Computational Statistics (1998) 13:447–461




# Bandwidth Selection for Local Polynomial Smoothing of Multinomial Data

I. Augustyns[1], M. P. Wand[2,3]

[1] Department of Statistics, Limburgs Universitair Centrum, B–3590 Diepenbeek, Belgium
[2] Australian Graduate School of Management, University of New South Wales, Sydney, 2052, Australia

**Summary**

We develop a rule for choosing bandwidths for local polynomial smoothing of ordered multinomial data. Our method is a variant of the double smoothing idea and is particularly geared towards good performance near the boundaries of the data, through the use of exact risk expressions.

**Keywords:** Categorical data, Contingency table, Double smoothing, Kernel estimator, Sparse multinomial data.

## 1 Introduction

It has been recognised for some time that smoothing can be beneficial in the estimation of cell probabilities from ordered multinomial data. Improvements over ordinary frequency estimates tend to be greater when the data are sparse among the cells. Simonoff (1995) provides a succinct summary of the large amount of literature that has dealt with this issue since the seminal paper of Aitchison and Aitken (1976).

[3]Research supported in part by the Limburgs Universitair Centrum (Belgium).

As the review of Simonoff shows, smoothing of multinomial data can be executed in a number of ways. Popular approaches include penalised likelihood and Bayesian ideas. Recently Aerts, Augustyns and Janssen (1997) investigated the application of local polynomial smoothing to the multinomial context. This is a relatively old approach to smoothing that has recently gained renewed popularity. One reason for this is the realization that local polynomial smoothers, in a certain sense, automatically correct for boundary effects (e.g. Fan and Gijbels, 1992, Hastie and Loader, 1993, Ruppert and Wand, 1994, Cheng, Fan and Marron, 1996). This is generally viewed as preferrable to the use of boundary kernel adjustments (e.g. Dong and Simonoff, 1994).

An example of multinomial smoothing where proper handling of the boundaries is very important is displayed in Figure 1. The data are based on time intervals between explosions in mines involving more than 10 miners killed in Great Britain from December 8, 1875 to May 29, 1951, and are given in Table 2 of Simonoff (1983). Figure 1a shows the cell proportions, while Figure 1b shows a local linear kernel smooth of these using a Gaussian kernel and bandwidth of 66, chosen by cross-validation. Clearly there is a considerable amount of structure near the boundary so proper boundary adjustment is desirable. To gain a better understanding of what "proper boundary adjustment" means for this type of data, we constructed Figure 2. In this example

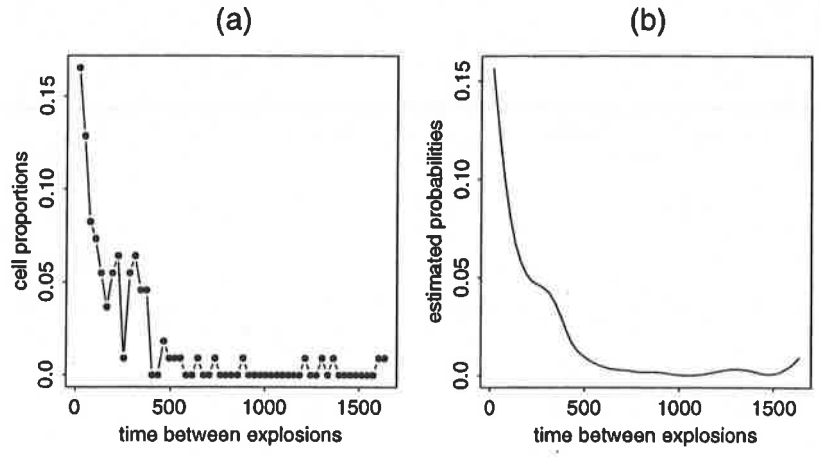


*Figure 1. (a) Cell probabilities for mine data. (b) Local linear kernel smooth of cell probabilities.*

there are 100 cells and the probabilities are obtained by integrating the function $e^{-5x}$ over the disjoint intervals surrounding each cell centre, and then renormalising. This function (divided by 5) is shown by the dotted curve and is chosen to roughly resemble the mine data structure. The circles correspond to the exact mean summed squared error optimal bandwidths over subsets of cells of size 5. In the interior, as one moves from the right to the left in Figure 2 it is seen that the optimal bandwidth decreases. This is consistent with the idea that the curvature is increasing and therefore a smaller bandwidth is optimal. However, near the boundary itself, the optimal amount of smoothing increases – reflecting the fact that near the boundary the variance of a local linear estimator increases. This indicates that one may wish to allow the bandwidth to vary across the cells since a bandwidth chosen for good performance in the interior is not necessarily good for estimation close to the boundary, despite the automatic boundary adjustments of local lines. This should be even more apparent in higher dimensional tables where the "boundary region" comprises a larger proportion of the cells (Dong and Simonoff, 1995).

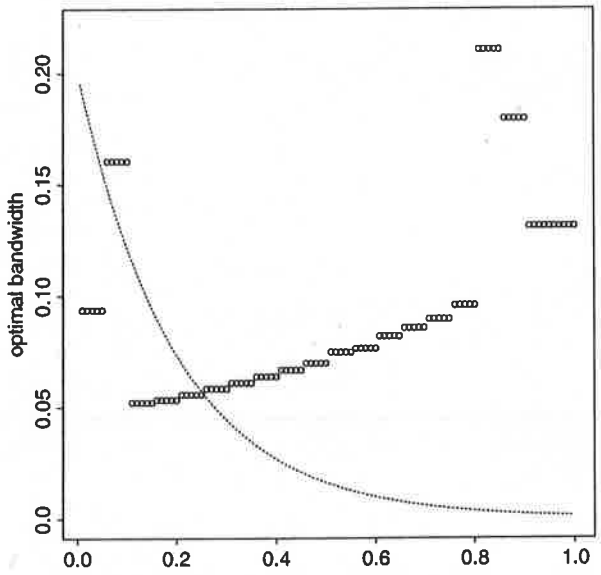


*Figure 2. Optimal bandwidths with respect to mean summed squared error over subsets of cells of size 5 for probabilities obtained from integrating* $e^{-5x}$.

Our goal in this paper is to develop methodology for selecting local and global bandwidths for local polynomial smoothing of multinomial data. For local bandwidth selection we follow the recommendations of Hall, Marron

and Titterington (1995) and focus on partial local smoothing, over subsets of cells, rather than "purely" local smoothing where a different bandwidth is chosen for each cell. Because of the fact that classical "small bandwidth" asymptotics tend to mask the effect of boundaries we aim to work with finite sample estimates of risk measures. Our resultant approach can be thought of as being a non-asymptotic version of the double smoothing idea used, for example, by Müller (1985), Staniswalis (1989) and Härdle, Hall and Marron (1992).

We describe our problem more formally in Section 2, and develop a solution in Section 3. Section 4 contains some heuristics on theoretical properties of our bandwidth selector while Section 5 investigates practical performance.

## 2 Local Polynomial Smoothers and Partially Local Bandwidths

Let $P = [P_1, \ldots, P_k]^T$ denote the probabilities of a $k$-cell multinomial distribution and let $N = [N_1, \ldots, N_k]^T$ represent the counts generated from this multinomial distribution with sample size $n = \sum_{i=1}^{k} N_i$. We denote the cell proportions by $\bar{P} = [\bar{P}_1, \ldots, \bar{P}_k]^T$.

Smoothing is appropriate when the cells have a natural ordering with respect to the index $i$ $(1 \leq i \leq k)$. The usual way to do this is to apply a nonparametric regression technique to the pairs $(x_1, \bar{P}_1), \ldots, (x_k, \bar{P}_k)$, where $x_i = (i - 1/2)/k$, $1 \leq i \leq k$, are fixed equidistant design points in $(0, 1)$. The local $p$th degree polynomial smoother for estimating $P_i$ is

$$\hat{P}_i(h) = e_1^T (X_i^T W_i X_i)^{-1} X_i^T W_i \bar{P}, \tag{2.1}$$

where

$$X_i = \begin{bmatrix} 1 & x_1 - x_i & \ldots & (x_1 - x_i)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k - x_i & \ldots & (x_k - x_i)^p \end{bmatrix}$$

$$W_i = \operatorname{diag}\left\{ K\left(\frac{x_1 - x_i}{h}\right), \ldots, K\left(\frac{x_k - x_i}{h}\right) \right\}$$

and $e_1$ is the $(p+1)$-vector $[1, 0, \ldots, 0]^T$. The kernel $K$ is assumed to be a symmetric, compactly supported density and $h > 0$ is the bandwidth. In matrix notation (2.1) can be written as $\hat{P}(h) = S_h \bar{P}$, where $S_h$ is a $k \times k$ matrix, often referred to as the smoother matrix.

For a subset $A$ of the set of indices $\{1, \ldots, k\}$ we let $\hat{P}^A(h)$ denote the vector of estimates for the cell probabilities with index in $A$. We may write $\hat{P}^A(h) = I_A \hat{P}(h)$ where $I_A$ is a matrix of zeros and ones that maps $[1, \ldots, k]^T$ to $A$. Also, let $S_h^A = I^A S_h$ so that $\hat{P}^A(h) = S_h^A \bar{P}$.

Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ be a partition of the set of all indices $\{1, \ldots, k\}$. The idea of partially local bandwidth selection is to determine $r$ bandwidths $h^{A_1}, \ldots, h^{A_r}$, all of which are optimal for their corresponding set $A_i$. Trivial partitions are $\mathcal{A} = \{\{1, \ldots, k\}\}$, the case of a global bandwidth, and $\mathcal{A} = \{\{1\}, \ldots, \{k\}\}$, the case of purely local bandwidths. All other partitions of $\{1, \ldots, k\}$ are examples of partially local bandwidths.

# 3 A Bandwidth Selection Strategy

In this section we focus on selecting the bandwidth for estimating $P^A = I_A P$. The performance of the estimator $\hat{P}^A(h)$ is measured by the mean sum of squared errors over $A$

$$\text{MSSE}_A(h, P) = \sum_{i \in A} E\{\hat{P}_i(h) - P_i\}^2.$$

With respect to this risk measure, the optimal bandwidth is

$$h_M^A = \underset{h>0}{\text{argmin}}\{\text{MSSE}_A(h, P)\}.$$

We will take $h_M^A$ as our goal.

It may be shown that

$$\text{MSSE}_A(h, P) = \|(S_h^A - I_A)P\|^2 - n^{-1}\|S_h^A P\|^2 + n^{-1}\text{tr}\{S_h^A \text{diag}(P)(S_h^A)^T\} \tag{3.1}$$

where $\|v\| = (v^T v)^{1/2}$ denotes the norm of the vector $v$ and $\text{diag}(P)$ the $k \times k$ matrix with $P_1, \ldots, P_k$ on the diagonal. A double smoothing bandwidth selection strategy involves replacing the unknown $P$ in (3.1) by a "pilot" estimate $S_{g^A}\bar{P}$ where $g^A$ is another bandwidth, which we will refer to as the pilot bandwidth. This leads to a class of bandwidth selectors $\{\hat{h}_{g^A}^A : g^A > 0\}$ where

$$\hat{h}_{g^A}^A = \underset{h>0}{\text{argmin}}\, \text{MSSE}_A(h, S_{g^A}\bar{P}).$$

The problem now boils down to the choice of $g^A$. The recent literature has seen a good deal of theory on the choice of the pilot bandwidth in double smoothing strategies, particularly in the density estimation context. See, for example, Jones, Marron and Park (1991), Hall, Marron and Park (1992), Härdle, Hall and Marron (1992) and Park and Marron (1992). In each case, attention has focussed on the asymptotic distribution of the relative error $(\hat{h}_{g^A}^A - h_M^A)/h_M^A$, with the choice of $g^A$ aimed at optimising the rate of convergence to a limiting normal distribution. While this is a reasonable approach in many smoothing contexts, it has its limitations in those contexts where the boundary effects are large since the usual asymptotic approximations can be quite inaccurate.

A natural alternative for selecting $g^A$ is to again use exact risk ideas as a guideline. One can define the optimal $g^A$ to be the one minimising the mean squared distance between $\hat{h}^A_{g^A}$ and $h^A_M$

$$g^A_M = \underset{g^A>0}{\operatorname{argmin}} E(\hat{h}^A_{g^A} - h^A_M)^2 = \underset{g^A>0}{\operatorname{argmin}} \operatorname{MSE}(\hat{h}^A_{g^A})$$

and use this to guide the choice of $g^A$. Unfortunately $\hat{h}^A_{g^A} - h^A_M$ does not have an explicit form which makes its MSE intractable. A way around this is to use the approximation

$$\hat{h}^A_{g^A} - h^A_M = -\operatorname{MSSE}^{[1]}_A(h^A_M, S_{g^A}\bar{P})/\operatorname{MSSE}^{[2]}_A(h^A_M, P) \tag{3.2}$$

where $\operatorname{MSSE}^{[i]}_A(h,P) = (\partial^i/\partial h^i)\operatorname{MSSE}_A(h,P)$, which is based on a Taylor expansion of $\operatorname{MSSE}^{[1]}_A(h,P)$ about $h^A_M$. This, in turn, leads to the following approximation to $g^A_M$:

$$\begin{aligned} \tilde{g}^A_M &= \underset{g>0}{\operatorname{argmin}} E[\{-\operatorname{MSSE}^{[1]}_A(h^A_M, S_g\bar{P})/\operatorname{MSSE}^{[2]}_A(h^A_M, P)\}^2] \\ &= \underset{g>0}{\operatorname{argmin}} L_A(g, h^A_M) \end{aligned}$$

where

$$L_A(g,h) = E\{\operatorname{MSSE}^{[1]}_A(h, S_g\bar{P})^2\}.$$

The advantage of working with $\tilde{g}^A_M$ is that $L_A(g,h)$ admits the following approximation:

$$\begin{aligned} L_A(g,h) \simeq\ & [(1-n^{-1})P^T D^A_{gh} P + n^{-1}\{\text{diagonal}(D^A_{gh})\}^T P]^2 \\ & + 4n^{-1}\operatorname{tr}\{{S'_h}^A \operatorname{diag}(P)(S^A_h)^T\}[(1-n^{-1})P^T D^A_{gh} P + \\ & n^{-1}\{\text{diagonal}(D^A_{gh})\}^T P] + \operatorname{Var}(\bar{P}^T D^A_{gh}\bar{P}) \end{aligned} \tag{3.3}$$

where

$$D^A_{gh} = S_g^T\{({S'_h}^A)^T(S^A_h - I_A) + (S^A_h - I_A)^T {S'_h}^A\}S_g,$$

$({S'_h}^A)_{ij} = (\partial/\partial h)(S^A_h)_{ij}$ and diagonal$(D^A_{gh})$ denotes the column vector containing the diagonal entries of the $k \times k$ matrix $D^A_{gh}$.

Derivation of this result, and an explicit approximation for $\operatorname{Var}(\bar{P}^T D^A_{gh}\bar{P})$ is given in the Appendix.

Not surprisingly $L_A(g,h)$ depends on the unknown probability vector $P$ and, in particular, $\tilde{g}^A_M$ depends on $h^A_M$ – the quantity that we are aiming to estimate in the first place. Therefore, to make this approach workable in practice some initial estimate for $P$ is required. We decided to use the Bayesian regression spline smoother of Smith and Kohn (1996) because of its very good simulation performance using default values of the tuning parameters.

A full description of the algorithm is :

1. Find $\hat{P}_{\text{init}}$ initial estimate for $P$, by applying a Bayesian regression spline smoother to $(x_1, \bar{P}_1), \ldots, (x_k, \bar{P}_k)$.
2. Set up a partition $\mathcal{A} = \{A_1, \ldots, A_r\}$ of the cell indices $\{1, \ldots, k\}$.
3. For each $A \in \mathcal{A}$:
   (a) Find $\hat{h}^A_{M,\text{init}} = \text{argmin}_{h>0} \text{MSSE}_A(h, \hat{P}_{\text{init}})$.
   (b) Find $\hat{g}^A_M = \text{argmin}_{g>0} L_A(g, \hat{h}^A_{\text{init}})$, where $\hat{P}_{\text{init}}$ is used as $P$ in (3.3).
   (c) Estimate $P$ by $\hat{P}_{\hat{g}^A_M}$.
   (d) The selected bandwidth for subset $A$ is $\hat{h}^A_M$, the minimiser of $\text{MSSE}_A(h, \hat{P}_{\hat{g}^A_M})$.
4. If $r > 1$ then fit a natural cubic spline to the pairs

$$(1, \log(\hat{h}^{A_1}_M)), (\kappa_2, \log(\hat{h}^{A_2}_M)), \ldots, (\kappa_{r-1}, \log(\hat{h}^{A_{r-1}}_M)), (k, \log(\hat{h}^{A_r}_M))$$

   where $\kappa_i$ is the mean of the indices in $A_i$. The values of the exponentiation of the spline are then used to give bandwidths for each of the $k$ cells.

The final step overcomes the problem of non-smooth pictures at the change-over from one interval in the partition to the next. The spline is fit to the logarithms of the $\hat{h}^{A_i}_M$ and then exponentiated to ensure that the final bandwidths are positive.

# 4 Theoretical Performance

Because our proposed selection rule is aimed at good performance when asymptotic approximations are not very accurate, it is difficult to assess its theoretical performance in complete generality. However, we are able to give some heuristic arguments that the rule has sound theoretical properties when the asymptotics do take affect.

From Aerts, Augustyns and Janssen (1997) we have

$$\text{MSSE}^{[1]}_A(h, P) \simeq h^3 B \int_A (f'')^2 + n^{-1} h^{-2} V \int_A f$$

where $f$ is the latent density corresponding to $P$ (as defined in Aerts et al, 1997), where

$$B = \{\#A/k^2\} \left( \int u^2 K(u)\, du \right)^2 \quad \text{and} \quad V = \{\#A/k^2\} \int K(u)^2\, du.$$

Since $\int_A f$ can be estimated root-$n$ consistently under sparse asymptotics,

$$\begin{aligned} \mathrm{MSSE}_A^{[1]}(h, \hat{P}_g) &\simeq h^3 B \int_A (\hat{f}_g'')^2 + n^{-1}h^{-2}V \int_A f \\ &\simeq \mathrm{MSSE}_A^{[1]}(h, P) + h^3 B \left\{ \int_A (\hat{f}_g'')^2 - \int_A (f'')^2 \right\} \end{aligned}$$

where $\hat{f}_g$ is the latent density corresponding to $\hat{P}_g$. Hence,

$$L_A(g, h) \simeq E\left( \left[ \mathrm{MSSE}_A^{[1]}(h, P) + h^3 B \left\{ \int_A (\hat{f}_g'')^2 - \int_A (f'')^2 \right\} \right]^2 \right).$$

Now $\mathrm{MSSE}_A^{[1]}(h_M^A, P) = 0$ so if $\hat{h}_{M,\mathrm{init}}^A$ is close to $h_M^A$ then $\hat{g}_M^A$, the minimiser of $L_A(g, \hat{h}_{\mathrm{init}}^A)$, will be close to the $g$ that minimises $E\left\{ \int_A (\hat{f}_g'')^2 - \int_A (f'')^2 \right\}^2$. Adaptation of the results in Park and Marron (1992) can be used to show that such a choice of $g$ is optimal in a double smoothing strategy.

# 5 Practical Performance

Figure 3 shows the results of applying the method to the mine data described in Section 1. The estimated bandwidth function is shown in Figure 3a, and is based on applying EDS to 5 partitions of size 11. Note that the method chooses smaller bandwidths near the boundaries, presumably because of the higher amount of curvature there. The resulting estimator is shown in Figure 3b.

Comparison with Figure 1b shows that there is little difference between the two estimates in this case, so it appears that a global bandwidth may be sufficient for these data. Nevertheless, it is useful to have the option of a locally adaptive bandwidth in those cases of more extreme changes in curvature.

We also conducted a simulation study to compare the global version of our bandwidth selector to existing selection rules. In first instance we considered cross-validatory (CV) rules, for which we implemented two versions. The first one is based on treating $(x_1, \bar{P}_1), \ldots, (x_k, \bar{P}_k)$ as an ordinary regression-type data set. Leaving an observation out in this case corresponds to leaving out a cell $(x_i, \bar{P}_i)$. The second version uses the fact that we are dealing with multinomial data. Leaving an observation out means working with $(N_1/(n-1), \ldots, (N_i - 1)/(n-1), \ldots, N_k/(n-1))$. The two versions are denoted by $\mathrm{CV}_{\mathrm{reg}}$ and $\mathrm{CV}_{\mathrm{mnl}}$ respectively. Hall and Titterington (1987) show that $\mathrm{CV}_{\mathrm{mnl}}$ produces consistent estimates in the case of kernel smoothing of multinomial data. As known in the density estimation context (e.g. Hall and Marron, 1987, Park and Marron, 1990) the performance, both theoretical and practical, of CV is somewhat disappointing.

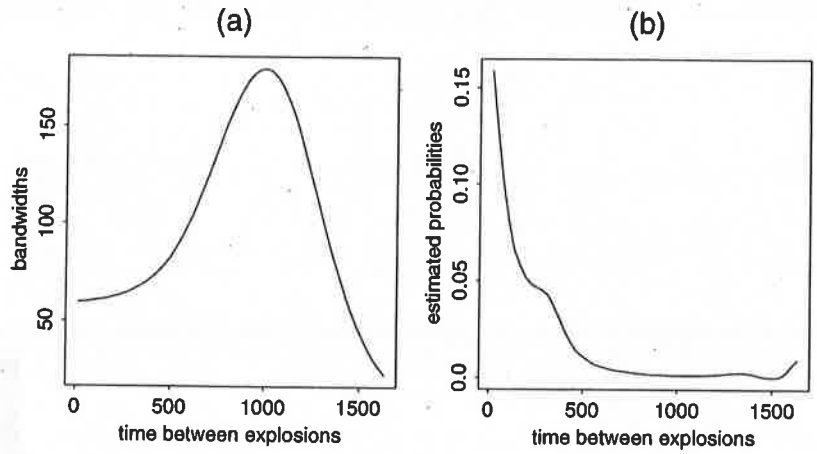


*Figure 3. (a) Estimated bandwidth function using exact double smoothing algorithm described in Section 3 and (b) resulting probability estimates for the mine data.*

We also included the direct plug-in (DPI) bandwidth selector of Ruppert, Sheather and Wand (1995) in our simulation study. This selection method is developed for use in local polynomial smoothing and uses an asymptotic version of the double smoothing strategy.

We assume that the cell probabilities are generated by an underlying latent density on [0,1], which is a standard assumption in the context of smoothing multinomial data. We considered the following latent densities :

1. a on [0,1] truncated exponential density
2. Beta(0.5,0.5) density
3. uniform density on [0,1].

Sparse tables are generated with number of cells $k = 50$ and sample sizes $n = 50$, $n = 100$, $n = 250$. The number of replications in each simulation was 500. The normal density truncated on $[-4, 4]$ was used as kernel.

The difficulties related with the first two latent densities are the very high boundary probabilities. The interesting of the uniform density is that the optimal bandwidth is infinity, i.e. ordinary least squares performs better than local linear regression. All bandwidth selection strategies, except DPI,

used in the simulations need a minimization step. This is done by using grid search on a logarithmically-equally-spaced grid around $h_M$. In view of the last setting also bandwidth zero, i.e. frequency estimators, and bandwidth infinity, i.e. least squares regression, are appended to the grid.

A graphical summary of the results is given in Figure 4. The plots show kernel density estimates of $\log(\hat{h}) - \log(h_{opt})$ for each rule. The solid line represents the EDS-selector, the long dashed line the DPI-selector, the dotted line the $\mathrm{CV}_{\mathrm{mnl}}$ and the short dashed line the $\mathrm{CV}_{\mathrm{reg}}$-selector. Figure 4a-c are the results for the different sparseness settings for the exponential latent density, and Figure 4d shows the result for the beta latent density when $n = 5k$.

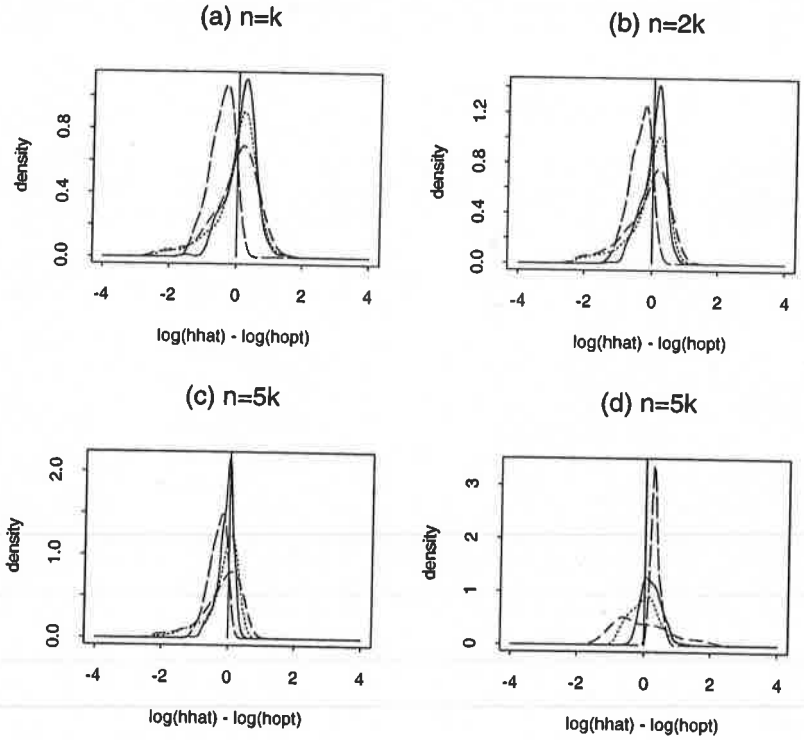


*Figure 4. Kernel density estimates of* $\log(\hat{h}) - \log(h_{opt})$ *for EDS-selector (solid line), DPI-selector (long dashed line),* $CV_{\mathrm{mnl}}$*-selector (dotted line) and* $CV_{\mathrm{reg}}$*-selector (short dashed line). (a)-(c) Exponential latent density. (d) Beta latent density.*

Numerical results of the final cell probability estimators are shown in Tables 1-4. Averages, medians and standard deviations of each of the sum of

squared errors (SSE) are given. We also include the results of the Bayesian regression spline smoother (BR) which is used as the initial estimator for the EDS algorithm.

Paired Wilcoxon tests were performed to determine whether the median SSE's were significantly different. Estimators shared the same SSE ranking when the test showed no difference at the 1% level. The results of these tests are presented in Table 4.

| SSE | $n = 50$ | $n = 100$ | $n = 250$ |
|---|---|---|---|
| EDS | 2.136e-3[a] (2.483e-3)[b] | 1.159e-3 (1.130e-3) | 5.342e-4 (4.721e-4) |
| | 1.364e-3[c] | 0.846e-3 | 3.906e-4 |
| $CV_{reg}$ | 2.713e-3 (3.523e-3) | 1.522e-3 (1.637e-3) | 7.244e-4 (6.829e-4) |
| | 1.644e-3 | 1.022e-3 | 5.112e-4 |
| $CV_{mnl}$ | 2.615e-3 (3.296e-3) | 1.536e-3 (1.626e-3) | 6.938e-4 (6.851e-4) |
| | 1.498e-3 | 0.977e-3 | 0.467e-3 |
| DPI | 2.286e-3 (3.008e-3) | 1.2e-3 (1.294e-3) | 5.571e-4 (4.98e-4) |
| | 1.337e-3 | 0.794e-3 | 4.162e-4 |
| BR | 3.194e-3 (3.908e-3) | 1.716e-3 (1.843e-3) | 7.116e-4 (7.615e-4) |
| | 1.908e-3 | 1.053e-3 | 4.633e-4 |

*Table 1: Numerical Summary of Simulation Study for the Exponential latent density. Average[a] (standard deviation)[b], and median[c] of SSE for each strategy.*

| SSE | $n = 100$ | $n = 250$ |
|---|---|---|
| EDS | 4.298e-3[a] (1.827e-3)[b] | 2.212e-3 (0.851e-3) |
| | 3.934e-3[c] | 2.065e-3 |
| $CV_{reg}$ | 5.143e-3 (2.166e-3) | 2.723e-3 (1.218e-3) |
| | 4.780e-3 | 2.406e-3 |
| $CV_{mnl}$ | 4.396e-3 (2.017e-3) | 2.222e-3 (0.862e-3) |
| | 3.950e-3 | 2.055e-3 |
| DPI | 3.75e-3 (1.637e-3) | 2.15e-3 (0.722e-3) |
| | 3.409e-3 | 2.05e-3 |
| BR | 4.515e-3 (2.152e-3) | 1.85e-3 (1.076e-3) |
| | 4.169e-3 | 1.597e-3 |

*Table 2: Numerical Summary of Simulation Study for the Beta latent density. Average[a] (standard deviation)[b], and median[c] of SSE for each strategy.*

It is clear from both Figure 4 and the tables that EDS exhibits good overall performance and offers significant improvement over the cross-validatory rules. From Figure 4 and Table 4 we conclude that the bandwidth selectors DPI and EDS have comparable behavior for the exponential latent density. For the uniform case DPI seems to perform rather poorly. The EDS method is sometimes bettered by its initial estimator based on BR. However, since the goal of this study is to provide high-quality automatic bandwidth choices

for the simple and intuitive local polynomial smoother, this does not concern us.

Based on the simulation results we finally conclude that it is worthwhile to use exact risk expressions to guide the choice of the bandwidth. We have shown that exact double smoothing, combined with a good initial estimate such as the Bayesian regression spline of Smith and Kohn, results in a reliable bandwidth selector for local polynomial smoothing of multinomial data.

| SSE | $n = 50$ | $n = 100$ | $n = 250$ |
|---|---|---|---|
| EDS | 3.979e-4[a] (5.857e-4[b]) | 2.118e-4 (3.008e-4) | 8.319e-5 (1.100e-4) |
| | 1.722e-4[c] | 1.055e-4 | 4.126e-5 |
| $CV_{reg}$ | 4.111e-4 (5.836e-4) | 2.19e-4 (3.011e-4) | 8.616e-5 (1.094e-4) |
| | 1.829e-4 | 1.161e-4 | 4.605e-5 |
| $CV_{mnl}$ | 4.870e-4 (1.796e-4) | 2.188e-4 (3.01e-4) | 8.609e-5 (1.093e-4) |
| | 1.829e-4 | 1.152e-4 | 4.605e-5 |
| DPI | 1.608e-3 (1.404e-3) | 8.712e-4 (6.759e-4) | 3.334e-4 (2.46e-4) |
| | 1.235e-3 | 6.988e-4 | 2.822e-4 |
| BR | 0.609e-3 (1.3e-3) | 2.439e-4 (4.638e-4) | 7.389e-5 (1.392e-4) |
| | 0.137e-3 | 0.625e-4 | 1.701e-5 |

*Table 3: Numerical Summary of Simulation Study for the Uniform latent density . Average[a] (standard deviation)[b], and median[c] of SSE for each strategy.*

| | | | | | | |
|---|---|---|---|---|---|---|
| Uniform | $n = k$ | EDS | BR | $CV_{mnl}$ | CV | DPI |
| | $n = 2k$ | BR | EDS | CV | $CV_{mnl}$ | DPI |
| | $n = 5k$ | BR | EDS | $CV_{mnl}$ | CV | DPI |
| Exponential | $n = k$ | DPI | EDS | $CV_{mnl}$ | CV | BR |
| | $n = 2k$ | DPI | EDS | $CV_{mnl}$ | CV | BR |
| | $n = 5k$ | EDS | DPI | BR | $CV_{mnl}$ | CV |
| Beta | $n = 2k$ | DPI | EDS | $CV_{mnl}$ | BR | CV |
| | $n = 5k$ | BR | DPI | EDS | $CV_{mnl}$ | CV |

*Table 4: The rankings based on paired Wilcoxon test. The best performer is ranked at the left. When there is no significant difference between different methods they are underlined.*

## Acknowledgement

We are grateful to a referee for helpful comments.

# Appendix: Derivation of (3.3)

In (3.1) the first term is the squared bias contribution and the second and the third term the variance contribution. The second term, which originates from the covariances among the $P_i$'s, is, for sparse tables, of lower order than the third, which is the leading term of the variances of the $P_i$'s. In the derivation for the $L$-function, this term will be omitted. Also we will only replace the $P$ in the squared bias part of $\mathrm{MSSE}_A$ by $S_g\bar{P}$. This is based on the fact that estimation of the variance has a lower order effect on the performance of a double smoothing bandwidth selector. Thus, we work with the approximation

$$\mathrm{MSSE}_A(h, S_g\bar{P}) \simeq \|(S_h^A - I_A)S_g\bar{P}\|^2 + n^{-1}\mathrm{tr}\left\{S_h^A \mathrm{diag}(P)(S_h^A)^T\right\}. \quad \text{(A.1)}$$

Taking the derivative of $\mathrm{MSSE}_A(h, S_g\bar{P})$ with respect to $h$ leads to

$$\mathrm{MSSE}_A^{[1]}(h, S_g\bar{P}) = \bar{P}^T D_{gh}^A \bar{P} + c(h) \quad \text{(A.2)}$$

where $D_{gh}^A$ is given in Section 3 and $c(h) = 2n^{-1}\mathrm{tr}\{S_h'^A \mathrm{diag}(P)(S_h^A)^T\}$. This results in

$$\begin{aligned} E\{\mathrm{MSSE}_A^{[1]}(h, S_g\bar{P})^2\} &= \mathrm{Var}(\bar{P}^T D_{gh}^A \bar{P}) + \left(E(\bar{P}^T D_{gh}^A \bar{P})\right)^2 \\ &\quad + 2c(h)E(\bar{P}^T D_{gh}^A \bar{P}) + c(h)^2. \quad \text{(A.3)} \end{aligned}$$

The last term does not involve $g$ and hence can be left out in the minimization procedure. It is easy to see that

$$E(\bar{P}^T D_{gh}^A \bar{P}) = (1 - n^{-1})P^T D_{gh}^A P + n^{-1}\{\mathrm{diagonal}(D_{gh}^A)\}^T P,$$

which leads to (3.3).

Before (3.3) can be implemented an explicit expression for $\mathrm{Var}(\bar{P}^T D_{gh}^A \bar{P})$ needs to be given. This will be done for a general symmetric $k \times k$ matrix $D$. Since the covariances between the $P_i$'s are of lower order than the variances we will assume that $Y_i$'s are independent in obtaining an approximate expression for $\mathrm{Var}(\bar{P}^T D_{gh}^A \bar{P})$. Thus, the $Y_i = n\bar{P}_i$ will be taken to be independent $\mathrm{Binomial}(n, P_i)$ binomial random variables.

We will use the tensor notation and results of McCullagh (1987). Let $d_{ij}$ denote the $(i, j)$ entry of $D$. Generalized cumulants will be denoted using partitioned superscript notation. For example,

$$\kappa^{i,j} = \mathrm{cum}(Y_i, Y_j) = \mathrm{cov}(Y_i, Y_j) \quad \text{and} \quad \kappa^{i,jk} = \mathrm{cum}(Y_i, Y_jY_k).$$

It is clear that

$$\mathrm{Var}(Y^T DY) = \sum_i \sum_j \sum_k \sum_\ell d_{ij} d_{k\ell} \kappa^{ij,k\ell}.$$

A fundamental identity for generalized cumulants, given on p.58 of McCullagh (1987), states that

$$\begin{aligned}\kappa^{ij,k\ell} &= \kappa^{i,j,k,\ell} + \kappa^{i}\kappa^{j,k,\ell} + \kappa^{j}\kappa^{i,k,\ell} + \kappa^{k}\kappa^{i,j,\ell} + \kappa^{\ell}\kappa^{i,j,k} + \kappa^{i,k}\kappa^{j,\ell} + \kappa^{i,\ell}\kappa^{j,k} \\ &\quad + \kappa^{i}\kappa^{k}\kappa^{j,\ell} + \kappa^{i}\kappa^{\ell}\kappa^{j,k} + \kappa^{j}\kappa^{k}\kappa^{i,\ell} + \kappa^{j}\kappa^{\ell}\kappa^{i,k}.\end{aligned}$$

This implies that, because of the mutual independence of the $Y_i$'s and the symmetry of $D$,

$$\begin{aligned}\mathrm{Var}(Y^T DY) &= \sum_i d_{ii}^2 \kappa^{i,i,i,i} + 4\sum_i \sum_j d_{ij} d_{jj} \kappa^{i} \kappa^{j,j,j} + 2\sum_i \sum_j d_{ij}^2 \kappa^{i,i} \kappa^{j,j} \\ &\quad + 4\sum_i \sum_j \sum_k d_{ij} d_{jk} \kappa^{i} \kappa^{k} \kappa^{j,j}.\end{aligned}$$

Expressions for the first four cumulants of a multinomial distribution are given by,

$$\begin{aligned}\kappa^{i} &= nP_i \\ \kappa^{i,i} &= n(P_i - P_i^2) \\ \kappa^{i,i,i} &= n(P_i - 3P_i^2 + 2P_i^3) \equiv n\gamma(P)_i \\ \kappa^{i,i,i,i} &= n(P_i - 7P_i^2 + 12P_i^3 - 6P_i^4) \equiv n\tau(P)_i.\end{aligned}$$

This results in

$$\begin{aligned}&\mathrm{Var}(\bar{P}^T D\bar{P}) = n^{-3}(\mathrm{diagonal}(D) \odot \mathrm{diagonal}(D))^T \tau(P) \\ &\quad + 4n^{-2} P^T D\{\mathrm{diagonal}(D) \odot \gamma(P)\} + 2n^{-2}\mathrm{tr}(DVDV) + 4n^{-1} P^T DVDP\end{aligned}$$

where $V = \mathrm{diag}(P - P \odot P)$ and $\odot$ means elementwise multiplication.

# References


Aerts, M., Augustyns, I. and Janssen, P. (1997) Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*, **8**, 127–147.

Aitchison, J. and Aitken, C.G.G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Cheng, M.-Y., Fan, J. and Marron, J.S. (1996) On automatic boundary corrections. *The Annals of Statistics*, to appear.

Dong, J. and Simonoff, J.S. (1994) The construction and properties of boundary kernels for smoothing sparse multinomials. *Journal of Computational and Graphical Statistics*, **3**, 57–66.

Dong, J. and Simonoff, J.S. (1995) A geometric combination estimator for $d$-dimensional ordinal contingency tables. *The Annals of Statistics*, **23**, 1143–1159.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371–394.

Hall, P., Marron, J.S. (1987) Extent to which least-squares cross-validation minimizes integrated squared error in nonparametric density estimation. *Probability Theory and Related Fields*, **74**, 567–581.

Hall, P., Marron, J.S. and Park, B.U. (1992) Smoothed cross-validation. *Probability Letters and Related Fields*, **92**,1–20.

Hall, P., Marron, J.S. and Titterington, D.M. (1995) On partial local smoothing rules for curve estimation. *Biometrika*, **82**, 575–588.

Hall, P. and Titterington, D.M. (1987) On smoothing sparse multinomial data. *Australian Journal of Statistics*, **29**, 19–37.

Härdle, W., Hall, P. and Marron, J.S. (1992) Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association*, **87**, 227–233.

Hastie, T. and Loader, C. (1993) Local regression : Automatic kernel carpentry. *Statistical Science*, **8**, 120–129.

Jones, M.C., Marron, J.S. and Park, B.U. (1991) A simple root-$n$ bandwidth selector. *The Annals of Statistics*, **4**, 1919–1932.

McCullagh, P. (1987) *Tensor Methods in Statistics*, London: Chapman and Hall.

Müller, H.-G. (1985) Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statistics and Decisions* Supplement no. **2**, 193–206.

Park, B.U. and Marron, J.S. (1990) Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.

Park, B.U. and Marron, J.S. (1992) On the use of pilot estimators in bandwidth selection. *Journal of Nonparametric Statistics*, **1**, 231–240.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Ruppert, D. and Wand, M.P. (1994) Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**, 1346–1370.

Simonoff, J.S. (1983) A penalty function approach to smoothing large sparse contingency tables. *Annals of Statistics*, **11**, 208–218.

Simonoff, J.S. (1995) Smoothing categorical data. *Journal of Statistical Planning and Inference*, **47**, 41–69.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection, *J. Econometrics*, **75**, 317–344.

Staniswalis, J.S. (1989) Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, **84**, 284–288.