JOURNAL OF MULTIVARIATE ANALYSIS 46, 283–308 (1993)

Kernel Estimators for Cell Probabilities

BIRGIT GRUND

University of Minnesota

Communicated by the Editors

Kernel density estimators for discrete multivariate data are investigated, using the notation framework of contingency tables. We derive large sample properties of kernel estimators and the least-squares cross-validation method for choosing the bandwidth, including the asymptotic bias, the mean summed squared error, the actual summed squared error, and the asymptotic distribution of the resulting non-parametric estimator. We show that the least-squares cross-validation procedure is superior to Kullback–Leibler cross-validation in terms of mean summed squared error, but that the least-squares cross-validation is still sub-optimal concerning actual summed squared error. © 1993 Academic Press, Inc.

1. INTRODUCTION

The estimation of cell probabilities is one of the central issues in the analysis of multivariate discrete (categorical) data. It is equivalent to density estimation for continuous data.

The present paper investigates kernel estimates for "densities" of multivariate discrete data. Their behaviour is strongly influenced by the bandwidth, in the same sense as we know it from their continuous data counterpart: the bandwidth determines the degree of smoothness of the resulting estimator. Here "smoothness" ranges from the extremely unsmooth frequency estimator to the extremely smooth case, where any data set produces the uniform distribution.

This paper is concerned with theoretical aspects of data-dependent bandwidth choice, with particular emphasis on least-squares cross-validation. Meanwhile, in continuous data density estimation much research has been done in this field; for an overview and references see Silverman [22], Hall and Marron [14], Marron [18], and Hall *et al.* [15]. Kernel

Received July 26, 1991; revised October 26, 1992.

AMS 1980 subject classification: primary 62G05; secondary 62G20.

Key words and phrases: kernel estimators, categorical data, cross-validation, smoothing parameter, data-driven bandwidth, discrete density estimation, mean summed squared error, summed squared error, contingency tables.

0047-259X/93 $5.00

estimators for multivariate discrete data are much less investigated; there are only a few theoretical results on data-dependent bandwidth procedures exceeding the fundamental paper by Wang and van Ryzin [27]. Nevertheless, the discrete data case clearly deserves separate treatment.

Our considerations here are motivated by the fact that basic large sample properties of kernel estimators are known to be quite different for the two classes of discrete and continuous densities. Among others the differences concern the convergence rates of the minimal possible mean summed squared error and mean integrated squared error, respectively, as well as the convergence rates of the optimal bandwidths. Therefore, we have no hope to get answers for data-driven bandwidth choice in the discrete setting by simply looking at the continuous data results.

In the following we derive large sample properties of kernel estimators for multivariate discrete densities (cell probabilities), concentrating on the least-squares cross-validation method (LS-method) for the bandwidth choice. Thereby, we assume the number of cells to be fixed, so that in fact we estimate a finite number of parameters (the cell probabilities) under increasing sample size. Hence, it is not surprising that we obtain parametric convergence rates for kernel density estimators, as opposed to the continuous density estimation case. An alternative asymptotic framework has been discussed by Sutherland *et al.* [25], Bishop *et al.* [2], Burman [5, 6], Simonoff [23], and other authors. Here, cell probabilities are represented by areas under a master density curve, whereby the number of cells increases with the sample size. This type of asymptotics can be used to model sparse data in large tables, a realistic setting in categorical data analysis. However, the method is not suited to investigating estimators that are tailored to reflect multivariate structures, as for example the classical kernel estimators proposed by Aitchison and Aitken [1]. For further discussion of the high-dimensional, sparse data problem we refer to Grund and Hall [10].

In the present paper, we consider the random cross-validation bandwith as an intrinsic part of a kernel estimator as opposed to the often used approach of first investigating kernel estimators under nonrandom bandwidths and later to average over the random effect due to the data-dependent bandwidth: an interesting, controversial discussion of both viewpoints can be found in Jones [16] and Mammen [17]. Sections 2 and 3 briefly introduce kernel estimators for discrete data (in the notation framework of contingency tables) and methods for choosing the bandwidth. General assumptions of the asymptotic setting are summarized in Section 4, while Sections 5 and 6 contain the main results. Here we give formulae for the leading terms of bias, variance, and mean summed squared error of kernel estimators with cross-validation bandwidth, and show that these estimators are best asymptotic normally distributed

(Theorem 5.1–5.3). The actual error is regarded in Theorem 5.4, whereas Theorem 5.5 contains generalizations for higher-dimensional bandwidths. In Section 6 the cross-validation bandwidth is compared to optimal parameters. We show that the least-squares cross-validation in average behaves better than the classical Kullback–Leibler method, but that least-squares cross-validation is still suboptimal concerning the actual error. All proofs are deferred to Section 7 and Appendixes A–C; more detailed proofs are given in Grund [9].

2. Kernel Estimators

Our aim is to estimate the density of an m-dimensional categorical random variable $X=(X_1, ..., X_m)'$, based on a sample of N independent realisations of X. Thereby each component X_v can take t_v values, so that X has $t=t_1 \cdot ... \cdot t_m$ possible outcomes.

The sample corresponds to an m-dimensional multinomial contingency table, with cells $\mathbf{i}=(i_1, ..., i_m) \in J=\{\mathbf{i}=(i_1, ..., i_m) : i_v \in \{1, ..., t_v\}, v=1, ..., m\}$, cell probabilities $p_{\mathbf{i}}=P(X=\mathbf{i})$, and N observations. Obviously, estimating the density of X is equivalent to estimating the cell probability vector (cpv) $p=(p_{\mathbf{i}})_{\mathbf{i} \in J}$. The well-known frequency estimator $\hat{\boldsymbol{p}}$ stands for the random observations,

$$N\hat{\boldsymbol{p}} \sim M_t(N, p),$$

where $M_t(N, p)$ denotes the corresponding multinomial distribution.

Throughout the paper we shall use this notation framework of contingency tables, and keep in mind that any estimator for the cpv p estimates the density of the underlying multivariate categorical variable X.

In estimating cell probabilities, the frequency estimator $\hat{\boldsymbol{p}}$ is known to be the best asymptotic normally distributed estimator, yielding

$$\mathcal{L}\{\sqrt{N}(\hat{\boldsymbol{p}}-p)\} \xrightarrow[N \to \infty]{} N_t(0, S(p)), \tag{2.1}$$

where $S(p)=\operatorname{Diag}\{p\}-pp'$, and $\operatorname{Diag}\{p\}$ denotes the diagonal matrix with the diagonal given by the vector p. However, $\hat{\boldsymbol{p}}$ has certain disadvantages for small and moderate sample sizes (zero estimates for nonobserved cells, large variance, etc.). Among the various approaches developed to improve the small sample behaviour, kernel estimators belong to the favoured nonparametric methods. Introduced by Aitchison and Aitken [1], they are based on the hope that cells, declared to be neighbours, really have similar probabilities, so that the observation of neighbour cells provides additional information.

683 46/2-8

According to Titterington [26], we define here kernel estimators as special linear estimators depending on the bandwidth ϑ. We call $\hat{k}$ a kernel estimator for the cpv p, iff

$$\hat{k} = A\hat{p},$$

where $A = A(\vartheta)$ is a kernel matrix (for each $\vartheta \in [0, 1]^r$), i.e., $A = ((a_{\mathbf{ij}}))_{\mathbf{ij} \in J} \in \mathbb{R}^{t \times t}$ is a double-stochastic matrix (nonnegative elements, row and column sum 1) with $a_{\mathbf{jj}} \geqslant a_{\mathbf{ij}}$ for all $\mathbf{i}, \mathbf{j} \in J$.

Examples include:

Pseudo-Bayes Estimators. The kernel estimator defined by

$$a_{\mathbf{ij}}(\vartheta) = \begin{cases} 1 - \vartheta \dfrac{t-1}{t} & \text{if } \mathbf{i} = \mathbf{j} \\ \vartheta / t & \text{else,} \end{cases} \tag{2.2}$$

where $\vartheta \in [0, 1]$, is equivalent to the so-called pseudo-Bayes estimator of Fienberg and Holland [8]:

$$(1 - \vartheta)\hat{\boldsymbol{p}} + \vartheta c_t, \tag{2.3}$$

where $c_t = (1/t)(1, ..., 1)'$ denotes the cpv with equal cell probabilities.

Aitchison and Aitken's Estimators. The classical kernel estimators proposed by Aitchison and Aitken [1] correspond to

$$a_{\mathbf{ij}}(\vartheta) \propto \vartheta^{\sum_{v=1}^{m} d_v(i_v, j_v)} \tag{2.4}$$

in the case of a one-dimensional smoothing parameter $\vartheta \in [0, 1]$, or to

$$a_{\mathbf{ij}}(\vartheta) \propto \vartheta_1^{d_1(i_1, j_1)} \cdot \ldots \cdot \vartheta_m^{d_m(i_m, j_m)} \tag{2.5}$$

with $\vartheta = (\vartheta_1, ..., \vartheta_m)' \in [0, 1]^m$. Thereby $d_v(i_v, j_v)$, $v = 1, ..., m$, measures the distance between the items i_v and j_v of variable X_v; in the case of nominal data e.g. by the 0–1-distance. See also Titterington [26].

Nearest Neighbour Estimators. Given a kernel estimator $A\hat{p}$, a symmetric function $d: J \times J \to \mathbb{R}^+$ measuring the distance between two cells, and a number $h > 0$, we define the corresponding nearest neighbour estimator by

$$a_{\mathrm{NN};\mathbf{ij}} \propto \begin{cases} a_{\mathbf{ij}} & \text{if } d(\mathbf{i}, \mathbf{j}) \leqslant h \\ 0 & \text{else,} \end{cases} \tag{2.6}$$

where a_{ij} and $a_{NN;ij}$ are the elements of the kernel matrices of the original estimator and the attached nearest neighbour estimator, respectively. Among others, this procedure includes the well-known nearest neighbour estimators of Aitchison and Aitken [1], given by (2.4)–(2.6) using

$$d(\mathbf{i}, \mathbf{j}) = \#\{v: i_v = j_v,\ v = 1, ..., m\}.$$

Remark. The nearest neighbour method for cell probabilities corresponds, in the context of estimating the density of a continuous random variable, to the use of kernel functions with compact support, not to the usual nearest neighbour estimators!

In all three examples the bandwidth ϑ has a high impact on the the performance of the entire kernel estimator. Small bandwidths correspond to little smoothing, and under $\vartheta = 0$ data are not smoothed at all; in the latter case the kernel estimators result in the frequency estimator $\hat{\boldsymbol{p}}$, as $A(0) = I$. The larger the bandwith, the more smoothing is involved, up to the extreme value $\vartheta = 1$, where both the pseudo-Bayes and Aitchison and Aitken's estimators degenerate to the constant c_l. The considerable flexibility of kernel density estimators makes bandwidth choice a central issue. Obviously, the optimal amount of smoothing, however measured, depends on the underlying unknown density. Under realistic conditions it has to be estimated, leading to data-driven bandwidth selectors.

3. Choosing the Bandwidth

Much work has been done on data-dependent bandwidth choice, and a variety of methods have been developed. For an overview see Titterington [20], Grund [9], and Santner and Duffy [21]. Here we concentrate on two cross-validation procedures:

The Kullback–Leibler Method. We take the bandwidth $\hat{\vartheta}_{KL}$ minimizing

$$\sum_{\mathbf{i} \in J} \hat{\boldsymbol{p}}_{\mathbf{i}} \ln(\hat{\boldsymbol{p}}_{\mathbf{i}} / \hat{k}_{-\mathbf{i};\mathbf{i}}(\vartheta)),$$

where $\hat{k}_{-\mathbf{i};\mathbf{i}}(\vartheta)$ denotes the element $\mathbf{i}$ of the kernel estimator $\hat{k}_{-\mathbf{i}}(\vartheta)$, computed with one observation missing from cell $\mathbf{i}$, i.e.,

$$\hat{k}_{-\mathbf{i}}(\vartheta) = A(\vartheta)\, \hat{\boldsymbol{p}}_{-\mathbf{i}},$$

where

$$\hat{\boldsymbol{p}}_{-\mathbf{i}} = \frac{N}{N-1}\hat{\boldsymbol{p}} - \frac{1}{N-1}e_{\mathbf{i}},$$

and $e_{\mathbf{i}}$ denotes the vector with "1" in the position of cell $\mathbf{i}$ (again in lexicographical order), and "0" else. Observe that $\hat{\boldsymbol{p}}_{-\mathbf{i}}$ represents the sample, with one observation deleted from cell $\mathbf{i}$.

This procedure is equivalent to the pseudo-likelihood method. In connection with cell probabilities it was first proposed by Aitchison and Aitken [1], who took their prescription from Habbema *et al.* [11] and Duin [7].

The Least-Squares Method. The LS-bandwidth $\hat{\vartheta}_{\mathrm{LS}}$ is defined to minimize

$$\sum_{\mathbf{i}\in J} \hat{\boldsymbol{p}}_{\mathbf{i}} \| \hat{k}_{-\mathbf{i}}(\vartheta) - e_{\mathbf{i}} \|^2, \tag{3.1}$$

where $\|\cdot\|$ is the Euclidean distance.

Remark. Formula (3.1) can be interpreted as an estimate of the mean squared error of prediction if $e_{\mathbf{i}}$ is considered to represent a single observation corresponding to cell $\mathbf{i}$. Therefore, $\hat{\vartheta}_{\mathrm{LS}}$ is expected to provide a small squared error for the resulting kernel estimator. Note that $\hat{\vartheta}_{\mathrm{LS}}$ is completely determined by the random cell frequencies $\hat{\boldsymbol{p}}$, though the cross-validation procedure (3.1) defines the dependence of $\hat{\vartheta}_{\mathrm{LS}}$ on $\hat{\boldsymbol{p}}$ differently for each sample size N.

The LS-cross-validation method was first used by Stone [24] in the context of the pseudo-Bayes estimator (2.3), and was considered again by Rudemo [20] and Bowman [4]. In the last years the LS-procedure has been discussed mainly in connection with kernel estimators for densities of continuous variables.

On the first view, kernel estimators with random bandwidths, say $\hat{k} = A(\hat{\vartheta})\hat{\boldsymbol{p}}$, are driven by two conceptually different sources of randomness. One is the random vector of relative cell frequencies $\hat{\boldsymbol{p}}$, which may be considered as a pilot estimate for p still to be smoothed by the kernel; the other random component is the amount of smoothing, determined by $\hat{\vartheta}$. In practical applications, such as the cross-validation procedures described above, the randomness of $\hat{\vartheta}$ is completely determined by $\hat{\boldsymbol{p}}$, so that the entire kernel estimator including random bandwidth is actually a non-linear, rather complicated function of $\hat{\boldsymbol{p}}$. Theoretical analysis is often tedious, but an honest investigation of bandwidth selectors and kernel estimators has to ackowledge that the same set of data (in our case $\hat{\boldsymbol{p}}$) that

we intend to smooth by the kernel simultaneously drives the amount of smoothing.

Optimal Bandwidth. In order to evaluate the cross-validation method, we compare $\hat{\vartheta}_{LS}$ with, in a certain sense, "optimal" smoothing parameters. For this purpose, we define ϑ_{opt}, the optimal bandwidth minimizing the mean summed squared error

$$\delta_M(\vartheta) = E \| \hat{k}(\vartheta) - p \|^2, \tag{3.2}$$

and $\hat{\vartheta}_{opt}$, the (random) actual optimal bandwidth minimizing the actual summed squared error

$$\delta_A(\vartheta) = \| \hat{k}(\vartheta) - p \|^2. \tag{3.3}$$

Note that (3.2) is the equivalent of the mean integrated squared error in continuous density estimation, whereas (3.3) is similar to both the average squared error and the integrated squared error, as discussed by Marron and Härdle [19]. The latter could be used if we were more interested in the sample at hand than in the average behaviour of a method. For further motivation see Hall and Marron [14] and Hall [13].

4. Assumptions and Notations

In Sections 5 and 6 we discuss asymptotic properties of kernel estimators $\hat{k}_{LS} = A(\hat{\vartheta}_{LS})\hat{\boldsymbol{p}}$, including the random effect of the cross-validation bandwidth. Therefore, we represent the random bandwidth $\hat{\vartheta}_{LS} = \vartheta_{LS}(N, \hat{\boldsymbol{p}})$ and the resulting kernel estimator $\hat{k}_{LS} = k(N, \hat{\boldsymbol{p}})$ as deterministic functions applied to the random sample (represented by $\hat{\boldsymbol{p}}$). Obviously, the functional relation between $\hat{\vartheta}_{LS}$ and $\hat{\boldsymbol{p}}$ is different for each sample size N; hence, we include N as an argument. In detail, our considerations are based on the following two assumptions (which will not be mentioned again):

1. $\{N\hat{\boldsymbol{p}}\}$ is a sequence of multinomial random variables yielding

$$N\hat{\boldsymbol{p}} \sim M_t(N, p) \qquad \text{and} \qquad p \in \mathcal{S}_t,$$

where $\mathcal{S}_t = \{p \in \mathbb{R}^t : p_i \geq 0,\ \sum_{i \in J} p_i = 1\}$ denotes the $(t-1)$-dimensional simplex of all possible cpv's.

2. The estimators $\hat{k}_{LS} = k(N, \hat{\boldsymbol{p}})$ are defined by the function $k : [2, \infty) \times \mathcal{S}_t \to \mathcal{S}_t$ with

$$k(N^*, \hat{p}) = A(\vartheta_{LS}(N^*, \hat{p}))\hat{p} \qquad \text{for all} \quad (N^*, \hat{p}) \in [2, \infty) \times \mathcal{S}_t. \tag{4.1}$$

Here $A(\vartheta)$ is a kernel matrix with $A(0)=I$ (see Section 2), continuous in ϑ on $[0, 1]^r$, and the function $\vartheta_{LS}: [2, \infty) \times \mathscr{S}_t \to [0, 1]^r$ yields pointwise the equation

$$c(N^*, \vartheta_{LS}(N^*, \hat{p}), \hat{p}) = \min_{\vartheta \in [0,1]^r} c(N^*, \vartheta, \hat{p}), \tag{4.2}$$

where the function $c: [2, \infty) \times [0, 1]^r \times \mathscr{S}_t \to \mathbb{R}$ is defined by

$$c(N^*, \vartheta, \hat{p}) = \sum_{i \in J} \hat{p}_i \| A(\vartheta) \hat{p}_{-i} - e_i \|^2 \tag{4.3}$$

with

$$\hat{p}_{-i} = \frac{N^*}{N^*-1} \hat{p} - \frac{1}{N^*-1} e_i. \tag{4.4}$$

Remarks. 1. For each sample size $N \geqslant 2$ the functions $k(N, \cdot)$ imply a kernel estimator with LS-bandwidth. All these estimators have the same kernel matrix. The definition of the estimators is influenced by the sample size N only via the cross-validation criterion; i.e., only via the smoothing parameter.

2. Throughout the paper the bold symbol $\hat{\boldsymbol{p}}$ denotes the random vector of observed cell proportions, whereas the nonbold $\hat{p}$ is used for arbitrary vectors in $\mathscr{S}_t$, including realisations of $\hat{\boldsymbol{p}}$.

3. Some of the theorems below are based on Taylor expansions of the functions k and ϑ_{LS} in both arguments. Hence, we define k and ϑ_{LS} by (4.1) and (4.2), respectively, on the entire convex set $[2, \infty) \times \mathscr{S}_t$. In any practical application, the first argument of a kernel estimator $\hat{k}_{LS} = k(N, \hat{\boldsymbol{p}})$ is always an integer, as N stands for the sample size, and the second argument $\hat{\boldsymbol{p}}$ takes values on a grid in $\mathscr{S}_t$.

4. The existence of ϑ_{LS} is guaranteed (pointwise) by (4.2), as we minimize over a compact domain. Uniqueness is not required!

Throughout the paper we use the following notation for derivatives of functions:

(i) Let $f: \mathbb{R}^n \to \mathbb{R}$ be a real-valued function, $x, x_0 \in \mathbb{R}^n$. Then $\partial_x f(x_0)$ denotes the row vector of the first partial derivatives at the point x_0 and $\partial_{x^2}^2 f(x_0)$ the Jacobian matrix of the second partial derivatives at x_0.

(ii) Let $f: \mathbb{R}^n \to \mathbb{R}^k$ be a vector-valued function, $x = (x_1, ..., x_n)'$, $f(x) = (f_1(x), ..., f_k(x))'$. Then

$$\partial_x f(x_0) = ((\partial_{x_j} f_i(x_0)))_{\substack{i=1,...,k \\ j=1,...,n}}$$

denotes the matrix of first partial derivatives with row index i, column index j.

(iii) In any case, for $l \geqslant 3$, $\partial^l_{x^l} f(x_0)$ denotes the matrix of corresponding partial derivatives. The order of elements within the matrix is not important.

We use the Euclidean matrix norm

$$\| A \| = (\operatorname{tr} A'A)^{1/2}$$

and write $|A|$ for the determinant.

5. Theorems on the Estimator

Let us consider the sequence of kernel estimators $\hat{k}_{\mathrm{LS}} = k(N, \hat{\boldsymbol{p}}) = A(\vartheta_{\mathrm{LS}}(N, \hat{\boldsymbol{p}}))\hat{\boldsymbol{p}}$ given in Section 4; i.e., we investigate kernel estimators, including an r-dimensional random smoothing parameter that minimizes the LS-criterion $c(N, \vartheta, \hat{\boldsymbol{p}})$ on $[0, 1]^r$.

The first three theorems concern one-dimensional bandwidths ($r = 1$). They provide the bias and mean summed squared error of $\hat{k}_{\mathrm{LS}}$ up to terms of order $O(N^{-2})$ as well as the asymptotic distribution. Proofs are long and tedious and are therefore concentrated in Section 7.

We use the following assumptions:

A1. $p \in \operatorname{int} \mathscr{S}_t \setminus \{c_t\}$.

A2. $A(\vartheta)$ is irreducible for $\vartheta > 0$.

A3. A is four times continuously differentiable on $[0, 1]$ and $\| \partial^4_{\vartheta^4} A \|$ is restricted there.

A4. $\| \partial_\vartheta A(0)\, p \| > 0$.

Remark. A nonnegative square matrix A is reducible if and only if, with a certain permutation matrix P,

$$PAP' = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B and D are square matrices. Otherwise A is irreducible. Obviously, all kernel matrices without zeros are irreducible.

Theorem 5.1. *Let $r = 1$. Under the assumptions* A1–A4 *the bias of a kernel estimator satisfies*

$$E\hat{k}_{\mathrm{LS}} - p = \frac{1}{N} b + O(N^{-2}), \tag{5.1}$$

where

$$b = -\frac{\operatorname{tr}\{\partial_\vartheta A(0) \cdot S(p)\}}{\|\partial_\vartheta A(0)\, p\|^2} \cdot \partial_\vartheta A(0)\, p. \tag{5.2}$$

THEOREM 5.2. *Let* $r = 1$. *Under the assumptions* A1–A4

$$\text{(i)} \quad E(\hat{k}_{\mathrm{LS}} - p)(\hat{k}_{\mathrm{LS}} - p)' = \frac{1}{N} S(p) + O(N^{-2}), \tag{5.3}$$

$$\text{(ii)} \quad \delta_{\mathrm{M}}(\hat{\vartheta}_{\mathrm{LS}}) = \frac{1}{N}(1 - \|p\|^2) + O(N^{-2}), \tag{5.4}$$

$$\text{(iii)} \quad \operatorname{var} \hat{k}_{\mathrm{LS}} = \frac{1}{N} S(p) + O(N^{-2}). \tag{5.5}$$

THEOREM 5.3. *Let* $r = 1$. *Under* A1–A4 *kernel estimators have the same asymptotic distribution as the frequency estimator*:

$$\mathscr{L}\{\sqrt{N}(\hat{k}_{\mathrm{LS}} - p)\} \xrightarrow[N \to \infty]{} N_t(0, S(p)). \tag{5.6}$$

Remarks. 1. The assumption A2 is realized, e.g., by the pseudo-Bayesian estimator, the one-parametric estimator of Aitchison and Aitken, and the corresponding nearest neighbour estimators proposed by Aitchison and Aitken [1]; see Section 2. Obviously, all these estimators fulfill A3 and, for $p \in \operatorname{int} \mathscr{S}_t \setminus \{c_t\}$, also A4.

2. The parts (ii) and (iii) of Theorem 5.2 follow immediately from Theorem 5.2(i) and Theorem 5.1.

The above theorems state that kernel estimators with LS-bandwidth have the same mean summed squared error and variance as the frequency estimator (up to terms of order $O(N^{-2})$) and the same asymptotic distribution. The frequency estimator belongs to the class of best asymptotic normally distributed estimators, so the same is true for kernel estimators with LS-bandwidth. Consequently, whatever makes us prefer the kernel estimator for small sample size, we know at least that it behaves well for large samples.

The connection between $\hat{k}_{\mathrm{LS}}$ and the frequency estimator $\hat{\boldsymbol{p}}$ is not surprising, considering that $\hat{\boldsymbol{p}}$ itself is a kernel estimator, obtained for the nonrandom bandwidth $\vartheta = 0$. In the case of random bandwidth procedures, kernel estimators behave like $\hat{\boldsymbol{p}}$ if the random bandwidth vanishes at the speed of $\hat{\vartheta}_N = o_p(N^{-1/2})$, so that $A(\hat{\vartheta}_N) \xrightarrow[N \to \infty]{} I$ fast enough, as shown by

Wang and van Ryzin [27]. In fact, we show in Section 7 that $\hat{\vartheta}_{\mathrm{LS}}$ is of the order $O_p(N^{-1})$; the assumptions A1–A4 are tailored to ensure the right convergence speed.

Assumption A1 excludes the uniform distribution, $p = c_t$, as well as any zero cell probabilities. Heuristically, the uniform distribution is the ideal background for kernel smoothing; the more we smooth the data, the better we estimate the true density. Actually, under $p = c_t$ it may happen that the limiting distribution of a kernel estimator with LS-bandwidth has a smaller variance than $S(p)$, so that $\hat{\vartheta}_{\mathrm{LS}}$ does not even converge with $o_p(N^{-1/2})$. The presence of zero cell probabilities, in the contrary, would call for a fast convergence rate of $\hat{\vartheta}_{\mathrm{LS}}$, as the mean summed squared error of $\hat{\boldsymbol{p}}$ is rather small on the boundary of $\mathscr{S}_t$. Nevertheless, the assumption $p \in \operatorname{int} \mathscr{S}_t$ is more a technical convenience, while $p \neq c_t$ is essential for all three theorems.

The convergence rate of the LS-bandwidth depends also on the shape of the kernel matrix $A(\vartheta)$ as a function of ϑ. Assumption A2 implies under any cpv $p \neq c_t$ that $A(\vartheta)p = p$ yields only for $\vartheta = 0$; it is introduced to ensure that the convergence of $\hat{\boldsymbol{p}} \xrightarrow[N \to \infty]{} p$ inevitably results in $\hat{\vartheta}_{\mathrm{LS}} \xrightarrow[N \to \infty]{} 0$. Further, violation of A4 would mean that the value $A(\vartheta)p$ could change only little in the neighborhood of $\vartheta = 0$, so that the root-N convergence of $|\hat{\boldsymbol{p}} - p| = O_p(N^{-1/2})$ might not be sufficient to force the rate $\hat{\vartheta}_{\mathrm{LS}} = O_p(N^{-1})$. On the other hand, the uniform bounds on $\| \partial_{\vartheta^4}^4 A \|$ in A3 prevent a faster convergence rate. As technical tools, assumptions A2–A4 provide the differentiability of ϑ_{LS} via the theorem on implicit functions.

THEOREM 5.4. *Let* $r = 1$. *Under* A1–A4

$$\text{(i)} \quad N\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}}) \xrightarrow[N \to \infty]{d.} U'S(p)\,U, \tag{5.7}$$

where $U \sim N_t(0, I)$ *is a t-dimensional standard normally distributed random vector,*

$$\text{(ii)} \quad N\{\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}}) - \delta_{\mathrm{A}}(0)\} \xrightarrow[N \to \infty]{a.s.} 0. \tag{5.8}$$

Remarks. 1. According to our definition every kernel estimator coincides with the frequency estimator for $\vartheta = 0$. Theorem 5.2 provides

$$\delta_{\mathrm{M}}(\hat{\vartheta}_{\mathrm{LS}}) - \delta_{\mathrm{M}}(0) = O(N^{-2}),$$

while (5.8) is weaker.

2. In our considerations $p = c_t$ is excluded, but for $p \approx c_t$ the distribution of $U'S(p)\,U$ can be approximated by a χ^2-distribution:

$$U'S(c_t)\,U \sim \frac{1}{t} \cdot \chi^2_{t-1}.$$

The results of Theorem 5.1–5.3 can be generalised to higher-dimensional smoothing parameters using stronger assumptions:

A5. There are a neighbourhood $U(p)$ of p and an integer N_0 such that

$$\sup_{\hat{p} \in U(p)} \| \vartheta_{\mathrm{LS}}(N, \hat{p}) \| \xrightarrow[N \to \infty]{} 0$$

and $\vartheta_{\mathrm{LS}}(N, \hat{p}) \in (0, 1)^r$ for all $N \geqslant N_0$, $\hat{p} \in U(p)$.

A6. A is four times continuously differentiable on $[0, 1]^r$ and the fourth derivative is bounded there.

A7. $$|(((\partial_{\vartheta_\nu} A(0)\,p)'\,\partial_{\vartheta_\mu} A(0)\,p))_{\substack{\nu=1,\ldots,r \\ \mu=1,\ldots,r}}| > 0.$$

THEOREM 5.5. *For any $r \geqslant 1$ the assumptions* A5–A7 *supply*

(i) $E\hat{k}_{\mathrm{LS}} - p = O(N^{-1})$,

(ii) (5.3)–(5.6) *are valid.*

The proof is similar to that in the one-dimensional case and is therefore not given here.

Remark. Condition A5 is a strong restriction on the cpv p for $r \geqslant 2$. So we find that the assumption $\vartheta_{\mathrm{LS}}(N, p) \in (0, 1)^r$ is not fulfilled in the case of the higher-dimensional estimator of Aitchison and Aitken (see (2.5)) for an open subset of cpv's unless $t_1 = \cdots = t_m$. Ways of avoiding A5 are proposed in Grund [9].

Finally, we comment on consistency. The results of Wang and van Ryzin [27] ensure the consistency of a kernel estimator provided the data-dependent bandwidth converges to zero fast enough. As part of the proofs in Section 7 we show that the convergence rate of $\hat{\vartheta}_{\mathrm{LS}}$ meets the requirements in Wang and van Ryzin [27], given that A1–A4 are valid. Nevertheless, we can state without these somewhat restrictive assumptions:

THEOREM 5.6.

$$\hat{k}_{\mathrm{LS}} \xrightarrow[N \to \infty]{a.s.} p.$$

The consistency of kernel estimators with Kullback–Leibler bandwidth was shown already by Bowman [3]. Theorem 5.6 could be proved in the same way, using

$$c(N, \vartheta, \hat{p}) = \| A(\vartheta)\hat{p} - \hat{p} \|^2 + 1 - \| \hat{p} \|^2 + R_1(N, \vartheta, \hat{p}) \tag{5.9}$$

with

$$\sup_{(\vartheta, \hat{p}) \in [0,1]^r \times \mathscr{S}_r} | R_1(N, \vartheta, \hat{p}) | = O(N^{-1}). \tag{5.10}$$

6. Theorems on the Bandwidth

First we compare the LS-bandwidth with the optimal one using simple kernels, namely the pseudo-Bayes and the one-dimensional Aitchison and Aitken estimators:

A8. $A(\vartheta)$ is defined by (2.2) or (2.4) with $t_1 = \cdots = t_m = 2$.

Theorem 6.1. *Under the assumptions* A1 *and* A8

$$\frac{\hat{\vartheta}_{\mathrm{LS}}}{\vartheta_{\mathrm{opt}}} \xrightarrow[N \to \infty]{P} 1.$$

Remark. Regarding Aitchison and Aitken's estimator Hall [12] pointed out that the Kullback–Leibler bandwidth $\hat{\vartheta}_{\mathrm{KL}}$ had the same convergence rate $O(N^{-1})$ as $\hat{\vartheta}_{\mathrm{LS}}$, but $\hat{\vartheta}_{\mathrm{KL}}/\vartheta_{\mathrm{opt}}$ did not converge to 1. The better behaviour of the LS-method w.r.t. quadratic risk is no surprise, as $\hat{\vartheta}_{\mathrm{LS}}$ minimizes an estimation of the quadratic risk of $A(\vartheta)\hat{\boldsymbol{p}}$, while $\hat{\vartheta}_{\mathrm{KL}}$ corresponds to Kullback–Leibler loss.

Theorem 6.2. *Under* A1–A4

$$\text{(i)} \quad 1 - \frac{\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{opt}})}{\delta_{\mathrm{A}}(0)} = \hat{\gamma} + O_p(N^{-1/2}), \tag{6.1}$$

$$\text{(ii)} \quad 1 - \frac{\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{opt}})}{\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}})} = \hat{\gamma} + O_p(N^{-1/2}), \tag{6.2}$$

where

$$\hat{\gamma} = \frac{[(\hat{\boldsymbol{p}} - p)' \, \partial_\vartheta A(0)\hat{\boldsymbol{p}}]^2}{\| \partial_\vartheta A(0)\hat{\boldsymbol{p}} \|^2 \, \| \hat{\boldsymbol{p}} - p \|^2} \cdot I_{\{\hat{\vartheta}_{\mathrm{opt}} > 0\}} \tag{6.3}$$

and

$$\hat{\gamma} \not\xrightarrow[N\to\infty]{P} 0, \tag{6.4}$$

yielding

$$0 \leqslant \hat{\gamma} \leqslant 1. \tag{6.5}$$

Remark. Observe that $\hat{\vartheta}_{\text{opt}}$ depends on the unknown cpv p, and if the sample $\hat{\boldsymbol{p}}$ lies within a certain neighbourhood of p, the maximum likelihood estimator of $\hat{\vartheta}_{\text{opt}}$ is 0. Theorem 6.2(ii) confirms that the LS-method is suboptimal if we are interested in the actual summed squared error (in distinction to referring to the mean summed squared error). However, it is comforting to know that $\hat{\vartheta}_{\text{LS}}$ is not worse than the maximum likelihood estimator of $\hat{\vartheta}_{\text{opt}}$ (asymptotically).

7. Proofs

Sketch of the Proof of Theorem 5.1. We use the decomposition

$$k(N, \hat{\boldsymbol{p}}) - p = \Delta_1 + \Delta_2,$$

where

$$\Delta_1 = k(N, \hat{\boldsymbol{p}}) - k(N, p), \tag{7.1}$$

$$\Delta_2 = k(N, p) - p, \tag{7.2}$$

check $E\Delta_1 = O(N^{-2})$ with the aid of Taylor series in $\hat{p}$, and compute Δ_2 up to terms of order $O(N^{-2})$ regarding something like Taylor expansions in N^{-1}.

The main problem is handling ϑ_{LS}. Differentiability conditions and derivatives of ϑ_{LS} are summarized in Appendix B. Lemmas 7.1–7.3 ensure that the assumptions of Appendix B are valid, while Lemma 7.4 helps to restrict $E\Delta_1$. Proofs of the lemmas are sketched in Appendix A.

Note that in some cases we omit arguments of functions. Then A and its derivatives are to be computed at the point $\vartheta_{\text{LS}}(N, p)$, as well as ϑ_{LS} and its derivatives at (N, p).

Lemma 7.1. *Under* A1 *and* A2

(i) $$\sup_{\hat{p} \in \mathscr{S}_l} \| A(\vartheta_{\text{LS}}(N, \hat{p}))\, \hat{p} - \hat{p} \|^2 = O(N^{-1}), \tag{7.3}$$

(ii) *there is a neighbourhood* $U(p)$ *such that*

$$\sup_{\hat{p} \in U(p)} \vartheta_{\text{LS}}(N, \hat{p}) \xrightarrow[N\to\infty]{} 0. \tag{7.4}$$

LEMMA 7.2. *Under* A1–A4 *there are a neighbourhood* $U(p)$ *and an integer* N_0 *such that*

$$\vartheta_{\mathrm{LS}}(N, \hat{p}) \in (0, 1) \qquad \textit{for all} \quad N \geqslant N_0 \qquad \textit{and} \qquad \hat{p} \in U(p).$$

LEMMA 7.3. *Under* A1–A4 *there are a neighbourhood* $U(p)$ *and an integer* N_0 *such that*

$$\inf_{N \geqslant N_0} \inf_{\hat{p} \in U(p)} \partial^2_{\vartheta^2} c(N, \vartheta_{\mathrm{LS}}(N, \hat{p}), \hat{p}) > 0.$$

LEMMA 7.4. *Under* A1-A4

$$A(\vartheta_{\mathrm{LS}}(N, p)) - I = O(N^{-1}).$$

Proof of Theorem 5.1. Let $U(p)$ and N_0 be chosen according to Lemmas 7.1–7.3, w.l.o.g. we assume $N \geqslant N_0$.

First, we consider Δ_1, given by (7.1). Applying Appendix C to $k(N, \cdot)$ we obtain

$$E\Delta_1 = \frac{1}{2N} ((\mathrm{tr}\{S(p) \cdot \partial^2_{\hat{p}^2} k_{\mathbf{i}}(N, p)\}))_{\mathbf{i}} + O(N^{-2}). \tag{7.5}$$

Thereby the assumptions of Appendix C are guaranteed by Appendix B:

..... The function $\tilde{\vartheta}_{\mathrm{LS}}$ (defined by (A.3)) is three times continuously differentiable on $(0, N_0^{-1}] \times U(p)$ and $\partial^3_{\hat{p}^3} \tilde{\vartheta}_{\mathrm{LS}}$ is uniformly bounded there. Consequently, the same holds for ϑ_{LS} and k on $[N_0, \infty) \times U(p)$.

— The properties of $\tilde{\vartheta}_{\mathrm{LS}}$ follow from Appendix B (cf. proof of Lemma 7.4).

The derivatives in (7.5) are

$$\begin{aligned} \partial^2_{\hat{p}^2} k_{\mathbf{i}}(N, p) = {} & w_2 \cdot (\partial_\vartheta A_{(\mathbf{i}\cdot)})\, p + w_1' \partial_\vartheta A_{(\mathbf{i}\cdot)} + (\partial_\vartheta A_{(\mathbf{i}\cdot)})'\, w_1 \\ & + (\partial^2_{\vartheta^2} A_{(\mathbf{i}\cdot)})\, p \cdot w_1' w_1, \end{aligned} \tag{7.6}$$

where $A_{(\mathbf{i}\cdot)}$ denotes the $\mathbf{i}$th row of A, and

$$w_1 = \partial_{\hat{p}} \vartheta_{\mathrm{LS}} = \frac{1}{\alpha} \{p' A'\, \partial_\vartheta A p \mathbf{1}\} + O(N^{-1}), \tag{7.7}$$

$$w_2 = \partial^2_{\hat{p}^2} \vartheta_{\mathrm{LS}} = \frac{1}{\alpha} \{\Gamma + \beta \cdot w_1' w_1 + d' w_1 + w_1' d\} + O(N^{-1}), \tag{7.8}$$

$$\alpha = \| \partial_\vartheta A p \|^2 + O(N^{-1}),$$

$$\beta = 3p'(\partial_\vartheta A)'\, \partial^2_{\vartheta^2} A p + O(N^{-1}),$$

$$d = 2p'(\partial_\vartheta A)'\, \partial_\vartheta A + \{p' A'\, \partial_\vartheta A p + \| \partial_\vartheta A p \|^2\}\, \mathbf{1}' + O(N^{-1}),$$

$$\Gamma = \mathbf{1} p' \{A'(\partial_\vartheta A) + (\partial_\vartheta A)'\, A\} + \{A'\, \partial_\vartheta A + (\partial_\vartheta A)'\, A\}\, p \mathbf{1}' + O(N^{-1})$$

with $\mathbf{1} = (1, ..., 1) \in \mathbb{R}^l$. The shape of the derivatives of ϑ_{LS} also results from Appendix B, setting $\psi(\vartheta, \hat{p}) = A(\vartheta)\hat{p}$, $\varepsilon = N^{-1}$, $p_{\varepsilon} = p + O(N^{-1})$ and applying Lemma 7.4.

Recall that $S(p)\mathbf{1} = 0$. Inserting (7.6)–(7.8) into (7.5) we obtain

$$E\Delta_1 = O(N^{-2}). \tag{7.9}$$

Second, we investigate Δ_2 given by (7.2). Let us consider the function $\tilde{k}: [0, \frac{1}{2}] \to \mathscr{S}_l$ defined by $\tilde{k}(\varepsilon) = A(\tilde{\vartheta}_{LS}(\varepsilon, p))\, p$. Obviously, $\tilde{k}(0) = A(0)\, p = p$. A Taylor expansion of $\tilde{k}$ about $\varepsilon = 0$ provides

$$\tilde{k}(\varepsilon) - p = \varepsilon \cdot \partial_{\varepsilon}\tilde{k}(0) + \tfrac{1}{2} \cdot \varepsilon^2 \cdot \partial_{\varepsilon^2}^2 \tilde{k}(\varepsilon^*) \tag{7.10}$$

for a certain $\varepsilon^* \in (0, \varepsilon)$. Thereby,

$$\partial_{\varepsilon}\tilde{k}(0) = \partial_{\vartheta}A(0)\, p \cdot \partial_{\varepsilon}\tilde{\vartheta}_{LS}(0, p), \tag{7.11}$$

$$\partial_{\varepsilon}\tilde{\vartheta}_{LS}(0, p) = -\frac{\operatorname{tr}\{\partial_{\vartheta}A(0) \cdot S(p)\}}{\| \partial_{\vartheta}A(0)\, p \|^2} \tag{7.12}$$

and

$$\sup_{\varepsilon \in (0, N_o^{-1}]} \| \partial_{\varepsilon^2}^2 \tilde{k}(\varepsilon) \| < \infty. \tag{7.13}$$

The differentiability of $\tilde{k}$ and $\tilde{\vartheta}_{LS}$ as well as (7.12)–(7.13) results from Appendix B; cf. the proof of Lemma 7.4.

Note that $\tilde{k}$ and $\tilde{\vartheta}_{LS}$ satisfy $k(N, p) = \tilde{k}(N^{-1})$. We put (7.11) and (7.12) into (7.10), set $\varepsilon = N^{-1}$, and obtain with (7.13) and (5.2)

$$\Delta_2 = \frac{1}{N} b + O(N^{-2}). \tag{7.14}$$

Result (5.1) follows directly from (7.14) and (7.9).

Proof of Theorem 5.2. Let us consider again $U(p)$ and N_o chosen according to Lemma 7.1–7.4, and Δ_1 and Δ_2 defined by (7.1) and (7.2), respectively. Taking into account $E\Delta_1 = O(N^{-2})$ and $\Delta_2 = O(N^{-1})$ we obtain

$$E(k(N, \hat{\boldsymbol{p}}) - p)(k(N, \hat{\boldsymbol{p}}) - p)' = E\Delta_1\Delta_1' + O(N^{-2}). \tag{7.15}$$

Appendix C(ii) applied to $k(N, \cdot)$ provides

$$E\Delta_1\Delta_1' = \frac{1}{N}\, \partial_{\hat{p}}k(N, p)\, S(p)(\partial_{\hat{p}}k(N, p))' + O(N^{-2}). \tag{7.16}$$

Combining

$$\partial_{\hat{p}} k(N, p) = (\partial_{\vartheta} A)\, p \cdot \partial_{\hat{p}} \vartheta_{\mathrm{LS}} + A,$$

Lemma 7.4 and (7.7) we obtain

$$\partial_{\hat{p}} k(N, p) = -\frac{p' A' (\partial_{\vartheta} A)\, p}{\| \partial_{\vartheta} A p \|^2} \cdot (\partial_{\vartheta} A)\, p \mathbf{1}' + I + O(N^{-1}). \tag{7.17}$$

The assertion follows with $S(p)\,\mathbf{1} = 0$ from (7.15)–(7.17).

Proof of Theorem 5.3. Let us consider again $U(p)$, N_o, Δ_1, and Δ_2 from the previous proof. We now apply Appendix C(iii) to $k(N, \cdot)$ and obtain

$$\mathscr{L}\{\sqrt{N}\Delta_1\} \xrightarrow[N \to \infty]{} N_t(0, B\, S(p)\, B'), \tag{7.18}$$

where $B = \lim_{N \to \infty} \partial_{\hat{p}} k(N, p)$. Checking the assumptions of Appendix C, note that

— the differentiability conditions on k are validated in the proof of Theorem 5.1,

— the convergence of the sequence $\{\partial_{\hat{p}} k(N, p)\}$ results from (7.17) considering A3, A4, and Lemma 7.1. So we find $\lim_{N \to \infty} \| \partial_{\vartheta} A p \| = \| \partial_{\vartheta} A(0)\, p \| > 0$.

Formulae (7.17), A3, A4, and Lemma 7.1 imply

$$B = -\frac{p' A'\, \partial_{\vartheta} A(0)\, p}{\| \partial_{\vartheta} A(0)\, p \|^2}\, \partial_{\vartheta} A(0)\, p \mathbf{1}' + I \tag{7.19}$$

and, consequently,

$$B\, S(p)\, B' = S(p). \tag{7.20}$$

The assertion follows with $\Delta_2 = O(N^{-1})$ from (7.18).

Proof of Theorem 5.4. We consider the same $U(p)$, N_o, Δ_1 and Δ_2.

(i) From our definition $\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}}) = \| k(N, \hat{\boldsymbol{p}}) - p \|^2 = \| \Delta_1 + \Delta_2 \|^2$, we obtain with (7.14)

$$N \delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}}) = N \| \Delta_1 \|^2 + 2 b' \Delta_1 + O(N^{-1}). \tag{7.21}$$

The assertion follows from (7.21) considering (7.18) and (7.20).

(ii) Combining (7.21) and (7.18) we get both

$$N\{\delta_{\mathrm{A}}(\hat{\vartheta}_{\mathrm{LS}}) - \delta_A(0)\} = N\{\| \Delta_1 \|^2 - \| \hat{\boldsymbol{p}} - p \|^2\} + O_p(N^{-1/2}) \tag{7.22}$$

and

$$\Delta_1 + (\hat{\boldsymbol{p}} - p) = O_p(N^{-1/2}). \tag{7.23}$$

Let us now consider $N\{\Delta_1 - (\hat{\boldsymbol{p}} - p)\}$. Since $U(p)$ and N_o were chosen such that $k(N, \cdot)$ is differentiable on $U(p)$ for $N \geqslant N_o$ and $\partial^2_{\hat{p}^2} k(N, \hat{p})$ is uniformly bounded on $[N_o, \infty) \times U(p)$ (cf. proof of Theorem 5.1), a Taylor expansion of $k(N, \cdot)$ provides

$$\begin{aligned}\Delta_1 - (\hat{\boldsymbol{p}} - p) = \{(\partial_{\hat{p}} k(N, p) - I)(\hat{\boldsymbol{p}} - p) + O_p(N^{-1})\}\, I_{\{\hat{\boldsymbol{p}} \in U(p)\}} \\ + R_2(N, \hat{\boldsymbol{p}}),\end{aligned} \tag{7.24}$$

where $R_2(N, \hat{\boldsymbol{p}}) = \{\Delta_1 - (\hat{\boldsymbol{p}} - p)\}(1 - I_{\{\hat{\boldsymbol{p}} \in U(p)\}})$. It is well known that the convergence rate $1 - I_{\{\hat{\boldsymbol{p}} \in U(p)\}} = O_p(N^{-2})$ yields. Hence, we obtain with (7.24), (7.17), and $\mathbf{1}'(\hat{\boldsymbol{p}} - p) \equiv 0$ the convergence speed $N\{\Delta_1 - (\hat{\boldsymbol{p}} - p)\} = O_p(1)$. The assertion follows with (7.23) and (7.22).

Proof of Theorem 6.1. It is sufficient to show that

$$\frac{\vartheta_{\mathrm{LS}}(N, \hat{\boldsymbol{p}}) - \vartheta_{\mathrm{LS}}(N, p)}{\vartheta_{\mathrm{opt}}} \xrightarrow[N \to \infty]{P} 0 \tag{7.25}$$

and

$$\frac{\vartheta_{\mathrm{LS}}(N, p)}{\vartheta_{\mathrm{opt}}} \xrightarrow[N \to \infty]{} 1. \tag{7.26}$$

Assumptions A1 and A8 ensure that the assumptions of Theorem 5.1 are fulfilled, and we choose again $U(p)$ and N_o according to Lemmas 7.1–7.4. Similar to the considerations concerning (7.24) we get

$$\begin{aligned}\vartheta_{\mathrm{LS}}(N, \hat{\boldsymbol{p}}) - \vartheta_{\mathrm{LS}}(N, p) = \partial_{\hat{p}} \vartheta_{\mathrm{LS}} \cdot (\hat{\boldsymbol{p}} - p) + \tfrac{1}{2} (\hat{\boldsymbol{p}} - p)' \, \partial^2_{\hat{p}^2} \vartheta_{\mathrm{LS}} (\hat{\boldsymbol{p}} - p) \\ + o_p(N^{-1}),\end{aligned} \tag{7.27}$$

and from (7.27) we obtain with (7.7), (7.8), $\mathbf{1}'(\hat{\boldsymbol{p}} - p) \equiv 0$, and Lemma 7.4

$$\vartheta_{\mathrm{LS}}(N, \hat{\boldsymbol{p}}) - \vartheta_{\mathrm{LS}}(N, p) = o_p(N^{-1}). \tag{7.28}$$

For both, the one-parameter Aitchison and Aitken estimator and the pseudo-Bayes estimator ϑ_{opt} yield

$$\vartheta_{\mathrm{opt}} = -\frac{1}{N} \frac{\mathrm{tr}\{\partial_\vartheta A(0) \cdot S(p)\}}{\| \partial_\vartheta A(0)\, p \|^2} + o(N^{-1}) \tag{7.29}$$

(according to Hall [12] and Sutherland *et al.* [25]). Since the dominant term is positive under A1, (7.25) follows from (7.28) and (7.29).

To prove (7.26) we develop a Taylor expansion of $\tilde{\vartheta}_{LS}(\cdot, p)$, defined by (A.3), about zero. Using (7.12) we obtain

$$\vartheta_{LS}(N, p) = -\frac{1}{N} \frac{\operatorname{tr}\{\partial_\vartheta A(0) \cdot S(p)\}}{\| \partial_\vartheta A(0)\, p \|^2} + O(N^{-2}). \tag{7.30}$$

Result (7.26) follows immediately.

Proof of Theorem 6.2. (i) Since $\hat{\vartheta}_{opt}$ minimizes $\delta_A(\vartheta)$, we obtain

$$\delta_A(0) - \delta_A(\hat{\vartheta}_{opt}) = \tfrac{1}{2}\{\hat{\vartheta}^2_{opt} \partial^2_{\vartheta^2} \delta_A(\hat{\vartheta}_{opt}) + o(\hat{\vartheta}^2_{opt})\} I_{\{\hat{\vartheta}_{opt} > 0\}} + o_p(N^{-1}), \tag{7.31}$$

taking into consideration, for $p \neq c_t$,

$$P(\hat{\vartheta}_{opt} = 1) \leqslant P(\| \hat{\boldsymbol{p}} - p \| \geqslant \| c_t - p \|) = o(N^{-1}). \tag{7.32}$$

Taylor expansions of $\partial_\vartheta \delta_A$ and $\partial^2_{\vartheta^2} \delta_A$ imply

$$\hat{\vartheta}_{opt} + o(\hat{\vartheta}_{opt}) = -\partial_\vartheta \delta_A(0) / \partial^2_{\vartheta^2} \delta_A(0) \tag{7.33}$$

and

$$\partial^2_{\vartheta^2} \delta_A(\hat{\vartheta}_{opt}) = \partial^2_{\vartheta^2} \delta_A(0) + o(\hat{\vartheta}_{opt}), \tag{7.34}$$

where

$$\partial_\vartheta \delta_A(0) = 2(\hat{\boldsymbol{p}} - p)' \, \partial_\vartheta A(0)\, \hat{\boldsymbol{p}}, \tag{7.35}$$

$$\partial^2_{\vartheta^2} \delta_A(0) = 2\{\| \partial_\vartheta A(0)\, \hat{\boldsymbol{p}} \|^2 + (\hat{\boldsymbol{p}} - p)' \, \partial^2_{\vartheta^2} A(0)\, \hat{\boldsymbol{p}}\} \tag{7.36}$$

and, consequently,

$$\hat{\vartheta}_{opt} = O_p(N^{-1/2}). \tag{7.37}$$

Combining (7.31)–(7.37) we obtain (6.1). Result (6.5) follows immediately from the Cauchy–Schwarz inequality.
It remains to show that

$$\frac{[(\hat{\boldsymbol{p}} - p)' \, \partial_\vartheta A(0)\, \hat{\boldsymbol{p}}]^2}{\| \partial_\vartheta A(0)\, \hat{\boldsymbol{p}} \|^2 \, \| \hat{\boldsymbol{p}} - p \|^2} \xrightarrow{P} 0 \tag{7.38}$$

and

$$I_{\{\hat{\vartheta}_{opt} > 0\}} \xrightarrow{P} 0. \tag{7.39}$$

Assume (7.38) is not true. Then the angle between $(\hat{\boldsymbol{p}} - p)$ and $\partial_\vartheta A(0)\, p$ has to converge to $\pi/2$, which contradicts (2.1).

To prove (7.39) consider the eigenvectors $h_1, ..., h_t$ and the eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_t$ of $A(\vartheta)$. Properties of irreducible double-stochastic matrices provide

$$h_1 = t^{-1/2}\mathbf{1}, \qquad \lambda_1 = 1, \qquad |\lambda_k| < 1 \qquad \text{for} \quad k = 2, ..., t,$$

and, consequently, there is a representation

$$p = c_t + \sum_{k=2}^{t} \beta_k h_k,$$

$$\hat{p} = c_t + \sum_{k=2}^{t} \hat{\beta}_k h_k.$$

Note that λ_k, h_k, β_k, $\hat{\beta}_k$ all depend on ϑ, whereby

$$\vartheta \to 0 \Rightarrow \lambda_k \to 1 \qquad \text{for all} \quad k = 2, ..., t.$$

We obtain

$$\begin{aligned} P(\hat{\vartheta}_{\text{opt}} > 0) &= P(\exists \vartheta > 0 : \|\hat{p} - p\|^2 > \|A(\vartheta)\hat{p} - p\|^2) \\ &= P\left(\bigcup_{\vartheta > 0} \left\{ \sum_{k=2}^{t} (\hat{\beta}_k - \beta_k)^2 > \sum_{k=2}^{t} (\lambda_k \hat{\beta}_k - \beta_k)^2 \right\} \right) \\ &\geqslant \sup_{\vartheta > 0} P\left(\sum_{k=2}^{t} (\hat{\beta}_k - \beta_k)^2 > \sum_{k=2}^{t} (\lambda_k \hat{\beta}_k - \beta_k)^2 \right) \\ &\geqslant \inf_{\{h_k\}} P((\hat{\beta}_k - \beta_k)\beta_k > 0 \qquad \text{for all} \quad k = 2, ..., t). \end{aligned} \tag{7.40}$$

Formula (7.40) describes the probability of a certain quadrant of the "least favourable" coordinate system, choosing among all those with origin p and axes according to any eigenvector system $\{h_k\}_{k=2,...,t}$. The right-hand side of (7.40) is bounded away from zero with increasing N because of (2.1). The assertion follows at once.

Result (ii) follows from (6.1) considering (5.8).

A. Appendix

Proof of Lemma 7.1. (i) Recall (5.9). Since $\vartheta_{LS}(N, \hat{p})$ is defined to minimize $c(N, \vartheta, \hat{p})$, we obtain

$$\| A(\vartheta_{LS}(N, \hat{p}))\, \hat{p} - \hat{p} \|^2 \leqslant \| A(\vartheta)\, \hat{p} - \hat{p} \|^2 + 2 \sup_{\vartheta \in [0, 1]} | R_1(N, \vartheta, \hat{p}) | \quad (A.1)$$

for any $\vartheta \in [0, 1]$. The assertion follows from (A.1) for $\vartheta = 0$, as $A(0) = I$.

(ii) Let $U(p)$ be any neighbourhood of p with the closure $\overline{U(p)} \subseteq \mathscr{S}_t \setminus \{c_t\}$. We prove (7.4) indirectly.

Assume there are a number $\varepsilon > 0$ and a sequence $\{(N_n, \hat{p}_n)\}$ with $N_n + 1 < N_{n+1}$ such that

$$\hat{p}_n \in U(p) \qquad \text{and} \qquad \vartheta_{LS}(N_n, \hat{p}_n) \geqslant \varepsilon \qquad \text{for all} \quad n = 1, 2, \ldots .$$

W.l.o.g. let $\lim_{n \to \infty} \hat{p}_n = p^*$. Using (7.3) we get

$$\limsup_{n \to \infty} \| A(\vartheta_{LS}(N_n, \hat{p}_n))\, p^* - p^* \| = 0$$

and therefore

$$\| A(\vartheta^*)\, p^* - p^* \| = 0 \quad (A.2)$$

for every accumulation point ϑ^* of the sequence $\{\vartheta_{LS}(N_n, \hat{p}_n)\}$.

It remains to show that (A.2) implies $\vartheta^* = 0$:

> Since $A(\vartheta)$ is irreducible for $\vartheta > 0$, c_t is the only eigenvector of $A(\vartheta)$ in $\mathscr{S}_t$, and the choice of $U(p)$ guarantees $p^* \neq c_t$. Therefore, (A.2) ensures $A(\vartheta^*) = I$ and $\vartheta^* = 0$.

Proof of Lemma 7.2. According to Lemma 7.1 there are $U(p)$ and N_o such that

$$\vartheta_{LS}(N, \hat{p}) < 1 \qquad \text{for all} \quad N \geqslant N_o, \hat{p} \in U(p).$$

In the following we consider

$$\Delta(\vartheta) = c(N, \vartheta, \hat{p}) - c(N, 0, \hat{p})$$

for any $N \geqslant 2$, $\hat{p} \in \operatorname{int} \mathscr{S}_t$, and show: There exists a $\vartheta^* > 0$ with $\Delta(\vartheta^*) < 0$. Inserting the Taylor series

$$A(\vartheta) = A(0) + \vartheta H + \vartheta^2\, \partial^2_{\vartheta^2} A(\tilde{\vartheta})$$

with $H = \partial_\vartheta A(0)$ and a certain $\tilde{\vartheta} \in [0, 1]$ into (4.3) we obtain

$$\Delta(\vartheta) = \vartheta\tau(\vartheta) - \vartheta \frac{2N}{(N-1)^2} \left\{ \hat{p}' H \hat{p} - \sum_{i \in J} \hat{p}_i h_{ii} \right\},$$

τ being a continuous function with $\tau(\vartheta) \xrightarrow[\vartheta \to 0]{} 0$.

The properties of $A(\vartheta)$ as a kernel matrix together with $A(\vartheta) \xrightarrow[\vartheta \to 0]{} I$ imply that $\mathbf{1}'H = 0$ and that the diagonal elements h_{ii} of H are nonpositive and that other elements are nonnegative. For each $\hat{p} \in \operatorname{int} \mathscr{S}_t$ it follows with A4 that

$$\hat{p}' H \hat{p} > \sum_{i \in J} \hat{p}_i h_{ii},$$

which had to be shown.

Proof of Lemma 7.3. Computing derivatives in a straightforward way and keeping in mind A3 and Lemma 7.1(i) we obtain, at the point $\vartheta = \vartheta_{LS}(N, \hat{p})$,

$$\partial^2_{\vartheta^2} c(N, \vartheta_{LS}(N, \hat{p}), \hat{p}) = \| \partial_\vartheta A(\vartheta_{LS}(N, \hat{p}))\, \hat{p} \|^2 + R_2(N, \hat{p}),$$

where $\sup_{\hat{p} \in \mathscr{S}_t} |R_2(N, \hat{p})| = O(N^{-1/2})$. The assertion follows immediately using A4, Lemma 7.1(ii), and A3.

Proof of Lemma 7.4. Let N_o and $U(p)$ be chosen according to Lemmas 7.1–7.3. Let us consider the function $\tilde{\vartheta}_{LS}: [0, \frac{1}{2}] \times \mathscr{S}_t \to [0, 1]$, defined by

$$\tilde{\vartheta}_{LS}(\varepsilon, \hat{p}) = \begin{cases} \vartheta_{LS}(\varepsilon^{-1}, \hat{p}), & \text{if } (\varepsilon, \hat{p}) \in (0, \frac{1}{2}] \times \mathscr{S}_t \\ 0 & \text{else.} \end{cases} \tag{A.3}$$

From Appendix B we conclude that (setting $\psi(\vartheta, \hat{p}) = A(\vartheta)\hat{p}$)

$$\tilde{\vartheta}_{LS}(\cdot, \hat{p}) \text{ is differentiable on } [0, N_0^{-1}] \text{ for all } \hat{p} \in U(p) \text{ and}$$

$$\sup_{\varepsilon \in [0, N_o^{-1}]} |\partial_\varepsilon \tilde{\vartheta}_{LS}(\varepsilon, p)| < \infty, \tag{A.4}$$

where the assumptions of Appendix B are ensured by Lemmas 7.1–7.3, A1, and A3. According to our definition we have $A(\tilde{\vartheta}_{LS}(0, p)) = A(0) = I$. A Taylor expansion of $A(\tilde{\vartheta}_{LS}(\cdot, p))$ about $\varepsilon = 0$ provides

$$A(\tilde{\vartheta}_{LS}(\varepsilon, p)) - I = \varepsilon \cdot \partial_\vartheta A(\tilde{\vartheta}_{LS}(\varepsilon^*, p)) \cdot \partial_\varepsilon \tilde{\vartheta}_{LS}(\varepsilon^*, p), \tag{A.5}$$

where $\varepsilon^* \in (0, \varepsilon)$. With respect to A3 and (A.4) we obtain from (A.5)

$$A(\tilde{\vartheta}_{LS}(\varepsilon, p)) - I = O(\varepsilon).$$

The assertion follows with (A.3), taking $\varepsilon = N^{-1}$.

B. Appendix

Let $p \in \operatorname{int} \mathscr{S}_t$, and let $\tilde{U}(p)$ be a neighbourhood of p such that the function $\psi: [0, 1] \times \mathscr{S}_t \to \mathscr{S}_t$ is four times continuously differentiable on $[0, 1] \times \tilde{U}(p)$.

Let the neighbourhood $U(p) \subset \tilde{U}(p)$ and $\varepsilon_0 > 0$ be chosen such that $\hat{p}_{-\mathbf{i}} \in \tilde{U}(p)$ is ensured for all $(\varepsilon, \hat{p}) \in (0, \varepsilon_0] \times U(p)$ and $\mathbf{i} \in J$, where

$$\hat{p}_{-\mathbf{i}} = \frac{\varepsilon^{-1}}{\varepsilon^{-1} - 1} \hat{p} - \frac{1}{\varepsilon^{-1} - 1} e_{\mathbf{i}}.$$

Let the function $\vartheta_{\mathrm{LS}}: [0, \frac{1}{2}] \times \mathscr{S}_t \to [0, 1]$ fulfill the following conditions:

a1. $$c(\varepsilon, \vartheta_{\mathrm{LS}}(\varepsilon, \hat{p}), \hat{p}) = \min_{\vartheta \in [0, 1]} c(\varepsilon, \vartheta, \hat{p})$$

for all $(\varepsilon, \hat{p}) \in (0, \varepsilon_0] \times U(p)$, where

$$c(\varepsilon, \vartheta, \hat{p}) = \sum_{\mathbf{i} \in J} \hat{p}_{\mathbf{i}} \| \psi(\vartheta, \hat{p}_{-\mathbf{i}}) - e_{\mathbf{i}} \|^2$$

and

$$\vartheta_{\mathrm{LS}}(0, \hat{p}) = 0 \qquad \text{for all} \quad \hat{p} \in U(p).$$

a2. $$\vartheta_{\mathrm{LS}}(\varepsilon, \hat{p}) \in (0, 1) \qquad \text{for all} \quad (\varepsilon, \hat{p}) \in (0, \varepsilon_0] \times U(p).$$

a3. $$\inf_{\varepsilon \in (0, \varepsilon_0]} \inf_{\hat{p} \in U(p)} \partial^2_{\vartheta^2} c(\varepsilon, \vartheta_{\mathrm{LS}}(\varepsilon, \hat{p}), \hat{p}) > 0.$$

a4. From a1 it follows that

$$\sup_{\hat{p} \in U(p)} |\vartheta_{\mathrm{LS}}(\varepsilon, \hat{p})| \xrightarrow[\varepsilon \to 0]{} 0.$$

Then

(i) ϑ_{LS} is three times continuously differentiable on $[0, \varepsilon_0] \times U(p)$. On $(0, \varepsilon_0] \times U(p)$, the derivatives are

$$\partial_{\hat{p}} \vartheta_{\mathrm{LS}} = -\alpha^{-1} b, \tag{B.1}$$

$$\partial^2_{\hat{p}^2} \vartheta_{\mathrm{LS}} = -\alpha^{-1} \{ \Gamma + \beta (\partial_{\hat{p}} \vartheta_{\mathrm{LS}})' \, \partial_{\hat{p}} \vartheta_{\mathrm{LS}} + (\partial_{\hat{p}} \vartheta_{\mathrm{LS}})' \, d + d' (\partial_{\hat{p}} \vartheta_{\mathrm{LS}}) \}, \tag{B.2}$$

$$\partial_{\varepsilon} \vartheta_{\mathrm{LS}} = -\alpha^{-1} \gamma, \tag{B.3}$$

where

$$\alpha = 2 \sum_{i \in J} \hat{p}_i \{(\psi_{-i} - e_i)' \partial_{\vartheta^2}^2 \psi_{-i} + \| \partial_\vartheta \psi_{-i} \|^2\},$$

$$\beta = 2 \sum_{i \in J} \hat{p}_i \{3(\partial_\vartheta \psi_{-i})' \partial_{\vartheta^2}^2 \psi_{-i} + (\psi_{-i} - e_i)' \partial_{\vartheta^3}^3 \psi_{-i}\},$$

$$\gamma = 2 \sum_{i \in J} \hat{p}_i \{(\partial_\vartheta \psi_{-i})' \partial_\varepsilon \psi_{-i} + (\psi_{-i} - e_i)' \partial_{\varepsilon\vartheta}^2 \psi_{-i}\},$$

$$b = 2(((\psi_{-i} - e_i)' \partial_\vartheta \psi_{-i}))'_{i \in J} + 2 \sum_{j \in J} \hat{p}_j \{(\psi_{-j} - e_j)' \partial_{\hat{p}\vartheta}^2 \psi_{-j} + (\partial_\vartheta \psi_{-j})' \partial_{\hat{p}} \psi_{-j}\},$$

$$d = 2 \sum_{i \in J} \hat{p}_i \{(\partial_{\vartheta^2}^2 \psi_{-i})' \partial_{\hat{p}} \psi_{-i} + (\psi_{-i} - e_i)' \partial_{\hat{p}\vartheta^2}^3 \psi_{-i} + 2(\partial_\vartheta \psi_{-i})' \partial_{\hat{p}\vartheta}^2 \psi_{-i}\} + 2(((\psi_{-i} - e_i)' \partial_{\vartheta^2}^2 \psi_{-i} + \| \partial_\vartheta \psi_{-i} \|^2))'_{i \in J}, \tag{B.4}$$

$$\Gamma = 2 \sum_{k \in J} \hat{p}_k \{(((\psi_{-k} - e_k)' \partial_{\hat{p}_i \hat{p}_j \vartheta}^3 \psi_{-k} + (\partial_\vartheta \psi_{-k})' \partial_{\hat{p}_i \hat{p}_j}^2 \psi_{-k}))_{i,j \in J} + H_{-k} + H'_{-k}\} + 2(G + G'),$$

with

$$H_{-k} = (\partial_{\hat{p}} \psi_{-k})' \partial_{\hat{p}\vartheta}^2 \psi_{-k}$$

and

$$G = (((\psi_{-i} - e_i)' \partial_{\hat{p}_j \vartheta}^2 \psi_{-i} + (\partial_{\hat{p}_j} \psi_{-i})' \partial_\vartheta \psi_{-i}))_{i,j \in J}.$$

Thereby all functions and derivatives are regarded in the point $(\varepsilon, \hat{p})$ or $(\varepsilon, \vartheta_{LS}(\varepsilon, \hat{p}), \hat{p})$, denoting

$$\psi_{-i}(\varepsilon, \vartheta, \hat{p}) = \psi(\vartheta, \hat{p}_{-i}).$$

For all $\hat{p} \in U(p)$ we obtain

$$\partial_\varepsilon \vartheta_{LS}(0, \hat{p}) = \lim_{\varepsilon \downarrow 0} \partial_\varepsilon \vartheta_{LS}(\varepsilon, \hat{p}).$$

(ii) If the 4th partial derivatives of ψ are bounded on $[0, 1] \times \tilde{U}(p)$, then the 3rd partial derivatives of ϑ_{LS} are bounded on $[0, \varepsilon_0] \times U(p)$.

The proof is based on the theorem on implicit functions; see Grund [9].

C. Appendix

For all integers $N \geqslant N_o$ let the functions $\psi(N, \cdot) : \mathscr{S}_t \to \mathscr{S}_t$ be three times continuously differentiable on $U(p)$ with

$$\sup_{N \geqslant N_o} \sup_{\hat{p} \in U(p)} \| \partial_{\hat{p}}^3 \psi(N, \hat{p}) \| < c < \infty$$

and $\Delta = \psi(N, \hat{\boldsymbol{p}}) - \psi(N, p)$.

Then

$$\text{(i)} \qquad E\Delta = \frac{1}{2N} ((\operatorname{tr}\{\partial_{\hat{p}}^2 \psi_i(N, p)\, S(p)\}))_{i \in J} + O(N^{-2}),$$

$$\text{(ii)} \qquad E\Delta\Delta' = \frac{1}{N} \partial_{\hat{p}} \psi(N, p)\, S(p) (\partial_{\hat{p}} \psi(N, p))' + O(N^{-2}).$$

(iii) If, furthermore, there is a unique limiting point $B = \lim_{N \to \infty} \partial_{\hat{p}} \psi(N, p)$, then

$$\mathscr{L}\{\sqrt{N}\Delta\} \xrightarrow[N \to \infty]{} N_t(0, B\, S(p)\, B').$$

References

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

[2] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, MA.

[3] Bowman, A. W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika* **67** 682–684.

[4] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

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

[6] Burman, P. (1987). Central limit theorem for quadratic forms for sparse tables. *J. Multivariate Anal.* **22** 258–277.

[7] Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1195.

[8] Fienberg, S. E., and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68** 683–691.

[9] Grund, B. (1987). *Estimates for Cell Probabilities in Multinomial Contingency Tables.* Ph.D. Thesis, Humboldt-Universität, Berlin. [In German]

[10] Grund, B., and Hall, P. (1993). On the performance of kernel estimates for high-dimensional, sparse binary data. *J. Multivariate Anal.* **44** 321–344.

[11] Habbema, J. D. F., Hermans, J., and van den Broeck, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat* 1974 (L. C. A. Corsten and J. Hermans, Eds.). Physica-Verlag, Wien.

[12] HALL, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **69** 287–294.
[13] HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.
[14] HALL, P., AND MARRON, S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15** 163–181.
[15] HALL, P., SHEATHER, S.J., JONES, M. C., AND MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 262–269.
[16] JONES, M. C. (1991). The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.* **12** 51–56.
[17] MAMMEN, E. (1990). A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Letters* **9** 23–25.
[18] MARRON, J.S. (1988). Automatic smoothing parameter selection: A survey. *Empirical Econ.* **13** 187–208.
[19] MARRON, S., AND HÄRDLE, W. (1986). Random approximations to some measures of accuracy in nonparametric estimation. *J. Multivariate Anal.* **20** 91–113.
[20] RUDEMO, M. (1982). Empirical choice of histogram and kernel density estimators. *Scand. J. Statist.* **9** 65–78.
[21] SANTNER, T. J., AND DUFFY, D. E. (1989). *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York.
[22] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.
[23] SIMONOFF, J. S. (1987). Probability estimation via smoothing in sparse contingency tables with ordered categories. *Statist. Probab. Letters* **5** 55–63.
[24] STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515.
[25] SUTHERLAND, M., HOLLAND, P., AND FIENBERG, S. E. (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter. In *Studies in Bayesian Econometrics and Statistics.* North-Holland, Amsterdam.
[26] TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22** 259–268.
[27] WANG, M. C., AND RYZIN, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika* **68** 301–309.