for smoothing, besides the arguments produced above, is that we can always think of $\mathbf{P}$ as the result of a discretization of a continuous underlying probability distribution. If this underlying distribution has $[0,1]$ support and density function f, each cell C_l may be interpreted as an interval $[(l-1)/k, l/k]$ and each P_l may be expressed as

$$P_l = \int_{(l-1)/k}^{l/k} f(t)\, dt.$$

The idea of smoothing such discretized distributions makes sense, as already argued in the older references cited above: Simonof [8], Burman [3], Hall and Titterington [6]. We refer the reader to Simonof [10] for a quite complete account of these arguments and earlier achievements on the estimation of $\mathbf{P}$.

Given an estimator $\mathbf{P}^* = (P_1^*, \ldots, P_k^*)^t$ the first error criteria studied was the mean sum of squared errors:

$$\mathrm{MSSE}(\mathbf{P}^*) = \mathrm{E}\left(\sum_{i=1}^{k} (P_i^* - P_i)^2\right). \tag{2}$$

More recently the error criteria introduced by Simonof attracted some interest:

$$\mathrm{SPSUP}(\mathbf{P}^*) = \sup_{1 \le i \le k} \left| \frac{P_i^*}{P_i} - 1 \right|. \tag{3}$$

An estimator $\mathbf{P}^*$ is said to be sparse consistent if this error criteria converges almost surely to 0. It is well known that the frequency estimator is not sparse consistent and the same holds for some Bayes estimators of the P_l (see Aerts, Augustyns and Janssen [1] and Simonof [9] for counter-examples).

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

$$\mathrm{NINF}(\mathbf{P}^*) = \sup_{1\leq i\leq k} |P_i^* - P_i| . \tag{4}$$

We stress again that we are not seeking asymptotic results for the estimators but are interested on their behaviour for finite samples with sparse observations.

2. The estimators

We identify each cell C_l with the interval $[(l-1)/k, l/k]$, $l = 1, \ldots, k$, so the center of each cell C_l is the point $x_l = \frac{2l-1}{2k}$. Let $p \geq 0$ be an integer identifying the degree of the polynomial used for the smoothing. For each $l = 1, \ldots, k$, we define the $k \times (p+1)$ matrix

$$\mathbf{X}_l = \begin{bmatrix} 1 & x_1 - x_l & \cdots & (x_1 - x_l)^p \\ \cdots\cdots & & \cdots\cdots & \cdots\cdots \\ 1 & x_i - x_l & \cdots & (x_i - x_l)^p \\ \cdots\cdots & & \cdots\cdots & \cdots\cdots \\ 1 & x_k - x_l & \cdots & (x_k - x_l)^p \end{bmatrix}, \tag{5}$$

and the $k \times k$ matrix

$$\mathbf{K}_l = \mathrm{diag}\left(\frac{1}{h}K\left(\frac{x_1 - x_l}{h}\right), \ldots, \frac{1}{h}K\left(\frac{x_k - x_l}{h}\right)\right), \tag{6}$$

where K is a symmetrical density function with bounded support and $h > 0$. Let $\beta_l = (\beta_{0,l}, \ldots, \beta_{p,l})^t$ and $\widehat{\beta}_l$ the minimizer of

$$(\overline{\mathbf{P}} - \mathbf{X}_l\beta)^t\mathbf{K}_l(\overline{\mathbf{P}} - \mathbf{X}_l\beta) = \\ = \frac{1}{h}\sum_{i=1}^{k}\left(\frac{N_i}{n} - \beta_{0,l} - \beta_{1,l}(x_i - x_l) + \cdots + \beta_{p,l}(x_i - x_l)^p\right)^2 K\left(\frac{x_i - x_l}{h}\right). \tag{7}$$

The estimator for P_l is the constant term $\widehat{\beta}_{0,l}$ of $\widehat{\beta}_l$. This is the polynomial smoother of degree p used throughout the literature and we will denote it by $\mathrm{PS}(p) = (\mathrm{PS}_1(p), \ldots, \mathrm{PS}_k(p))$. The estimator of the l^{th}-coordinate $\mathrm{PS}_l(p)$ is representable in the form $\sum_{j=1}^k s_{lj}\overline{P}_j$, where the s_{lj} depend on the degree p of the polynomial and the positions of the cells C_l, C_j. We refer to Aers, Augustyns and Janssen [2] for an explicit expression. For a more general framework, where the design points are random, explicit expressions may be found in Fan and Gijbels [5] or Ruppert and Wand [7]. As polynomial smoothers, these estimators automatically correct border effects at the cost of somewhat more intricate expression for the coefficients s_{lj} for the cells C_j near the borders. The classification of a cell being near the border or on the interior depends on the support of the kernel K. As we shall explain next, the expression of the coefficients is quite simplified for interior cells, so we will start by a precise definition of interior cell.

We say that a cell C_l is an interior cell if

$$K\left(\frac{x-x_l}{h}\right)=0, \qquad x\notin[0,1],$$

and that C_l is a border cell if the support of $K\left(\frac{\bullet-x_l}{h}\right)$ is not a subset of $[0,1]$.

For an interior cell C_l, as the kernel K is supposed symmetrical, the distribution of the weights s_{lj} is symmetrical with respect to x_l. This is the source of the simplifications referred above. For polynomial smoothers of degree, PS(0) and PS(1), and for interior cells, the symmetry implies that the minimizer is the Nadaraya-Watson estimator, representable on the form

$$\mathrm{NW}_l=\mathrm{PS}_l(0)=\mathrm{PS}_l(1)=\sum_{j=-u}^{u}\overline{P}_{l-j}p(j), \tag{8}$$

where

$$p(j)=\frac{K\left(\frac{j}{kh}\right)}{\sum_{i=-u}^{u}K\left(\frac{i}{kh}\right)}, \qquad j=-u,\ldots,u.$$

For the polynomial smoothers of degree $p=2,3$ and again for interior cells, the representation of the estimator is of the same form, just changing the weights. We have then

$$\mathrm{PS}_l(2)=\mathrm{PS}_l(3)=\sum_{j=-u}^{u}\overline{P}_{l-j}q(j), \tag{9}$$

where

$$q(j)=\frac{\tau^4-\sigma^2j^2}{\tau^4-\sigma^4}\,p(j), \qquad j=-u,\ldots,u, \tag{10}$$

and σ^2, τ^4 are the second and fourth order moments of the weight distribution $p(j)$, $j=-u,\ldots,u$:

$$\sigma^2=\sum_{j=-u}^{u}j^2p(j) \qquad \text{and} \qquad \tau^4=\sum_{j=-u}^{u}j^4p(j). \tag{11}$$

Thus, the polynomial smoothers of orders $p = 2, 3$ appear as kernel estimates associated to a redefined weight function $q(\cdot)$. This new weight function is still symmetric and it is easy to verify that

$$\sum_{j=-u}^{u} j^2 q(j) = 0,$$

so this redefined weight function is a 4^{th} order kernel. This means that this weight function $q(\cdot)$ assumes negative values, so the estimator PPS(2) may be negative for some cells C_l. As we are trying to estimate probabilities, this is a quite inconvenient property.

We stress that the above expressions apply only to interior cells. In this paper we tried to avoid boundary modifications in comparing the estimators described above and the ones we will introduce later. So, instead of considering the expressions of the estimators as referred in Aerts, Augustyns and Janssen [2], we used the well-known replication device of introducing fictitious cells C_l, $l = 1-k, \dots, 0$, to the left of the initial cells, and C_l, $l = k+1, \dots, 2k$, to the right. The frequency in each of these new cells is the one observed on the real cell which is symmetrically situated with respect to the origin or to the last original cell, respectively.

C_{1-k}	$\dots$	C_0	C_1	$\dots$	C_k	C_{k+1}	$\dots$	C_{2k}
$\overline{P}_k$	$\dots$	$\overline{P}_1$	$\overline{P}_1$	$\dots$	$\overline{P}_k$	$\overline{P}_k$	$\dots$	$\overline{P}_1$

That is, for the new cells we define

$$\overline{P}_{1-j} = \overline{P}_{2k+1-j} = \overline{P}_j, \qquad j = 1, \dots, k.$$

For this enlarged support we must redefine the matrices $\mathbf{X}_l$ and $\mathbf{K}_l$ allowing the indexes to range from $1-k$ to $2k$. That is, $\mathbf{X}_l$ is now $(3k) \times (p+1)$ matrix with entries

$$\mathbf{X}_l = \begin{bmatrix} 1 & x_{1-k} - x_l & \cdots & (x_{1-k} - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 1 & x_0 - x_l & \cdots & (x_0 - x_l)^p \\ 1 & x_1 - x_l & \cdots & (x_1 - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 1 & x_k - x_l & \cdots & (x_k - x_l)^p \\ 1 & x_{k+1} - x_l & \cdots & (x_{k+1} - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 1 & x_{2k} - x_l & \cdots & (x_{2k} - x_l)^p \end{bmatrix}, \qquad (12)$$

and $\mathbf{K}_l$ becomes the $(3k) \times (3k)$ matrix

$$\mathbf{K}_l = \operatorname{diag}\left(\frac{1}{h}K\left(\frac{x_{1-k}-x_l}{h}\right), \dots, \frac{1}{h}K\left(\frac{x_{2k}-x_l}{h}\right)\right). \tag{13}$$

With this device, all the original cells become interior cells, so the formulas (8) and (9) apply to each one of them. In order to introduce our estimators it is convenient to define the $(3k) \times (3k)$ matrix

$$\mathbf{W}_l = \operatorname{diag}\left(\frac{K\left(\frac{x_j-x_l}{h}\right)}{\sum_{i=1}^{k} K\left(\frac{x_i-x_l}{h}\right)},\ j = 1-k, \dots, 2k\right).$$

Notice that the denominator in the entries of $\mathbf{W}_l$ depends only on l. This matrix describes the symmetric weights $p(\cdot)$ recentered at the cell C_l. Note further that the diagonal entries are zero whenever $|j-l| > u$.

Our goal is to achieve more precise approximations for small probabilities and, at the same time, guarantee their positivity by minimizing expressions such as

$$H_l = \frac{1}{\beta_{0,l}}(\overline{\mathbf{P}} - \mathbf{X}_l\beta)^t\mathbf{W}_l(\overline{\mathbf{P}} - \mathbf{X}_l\beta). \tag{14}$$

This is a penalized error criteria, so we will call the estimator that we will derive a penalized polynomial smoother of degree p, denoted by $\mathrm{PPS}(p) = (\mathrm{PPS}_1(p), \dots, \mathrm{PPS}_k(p))$. The idea of considering errors that are relative to probability we are trying to estimate will contribute to keep the nonnegativity and to some overestimation of very small probabilities. This error function is, apart from the relativization of the error, the same as considered for the polynomial smoothers (7). However, the minimization of (14), for each $l = 1, \dots, k$, does not necessarily produce a probability distribution over the initial cells. So we introduce this as a global condition and minimize the sum of the errors H_l. Thus our estimator appears as the solution of the optimization problem

$$\begin{gathered}\text{minimize } \sum_{l=1}^{k} H_l \\ \text{subject to } \textstyle\sum_{l=1}^{k} \beta_{0,l} = 1.\end{gathered}$$

Introducing the Lagrange multiplier, we need to minimize

$$H = \sum_{l=1}^{k} H_l + \lambda\left(\sum_{l=1}^{k} \mathbf{h}^t\beta_l - 1\right), \tag{15}$$

where $\mathbf{h} = (1, 0, \ldots, 0)^t$ is a $(p+1)$-dimensional vector. We defer the computational details Appendix A. The estimator is, as before, the first coordinate of the minimizer of (15) and may expressed as

$$\mathrm{PPS}_l(p) = \frac{\left(\overline{\mathbf{P}}^t(\mathbf{W}_l - \mathbf{A}_l)\overline{\mathbf{P}}\right)^{1/2}}{\sum_{j=1}^k \left(\overline{\mathbf{P}}^t(\mathbf{W}_l - \mathbf{A}_l)\overline{\mathbf{P}}\right)^{1/2}}, \qquad l = 1, \ldots, k, \tag{16}$$

where $\mathbf{A}_l$ is a matrix to be described in the next section. In order to give explicit expression we need, besides the second and fourth moments of the weight function introduced before (11), the sixth moment

$$\gamma^6 = \sum_{j=-u}^{u} j^6 p(j),$$

and

$$m_t = \sum_{j=-u}^{u} \overline{P}_{l-j} j^t p(j), \qquad t = 0, 1, 2, \ldots. \tag{17}$$

Now we are able to produce explicit formulas for each value of the polynomial degree p.

- $p = 0$: $\mathbf{A}_l = 0$, thus $\overline{\mathbf{P}}^t(\mathbf{W}_l - \mathbf{A}_l)\overline{\mathbf{P}} = \overline{\mathbf{P}}^t\mathbf{W}_l\overline{\mathbf{P}} = \sum_{j=-u}^{u} \overline{P}_{j-l}^2 p(j)$;
- $p = 1$: $\overline{\mathbf{P}}^t\mathbf{A}_l\overline{\mathbf{P}} = \frac{1}{\sigma^2} m_1^2$;
- $p = 2$: $\overline{\mathbf{P}}^t\mathbf{A}_l\overline{\mathbf{P}} = \frac{1}{\sigma^2} m_1^2 + \frac{1}{\tau^4} m_2^2$;
- $p = 3$:
$$\overline{\mathbf{P}}^t\mathbf{A}_l\overline{\mathbf{P}} = \frac{\gamma^6}{\sigma^2\gamma^6 - \tau^8} m_1^2 - \frac{2\tau^4}{\sigma^2\gamma^6 - \tau^8} m_1 m_3 + \frac{\gamma^6}{\sigma^2\gamma^6 - \tau^8} m_3^2 + \frac{1}{\tau^4} m_2^2.$$

As it will be evident when analyzing simulation results these penalized smoothers perform well when the probabilities to be estimated are not to close to zero. When the discrete distribution has cells with probabilities of the order of 10^{-4} the sparse error criteria SPSUP and NINF start showing a worse empirical behaviour than the polynomial smoothers NW or PS(2). This seems to be due to an overestimation of these probabilities. Having in mind distributions with cell probabilities of reasonable magnitude we considered a quadratic penalized polynomial smoother obtained by optimizing

$$H_l^* = \frac{1}{\beta_{0,l}^2}\,(\overline{\mathbf{P}} - \mathbf{X}_l\beta)^t\mathbf{W}_l(\overline{\mathbf{P}} - \mathbf{X}_l\beta). \tag{18}$$

The estimator, that will be denoted $\mathrm{P2PS}(p) = (\mathrm{P2PS}_1(p), \ldots, \mathrm{P2PS}_k(p))$, should thus be the minimizer of

$$H^* = \sum_{l=1}^{k} H_l^*. \tag{19}$$

As before, the solution of this minimization problem need not to be a probability distribution. Considering the constraint $\sum_{l=1}^{k}\beta_{0,l} = 1$ leads to a complex solution, so we propose another method to make the estimator explicit. Regarding the expressions for $\mathrm{PPS}(p)$ we notice that the penalized smoother is just the unconstrained solution of the minimization problem normalized by the sum of the values corresponding to each cell. So we propose the estimator obtained by the same normalizing procedure applied to the minimizer of (19). This leads to a simple explicit expression:

$$\mathrm{P2PS}_l(p) = \frac{\frac{\overline{\mathbf{P}}^t(\mathbf{W}_1-\mathbf{A}_l)\overline{\mathbf{P}}}{\overline{\mathbf{P}}^t(\mathbf{W}_1-\mathbf{A}_l)\mathbf{e}}}{\sum_{l=1}^{k}\frac{\overline{\mathbf{P}}^t(\mathbf{W}_1-\mathbf{A}_l)\overline{\mathbf{P}}}{\overline{\mathbf{P}}^t(\mathbf{W}_1-\mathbf{A}_l)\mathbf{e}}}\,, \tag{20}$$

where $\mathbf{e} = (1, \ldots, 1)^t$.

3. Simulation results

In this section we compare the performance of the polynomial smoothers, penalized and not penalized, under different error criteria. We consider distributions that have been used in the literature for simulation purposes, say discretizations over k equally spaced points based on Beta distributions with parameters (3,3) and (.6,.6). We decided to include another family of distributions based on the discretization of the density function $g_1(t) = .4{\times}2g(2(t+.5)) + .4{\times}g(t) + .2{\times}2g(2(t-.5))$, where $g(t) = \cos^2(\frac{t\pi}{2})$, if $t \in [-1, 1]$ and $g(t) = 0$ elsewhere, which is smoother on the boundary. We simulated the behaviour of each estimator for discrete distributions with $k = 50$, 100, 500 cells in their support and considering $n = k/2$ observations, that is a sparse situation, or $n = 2k$ observations, a not so sparse situation, to judge of the performance when the size of the sample increases (with a fixed support for the distribution). We considered a weight function $p(\cdot)$ with a five point support, that is, with $u = 2$, referring to the notation used above.

We stress that in the literature, authors were mainly concerned with the asymptotic behaviour when both the number of cells of the distribution and the number of observations were converging to infinity, maintaining some sort of relation implying that the number of cells in the support would not become to small with respect to the size of the sample. In this paper we are mainly interested in the behaviour when the discrete distribution is fixed and trying to find estimators that have good performance when there are few observations. We have no objective quantification of what *few* means, so we found instructive to collect data about the possibilities described above. All the numerical results included were obtained from the result of 500 simulated samples in each of the considered situations.

In Table 1 we find the simulated values for the MSSE of the estimators considered. We may verify that the penalized polynomial smoothers and the quadratic penalized smoother perform better than the polynomial smoothers in every situation. The quadratic penalized smoother P2PS(0) performs specially well in sparse situations ($n = k/2$), regardless of the size or type of distribution. Among the penalized polynomial smoothers the one of degree $p = 0$, that we might think of as penalized Nadaraya-Watson estimator, is always the best of the penalized smoothers, as what the MSSE regards. Because of this behaviour we decided not to include simulations results for quadratic penalized smoothers of degree geater than 0. The performance of the penalized or quadratic penalized polynomial smoothers P2PS(0), PPS(0) or PPS(1) are specially better for really sparse observations: their MSSE's are clearly smaller for every support size of the distribution when $n = k/2$. When the sample size increases the performance of P2PS(0), PPS(0) and PPS(1), although remaining better, is similar to the one of the Nadaraya-Watson estimator. This seems to indicate that the penalized or quadratic penalized smoothing are particularly well suited for sparse situations. In fact, the penalized or quadratic penalized smoothers show a better MSSE in every simulation. The MSSE for nonsparse situations is of same order for the Nadaraya-Watson estimator and for the penalized or quadratic penalized smoothers of degrees $p = 0, 1$, although with some advantage for these last estimators. For this error criteria the polynomial smoother of degree 2, PS(2), is consistently the worse of the considered estimators.

	$\beta(3,3)$		$\beta(.6,.6)$		$g_1(t)$	
	$n=k/2$	$n=2k$	$n=k/2$	$n=2k$	$n=k/2$	$n=2k$
$k=50$						
NW	0.010395	0.002601	0.012165	0.003292	0.010357	0.002646
PS(2)	0.020110	0.005051	0.021735	0.005418	0.020154	0.005062
PPS(0)	0.006987	0.002228	0.008564	0.002985	0.006913	0.002245
PPS(1)	0.007455	0.002286	0.008921	0.002968	0.007407	0.002308
PPS(2)	0.010280	0.003879	0.011093	0.004109	0.010072	0.003821
P2PS(0)	0.006226	0.002579	0.007590	0.003252	0.005939	0.002597
$k=100$						
NW	0.005484	0.001377	0.006119	0.001645	0.005481	0.001365
PS(2)	0.010381	0.002598	0.010935	0.002714	0.010328	0.002598
PPS(0)	0.003702	0.001187	0.004349	0.001473	0.003650	0.001158
PPS(1)	0.003952	0.001218	0.004524	0.001473	0.003905	0.001194
PPS(2)	0.005412	0.002014	0.005643	0.002068	0.005265	0.001976
P2PS(0)	0.003214	0.001334	0.003966	0.001601	0.003066	0.001267
$k=500$						
NW	0.001133	0.000285	0.001184	0.000316	0.001145	0.000284
PS(2)	0.002106	0.000529	0.002144	0.000542	0.002123	0.000528
PPS(0)	0.000763	0.000244	0.000845	0.000279	0.000763	0.000240
PPS(1)	0.000814	0.000251	0.000881	0.000282	0.000814	0.000248
PPS(2)	0.001102	0.000411	0.001121	0.000412	0.001085	0.000403
P2PS(0)	0.000661	0.000267	0.000808	0.000306	0.000638	0.000258

TABLE 1. MSSE for the polynomial, penalized polynomial and quadratic penalized smoothers.

In Tables 2, 3 and 4 we describe the empirical distribution function of the simulated values for SPSUP. These graphs give a better impression of the overall behaviour of this error criteria. Generally speaking, the faster the curve grows the better the associated estimator is: this means that larger values for the SPSUP criteria appear with less frequency. The penalized and quadratic penalized smoothers always show their better performance for sparse situations ($n = k/2$). For the distributions obtained by discretizing $\beta(3, 3)$ or $g_1(\cdot)$ densities the cell probabilities may be to close to zero. As already mentioned, this is responsible for a poorer performance of the penalized estimators as these tend to overestimate probabilities that are to small. Then the presence of the very small true probability p_l in the denominator produces the results described. Looking at the results obtained for the discretization

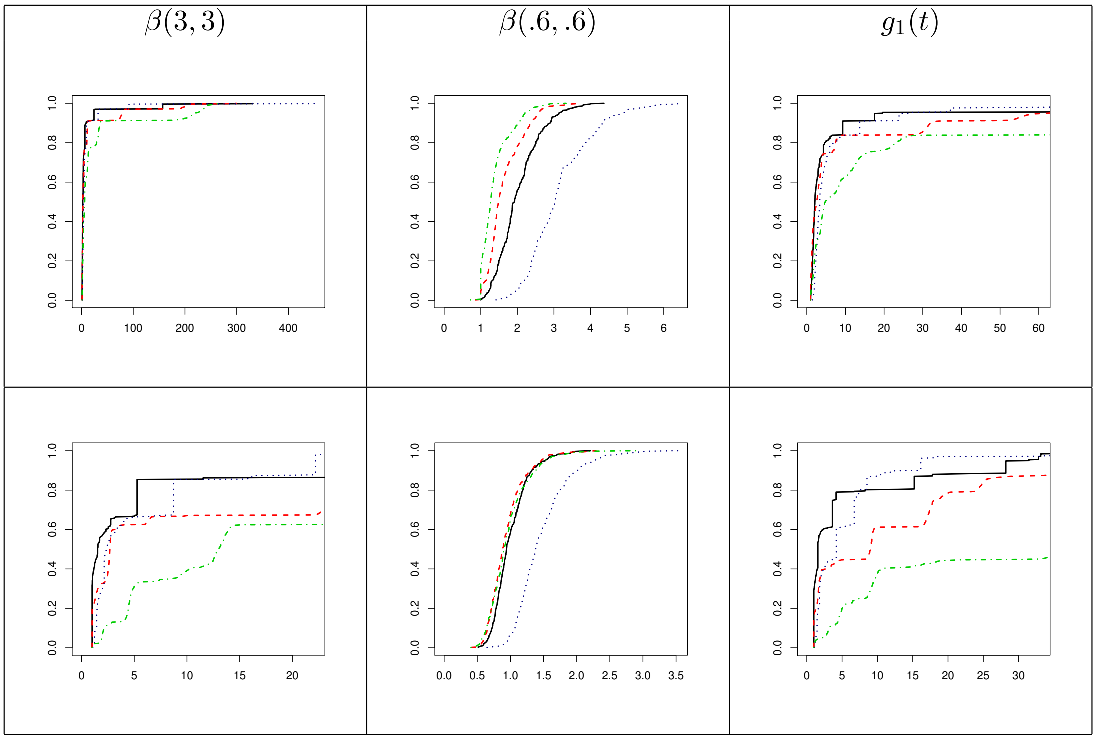

TABLE 2. Empirical distribution of SPSUP ($k = 50$, first line $n = 25$, second line $n = 100$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

of the $\beta(.6,.6)$ density we observe that the penalized smoothers P2PS(0) and PPS(0) improve significantly on the polynomial smoothers NW and PS(2), particularly in sparse situations. The polynomial smoother PS(2) shows its best performance with respect to the SPSUP criteria for distributions with very small cell probabilities and not so sparse observations ($n = 2k$).

To confirm the fact that the penalized smoothers seem not very well suited for the estimation of very small probability values it is instructive to look at the same kind of information as before but taking the sup in (3) avoiding the boundaries where the too small probabilities appear. In this way we try to identify where the bad behaviour of the SPSUP happens. As expected the better behaviour with respect to the discretization of $\beta(.6,.6)$ is confirmed. The penalized and quadratic penalized smoothers improve significantly for the estimation of the discretized $g_1(\cdot)$ density, which has a smoother boundary behaviour. These estimators are of equivalent quality as the Nadaraya-Watson estimator and tend, for sparse observations ($n = k/2$), to behaviour

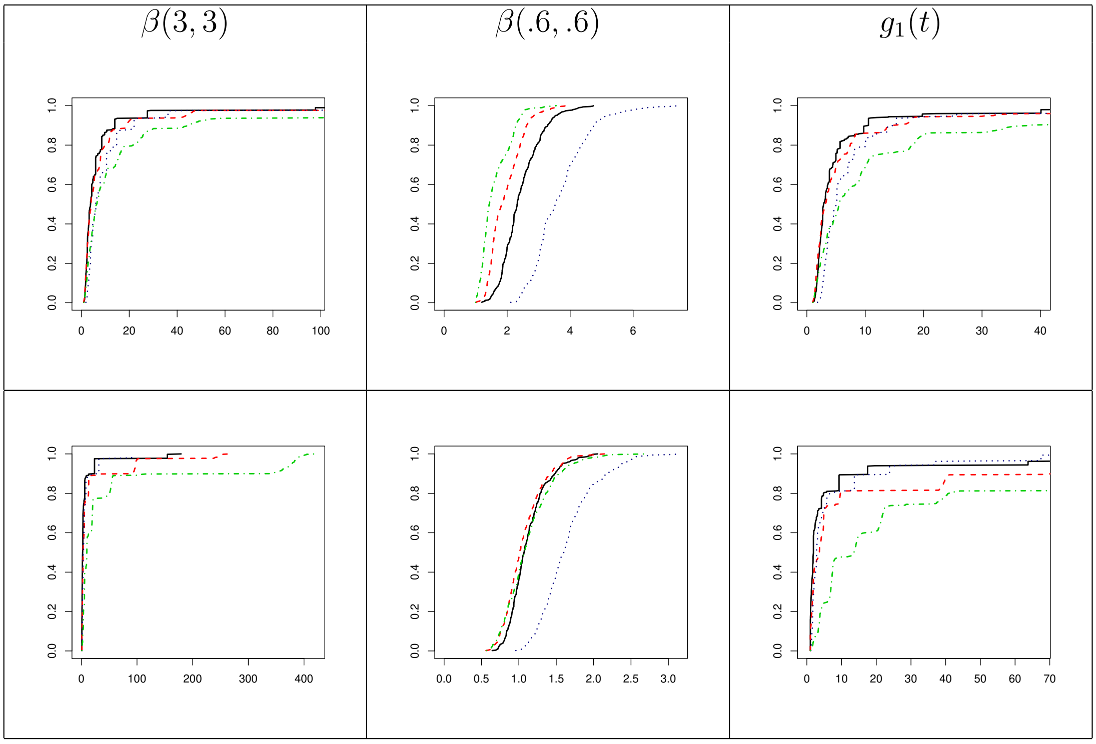

TABLE 3. Empirical distribution of SPSUP ($k = 100$, first line $n = 50$, second line $n = 200$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

better as what the right tail of the SPSUP distribution is regarded. So the penalized estimators are less likely to produce large values of this error criteria. Among the penalized smoothers it is PPS(0) that seems to perform better in the presence of very small probabilities. This confirms again the impression that the more one penalizes the error functions the more one overestimates small probabilities. This could be useful if we have some prior knowledge concerning the underlying distribution. For nonsparse observations ($n = 2k$) the penalized smoothers do not perform so well. The empirical results for sparse observations are described in Table 5.

We now comment on the behaviour of the estimators with respect to the NINF error criteria. As this is not a relative error this criteria is not affected by the existence of very small probabilities. For sparse observations ($n = k/2$) the penalized smoothers PPS(0) and P2PS(0) show the better performance, being for the discretized $\beta(.6, .6)$ approached by the Nadaraya-Watson estimator in some cases. When the degree of sparseness decreases ($n = 2k$) the behaviour of the Nadaraya-Watson becomes similar to the one

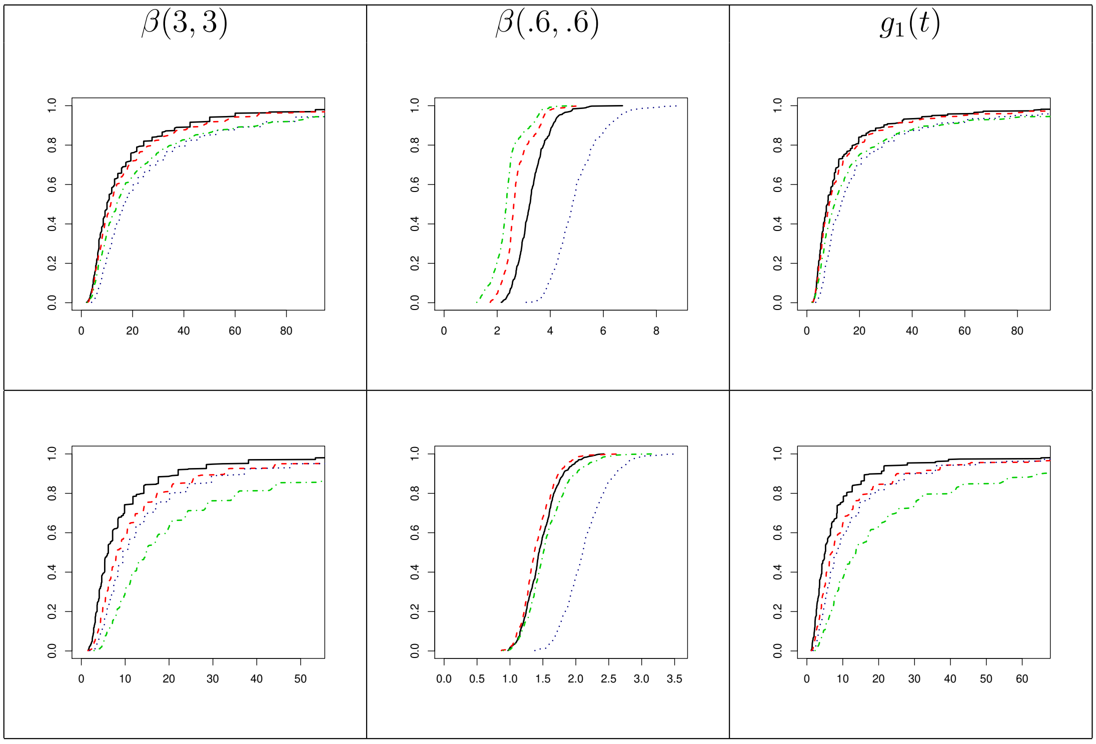

TABLE 4. Empirical distribution of SPSUP ($k = 500$, first line $n = 250$, second line $n = 1000$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

of PPS(0) and P2PS(0), althought the classical Nadaraya-Watson tends to produce larger values for NINF than the penalized smoothers, as is visible on the right tail of the empirical distribution. In each case the polynomial smoother of degree $p = 2$, PS(2) tends to show the worse behaviour in what this error criteria is concerned. The simulation results are shown in Tables 6 and 7.

Appendix

Appendix A. Derivation of the PPS estimators

We start by computing the derivative of H, given by (15), with respect to the polynomial parameters in each cell, that is, with respect to β_l, $l =$

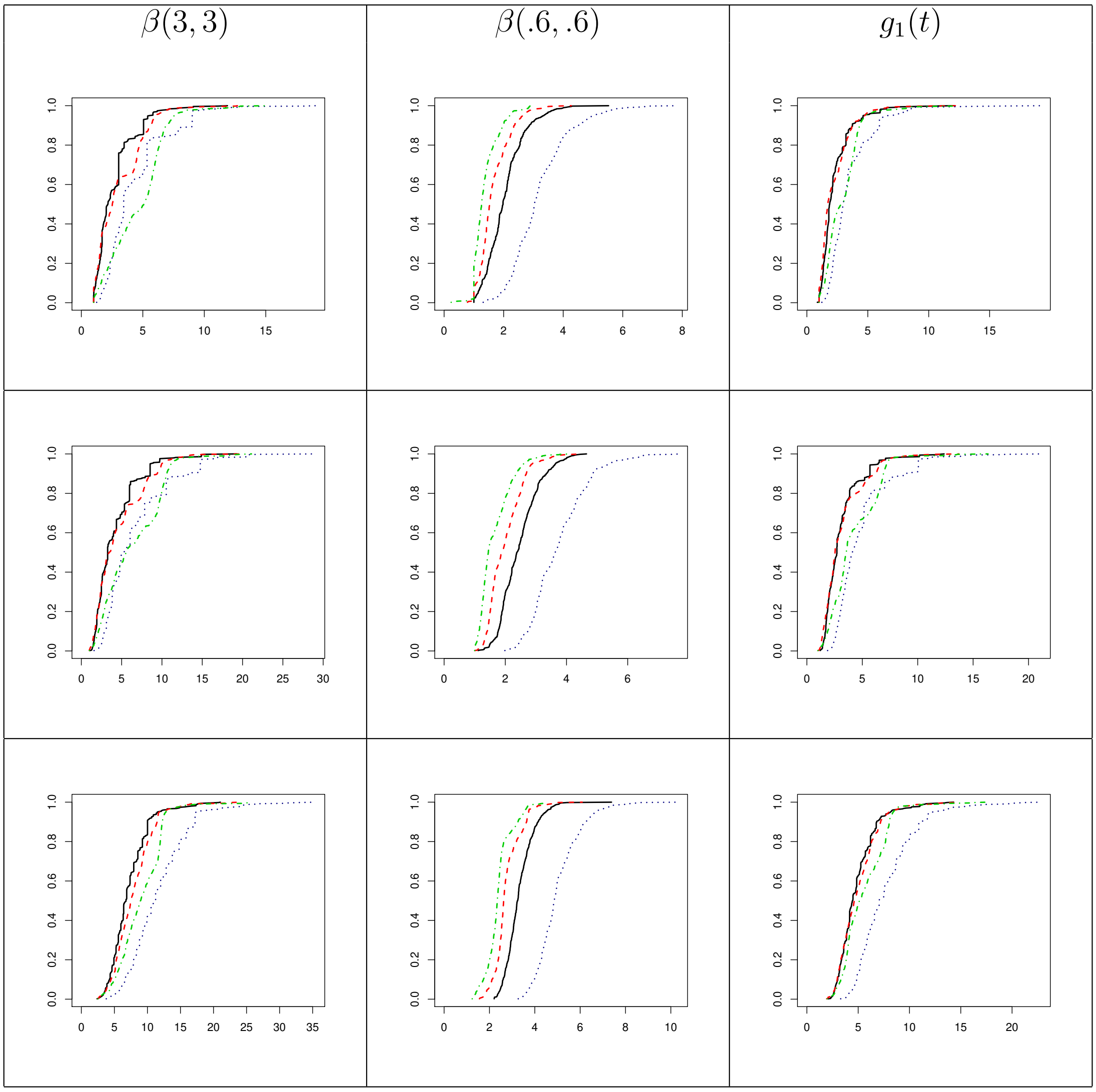

TABLE 5. Empirical distribution of SPSUP ignoring small probabilities (first line $k = 50$, second line $k = 100$, third line $k = 500$, $n = k/2$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

$1, \ldots, k$. This gives rise to the following system of $(p+1)$ equations:

$$-2\mathbf{X}_l^t\mathbf{W}_l(\overline{\mathbf{P}} - \mathbf{X}_l\beta_l) - \frac{\mathbf{h}}{\beta_{0,l}}(\overline{\mathbf{P}} - \mathbf{X}_l\beta_l)^t\mathbf{W}_l(\overline{\mathbf{P}} - \mathbf{X}_l\beta_l) = -\lambda\beta_{0,l}\mathbf{h}. \quad (21)$$

We remark that only the first of these equations is nonlinear. Let us start by solving the linear part of this system of equations. For this purpose define $\widetilde{\beta}_l = \beta_l - \beta_{0,l}\mathbf{h}$, and $\widetilde{\mathbf{X}}_l$ the $(p+1)$-column matrix obtain from $\mathbf{X}_l$ by replacing

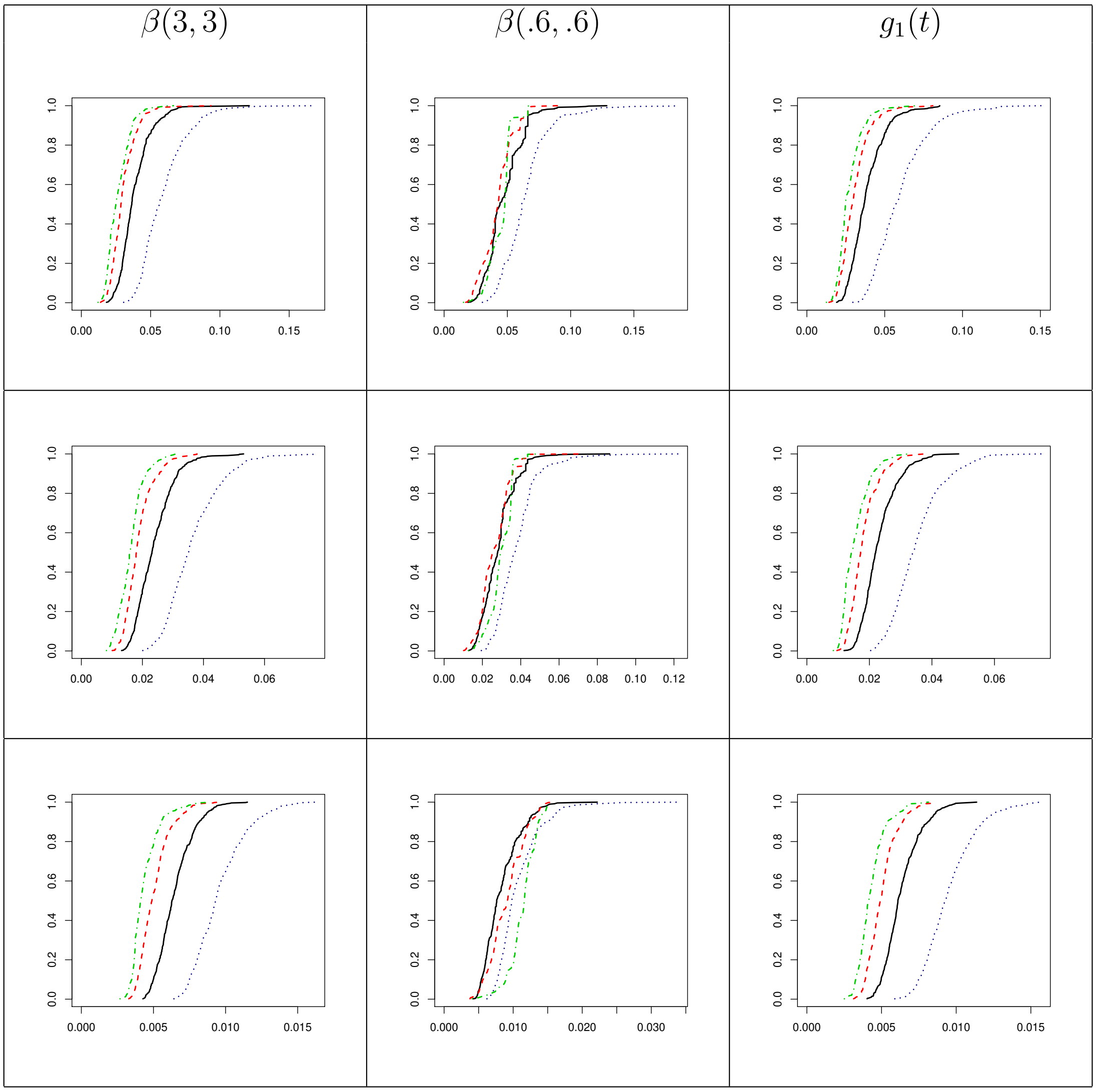

TABLE 6. Empirical distribution of the NINF for sparse observations (first line $k = 50$, second line $k = 100$, third line $k = 500$, $n = k/2$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

its first column entries by zero's, that is

$$\mathbf{X}_l = \begin{bmatrix} 0 & x_{1-k} - x_l & \cdots & (x_{1-k} - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & x_0 - x_l & \cdots & (x_0 - x_l)^p \\ 0 & x_1 - x_l & \cdots & (x_1 - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & x_k - x_l & \cdots & (x_k - x_l)^p \\ 0 & x_{k+1} - x_l & \cdots & (x_{k+1} - x_l)^p \\ & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & x_{2k} - x_l & \cdots & (x_{2k} - x_l)^p \end{bmatrix}.$$

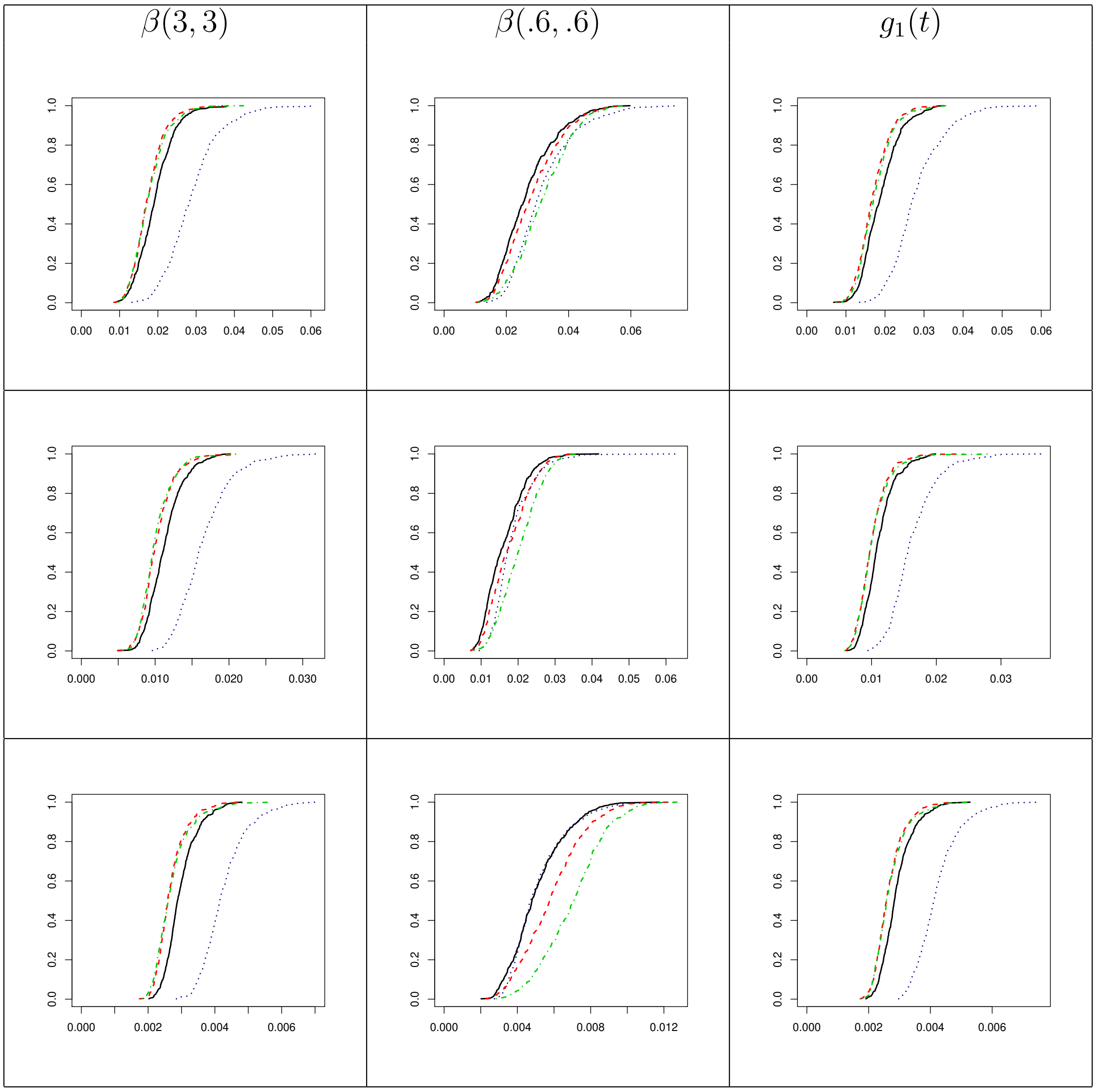

TABLE 7. Empirical distribution of the NINF for nonsparse observations (first line $k = 50$, second line $k = 100$, third line $k = 500$, $n = 2k$). Smoothers: NW (solid), PS(2) (dotted), PPS(0) (dashed), P2PS(0) (dotdash).

Then, we have that

$$\mathbf{X}_l\beta_l = \widetilde{\mathbf{X}}_l\widetilde{\beta}_l + \beta_{0,l}\mathbf{e}, \tag{22}$$

where $\mathbf{e} = (1, \ldots, 1)^t$, and the linear part of the system (21) may be rewritten as

$$\widetilde{\mathbf{X}}_l^t\mathbf{W}_l(\overline{\mathbf{P}} - \beta_{0,l}\mathbf{e}) = \widetilde{\mathbf{X}}_l^t\mathbf{W}_l\widetilde{\mathbf{X}}_l\widetilde{\beta}_l.$$

Now the matrix $\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l$ is of the form

$$\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & & A \end{bmatrix}.$$

Whenever the block A is nonsingular we shall say the matrix has a generalized inverse and write

$$\left(\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l\right)^{\leftarrow} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & & A^{-1} \end{bmatrix}.$$

With this remark, we have that

$$\widetilde{\beta}_l = \left(\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l\right)^{\leftarrow} \widetilde{\mathbf{X}}_l^t \mathbf{W}_l (\overline{\mathbf{P}} - \beta_{0,l}\mathbf{e}). \tag{23}$$

Now we go back to the first equation of (21). To isolate this equation we multiply (21) by $\beta_{0,l}\mathbf{h}^t$. Using (22), it follows that

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

Expanding and noting that $\mathbf{e}^t\mathbf{W}_l\mathbf{e} = 1$, the previous equation is equivalent to

$$\left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l\widetilde{\beta}_l\right)^t \mathbf{W}_l \left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l\widetilde{\beta}_l\right) - \beta_{0,l}^2 = \lambda\beta_{0,l}^2. \tag{24}$$

Now we replace $\widetilde{\beta}_l$ using (23) to find, after some simplification,

$$\overline{\mathbf{P}}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \overline{\mathbf{P}} - \beta_{0,l}^2 \mathbf{e}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \mathbf{e} = \lambda\beta_{0,l}^2, \tag{25}$$

where

$$\mathbf{A}_l = \mathbf{W}_l^t \widetilde{\mathbf{X}}_l \left(\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l\right)^{\leftarrow} \widetilde{\mathbf{X}}_l^t \mathbf{W}_l.$$

Now the replication device we described in the previous section comes into action: $\mathbf{e}^t\mathbf{A}_l\mathbf{e}$ does not depend on l, due to the fact that we always deal with interior cells.

Before continuing the analysis we note a useful fact: of we remove the constraint $\sum_{l=1}^k \beta_{0,l} = 1$, which is equivalent to setting $\lambda = 0$, we find that

$$\overline{\mathbf{P}}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \overline{\mathbf{P}} = \beta_{0,l}^2 \mathbf{e}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \mathbf{e},$$

thus the random variable $\overline{\mathbf{P}}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \overline{\mathbf{P}}$ has, for each $l = 1, \ldots, k$, the same sign as $\mathbf{e}^t \left(\mathbf{W}_l - \mathbf{A}_l\right) \mathbf{e} = 1 - \mathbf{e}^t\mathbf{A}_l\mathbf{e}$, which not only has always the same sign but, as noted before, is even constant for the cells C_l, $l = 1, \ldots, k$. We prove, in Appendix C, that for $p = 0, 1, 2, 3$, $\mathbf{e}^t (\mathbf{W}_l - \mathbf{A}_l) \mathbf{e} > 0$. Thus, for $p = 0, 1, 2, 3$, it follows from (25), that

$$\beta_{0,l}^2 \left(\lambda + \mathbf{e}^t (\mathbf{W}_l - \mathbf{A}_l) \mathbf{e}\right) > 0.$$

Finally, taking into account that $\beta_{0,l}$, being an estimator for a probability, should positive, we derive

$$\beta_{0,l} = \frac{\left(\overline{\mathbf{P}}^t (\mathbf{W}_l - \mathbf{A}_l) \overline{\mathbf{P}}\right)^{1/2}}{(\lambda + \mathbf{e}^t (\mathbf{W}_l - \mathbf{A}_l) \mathbf{e})^{1/2}}. \tag{26}$$

Using the constraint to optimization problem, we find that

$$\left(\lambda + \mathbf{e}^t (\mathbf{W}_l - \mathbf{A}_l) \mathbf{e}\right)^{1/2} = \sum_{l=1}^{k} \left(\overline{\mathbf{P}}^t (\mathbf{W}_l - \mathbf{A}_l) \overline{\mathbf{P}}\right)^{1/2}.$$

From this equation and (26) the expression (16) for the penalized polynomial smoother of degree p, PPS(p) follows.

Appendix B. Derivation of the P2PS estimators

We now give some details on the computation of the P2PS(p) estimators given by (20). Putting the derivative of H^* with respect to $\beta = (\beta_1, \ldots, \beta_p)$ equal to 0 produces the following $(p + 1)$ equations:

$$\frac{-2\mathbf{X}_l^t \mathbf{W}_l \left(\overline{\mathbf{P}} - \mathbf{X}_l \beta_l\right)}{\beta_{0,l}^2} - \frac{2\mathbf{h}}{\beta_{0,l}^3} (\overline{\mathbf{P}} - \mathbf{X}_l \beta_l)^t \mathbf{W}_l (\overline{\mathbf{P}} - \mathbf{X}_l \beta_l) = 0. \tag{27}$$

Just as for the previous case, this system of equations is linear except for the first equation. Moreover, assuming that the $\beta_{0,l}$ are nonzero, the linear part of this system of equations coincides with the linear part of (21). Thus, setting as before $\widetilde{\beta}_l = \beta_l - \beta_{0,l}\mathbf{h}$, the representation (23) still holds. Going back to (27) we isolate the first equation by left-multiplying the equation by $\beta_{0,l}^3 \mathbf{h}^t$. After using (22), this first equation rewrites as

$$\left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l \widetilde{\beta}_l\right)^t \mathbf{W}_l \left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l \widetilde{\beta}_l\right) - \beta_{0,l} \left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l \widetilde{\beta}_l\right)^t \mathbf{W}_l \mathbf{e} = 0. \tag{28}$$

Now we need some simplification of this expression to recover (20). Using the matrix algebra of the derivation of the previous family of estimators we know that

$$\left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l \widetilde{\beta}_l\right)^t \mathbf{W}_l \left(\overline{\mathbf{P}} - \widetilde{\mathbf{X}}_l \widetilde{\beta}_l\right) = \overline{\mathbf{P}}^t (\mathbf{W}_l - \mathbf{A}_l) \overline{\mathbf{P}} + \beta_{0,l}^2 \mathbf{e}^t \mathbf{A}_l \overline{\mathbf{P}}.$$

Expanding the other term in (28), and using (23), it now easily follows that

$$\beta_{0,l} = \frac{\overline{\mathbf{P}}^t(\mathbf{W}_l - \mathbf{A}_l)\overline{\mathbf{P}}}{\overline{\mathbf{P}}^t(\mathbf{W}_l - \mathbf{A}_l)\mathbf{e}} \,.$$

Finally, recall that this is the solution to optimization problem without the constraint so, to obtain as solution a probability distribution, we normalize by the sum of the $\beta_{0,l}$, $l = 1, \dots, k$.

Appendix C. Explicit computation of $\mathbf{A}_l$

We give brief indications about the explicit computation of $\mathbf{A}_l$ for the cases $p = 1, 2, 3$ and verify that, for these values of p, $\mathbf{e}^t\left(\mathbf{W}_l - \mathbf{A}_l\right)\mathbf{e} > 0$. Given $\alpha \in \mathbb{N}$, define

$$S_\alpha = \sum_{j=-u}^{u} (x_l - x_{l+j})^\alpha p(j) = \sum_{j=-u}^{u} \frac{j^\alpha}{k^\alpha}\, p(j).$$

It follows by symmetry that $S_1 = S_3 = S_5 = 0$. As for the even values of α,

$$S_2 = \frac{\sigma^2}{k^2}\,, \qquad S_4 = \frac{\tau^4}{k^4}\,, \qquad S_6 = \frac{\gamma^6}{k^6}\,.$$

For $p = 3$, it follows immediately that

$$\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & S_2 & 0 & S_4 \\ 0 & 0 & S_4 & 0 \\ 0 & S_4 & 0 & S_6 \end{bmatrix}.$$

For the cases $p = 1, 2$, the matrix is just the submatrix of the previous one corresponding to the square block of order 2 or 3, respectively, situated on the north-east corner. The generalized inverse $\left(\widetilde{\mathbf{X}}_l^t \mathbf{W}_l \widetilde{\mathbf{X}}_l\right)^{\leftarrow}$ is easily identified:

$$p = 1 \qquad \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{S_2} \end{bmatrix}$$

$$p = 2 \qquad \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{S_2} & 0 \\ 0 & 0 & \frac{1}{S_4} \end{bmatrix}$$

$$p = 3 \qquad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{S_6}{\Delta} & 0 & -\frac{S_4}{\Delta} \\ 0 & 0 & \frac{1}{S_4} & 0 \\ 0 & -\frac{S_4}{\Delta} & 0 & \frac{S_6}{\Delta} \end{bmatrix}$$

where $\Delta = S_2 S_6 - S_4^2$. Explicit expressions may now be given for the entries $\mathbf{A}_l(i, j)$ of the matrix:

$p = 1$: $\mathbf{A}_l(i,j) = \frac{1}{\sigma^2} p(i-l)p(j-l)(i-l)(j-l)$;

$$p = 2\text{: } \mathbf{A}_l(i,j) = p(i-l)p(j-l)\left[\frac{(i-l)(j-l)}{\sigma^2} + \frac{(i-l)^2(j-l)^2}{\tau^4}\right];$$

$$p = 3\text{: } \mathbf{A}_l(i,j) = p(i-l)p(j-l)\left[\frac{(i-l)(j-l)\gamma^6}{\sigma^2\gamma^6 - \tau^8} - \frac{(i-l)^3(j-l)\tau^4}{\sigma^2\gamma^6 - \tau^8}\right.$$
$$\left. + \frac{(i-l)^2(j-l)^2}{\tau^4} - \frac{(i-l)(j-l)^3\tau^4}{\sigma^2\gamma^6 - \tau^8} + \frac{(i-l)^3(j-l)^3\gamma^6}{\sigma^2\gamma^6 - \tau^8}\right].$$

It is now easy to verify that

$$\mathbf{e}^t\left(\mathbf{W}_l - \mathbf{A}_l\right)\mathbf{e} = 1 - \mathbf{e}^t\mathbf{A}_l\mathbf{e} = 1 - \sum_{i,j}\mathbf{A}_l(i,j) > 0.$$

In fact, up to the multiplication by a constant, the sum reduces to terms of the form

$$\sum_{i,j} p(i-l)p(j-l)(i-l)^{\alpha}(j-l)^{\alpha'}$$
$$= \left(\sum_i p(i-l)(i-l)^{\alpha}\right)\left(\sum_j p(j-l)(j-l)^{\alpha'}\right).$$

Thus, whenever α or α' are odd this sum is equal to 0. So, for $p = 1$, we have $\mathbf{e}^t\left(\mathbf{W}_l - \mathbf{A}_l\right)\mathbf{e} = 1$, while for $p = 2, 3$, we have $\mathbf{e}^t\left(\mathbf{W}_l - \mathbf{A}_l\right)\mathbf{e} = 1 - \frac{\sigma^4}{\tau 4} > 0$, using the Cauchy-Schwarz inequality.

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