JOURNAL OF MULTIVARIATE ANALYSIS 44, 321–344 (1993)

On the Performance of Kernel Estimators for High-Dimensional, Sparse Binary Data*

BIRGIT GRUND

University of Minnesota

AND

PETER HALL

Australian National University, Sydney, Australia

Communicated by the Editors

We develop mathematical models for high-dimensional binary distributions, and apply them to the study of smoothing methods for sparse binary data. Specifically, we treat the kernel-type estimator developed by Aitchison and Aitken (*Biometrika* **63** (1976), 413–420). Our analysis is of an asymptotic nature. It permits a concise account of the way in which data dimension, data sparseness, and distribution smoothness interact to determine the over-all performance of smoothing methods. Previous work on this problem has been hampered by the requirement that the data dimension be fixed. Our approach allows dimension to increase with sample size, so that the theoretical model may accurately reflect the situations encountered in practice; e.g., approximately 20 dimensions and 40 data points. We compare the performance of kernel estimators with that of the cell frequency estimator, and describe the effectiveness of cross-validation.

1. INTRODUCTION

The problem of estimating cell probabilities for multivariate, binary data is an important one, with applications in a variety of problems where data are recorded only as yes/no, or zero/one, responses. When the dimension is large, data may be distributed quite sparsely among cells, and cells with

Received June 10, 1991; revised June 22, 1992.

AMS subject classifications: 62G05, 62M15.

Key words and phrases: binary data, cell frequency estimator, cross-validation, use of dimensionality, kernel estimator, Kullback-Leibler loss, mean squared error, smoothing, sparseness, squared error.

* This work was supported in part by Grant N26 from the program "Pôle d'Attraction Interuniversitaire—Deuxième phase" to CORE, Université Catholique de Louvain.

0047-259X/93 $5.00

683 44 2-10

no data may arise. In such cases it can be advantageous to make use of relationships among probabilities in neighbouring cells. If cells which are "close" in some sense have probabilities which are also close, then data from neighbouring cells can be employed to help estimate cell probabilities.

The kernel estimator introduced by Aitchison and Aitken [1] is designed to make use of just such neighbourhood information. It incorporates a smoothing parameter, designated by θ in the present paper, which may be chosen empirically to optimise the extent to which information from neighbouring cells can be employed to estimate cell probabilities. The context of Aitchison and Aitken's estimator is that of a relatively high number of dimensions, with relatively few data per cell. For example, the binary data sets to which Aitchison and Aitken applied their estimator were of $m = 17$ dimensions, with sample sizes $n = 37$ and $n = 40$. Thus, the number of data per cell was at most $n/2^m = 40/2^{17} \approx 3 \times 10^{-3}$. Estimation in such high-dimensional, sparse contexts is quite feasible, provided the underlying probability distribution is sufficiently smooth.

Theoretical accounts of the performance of Aitchison and Aitken's estimator have hitherto been restricted to the case of a fixed value of m, with sample size increasing (e.g., Bowman [3], Hall [9]). In this setting the average number of data values per cell diverges to $+\infty$, and so the context of Aitchison and Aitken's estimator is not preserved. The aim of the present paper is to develop an entirely different mathematical model for describing the performance of Aitchison and Aitken's estimator, reflecting the high-dimensional, multivariate character of the problem, and describing the way in which distribution smoothness may be used to overcome difficulties caused by data sparseness.

We describe smoothness by an adjustable parameter, δ, with the property that smoothness increases as $\delta \to 0$; we allow the number of dimensions, m, to increase with sample size, n; and we permit δ, m, and n to vary simultaneously. Tractable theoretical models for cell probabilities are developed within this framework, and the performance of Aitchison and Aitken's estimator is studied under those models. In particular, our models permit $n/2^m$ to decrease to zero. We show theoretically that, depending on the smoothness of the underlying distribution, Aitchison and Aitken's estimator can produce substantial improvements over the cell frequency estimator, which estimates cell probabilities as the observed data frequencies.

Our models allow us to identify concisely the way in which dimension m, data sparseness (represented by $n/2^m$), and distribution smoothness (represented by δ^{-1}) affect performance. We discuss three distinct cases: where there is a significant improvement to be gained by using Aitchison and Aitken's estimator, or where the amount of improvement is marginal, or where there is no possibility of improvement in a first-order sense. We

investigate these three different situations using both isotropic and anisotropic models for cell probabilities, ensuring that the models cover a broad range of distributions, and we show that cross-validation can be used to determine empirically the optimal smoothing parameter even in contexts of extreme data sparseness (i.e., small values of $n/2^m$).

In problems where dimension is large, individual cell probabilities are usually small, and consistency of estimation of those probabilities is not usually the central issue. Instead, one is typically interested in a global measure of performance, such as mean summed squared error,

$$MSSE = \sum_i E(\hat{p}_i - p_i)^2.$$

Here, p_i denotes the probability that an arbitrary data value falls into cell i, the cell index $i \in \{0, 1\}^m$ standing for the outcomes of the m underlying binary variables; $\hat{p}_i$ represents Aitchison and Aitken's estimator of p_i; and enumeration is taken over all 2^m cells. An alternative measure of fidelity is mean Kullback–Leibler loss,

$$MKL = \sum_i p_i E\left\{\log\left(\frac{p_i}{\hat{p}_i}\right)\right\}.$$

Indeed, provided the function f in our mathematical model for cell probabilities (see (2.5) and (2.6) below) is bounded away from zero, it may be shown using the techniques of Section 5 that first-order properties of MKL are asymptotically equivalent to those of a particular form of weighted mean summed squared error,

$$WMSSE = \tfrac{1}{2} \sum_i E(\hat{p}_i - p_i)^2 \, p_i^{-1}.$$

Furthermore, Kullback–Leibler cross-validation leads to asymptotic minimisation of mean Kullback–Leibler loss, for a wide range of values of dimension, data sparseness and distribution smoothness. We confine attention to verifying this result in the case of squared-error cross-validation and squared-error loss.

Section 2 introduces our models for the cell probabilities p_i, and describes mean squared error properties of Aitchison and Aitken's estimator under those models. Our main results there identify the roles played by dimension, data sparseness, and smoothness in determining the relative merits of the kernel estimator and the cell frequency estimator. Section 3 shows that minimising loss, minimising risk, and minimising the cross-validation criterion are all asymptotically equivalent, to first order. Section 4 demonstrates that the models of Section 2, though looking at first

sight rather restrictive, indeed represent any finite-dimensional distribution. All proofs are deferred to the Appendix.

Related work on the smoothing of sparse multinomial models includes that of Sutherland *et al.* [15], Bishop *et al.* [2], [Chap. 12], Ighodaro and Santner [11], Simonoff [13, 14], Burman [5, 6], and Hall and Titterington [10]. However, none of the probability models in these papers admits the case of *multivariate* binary data. Hall and Titterington [10] do allow a degree of sparseness in their one-dimensional model, but neighbour relationships among cells, inherent to that work, are all strictly one-dimensional. While we employ univariate distributions as part of our m-variate model, the neighbour relationships in that model are intrinsically m-variate. In particular, our formula (2.6), in Section 2 below, cannot be written in the form of Hall and Titterington's [10] formula (3.1), since in our model there is no natural linear ordering of the cells. Of more incidental interest is an observation made by Titterington [16] that a special form of Aitchison and Aitken's estimator falls into a class considered by Sutherland *et al.* [15]. It would be possible to develop versions of our results for estimators in that class.

There is an extensive literature on smoothing methods for categorical data, which has been surveyed by Titterington [16], and Santner and Duffy [12]. The work of Brown and Rundall [4] on smoothing in two-dimensional contingency tables should also be mentioned. Grund [7, 8] has developed a detailed theoretical account of smoothing for Aitchison and Aitken's estimator, in the case of fixed dimension.

2. Properties of Mean Squared Error

2.1. *Summary*

Subsection 2.2 introduces mathematical models for cell probabilities in high-dimensional binary distributions. Two different cases are treated in detail, representing isotropic and anisotropic distributions respectively. Aitchison and Aitken's [1] kernel estimator is defined in Subsection 2.3, and its mean squared error properties are described in Subsection 2.4. There it is shown that the extent to which the kernel estimator improves on the cell frequency estimator depends in a monotone increasing way on smoothness of the binary probability distribution and on sparseness of the data.

2.2. *Models for Cell Probabilities*

The sample space of possible data values is the set $\mathscr{I}$ of all m-vectors $\underline{i} = (i_1, ..., i_m)^T$, where each i_j may be either 0 or 1. Thus, $\mathscr{I}$ contains 2^m

elements. Given an m-vector $v = (v_1, ..., v_m)^T$, define $|v| = \sum |v_j|$. If $v = \underline{i} - \underline{i}$, where $\underline{i} = (i_1, ..., i_m)^T$ and $\underline{i}' = (i_1', ..., i_m')$ are elements of $\mathscr{I}$, then $|v|$ equals the number of indices j such that $i_j \neq i_j'$. If $|\underline{i} - \underline{i}'| = r$ then we say that $\underline{i}'$ is an rth order neighbour of $\underline{i}$. Data are drawn from a distribution p on $\mathscr{I}$, with probability p_i ascribed to cell $\underline{i}$.

Our mathematical model for the p_i's may be motivated as follows. To ensure tractability, we wish to be able to adjust the p_i's through a fixed number of parameters no matter how large the number of dimensions. This indicates that we should define the p_i's in terms of low-dimensional functions, for example

$$p_i = Cf(\omega_1^T \underline{i}, ..., \omega_t^T \underline{i}), \qquad \underline{i} \in \mathscr{I}, \tag{2.1}$$

where $t \geqslant 1$ is fixed; f is a nonnegative, t-variate function; $\omega_1, ..., \omega_t$ are m-vectors; and C is chosen to ensure that $\sum \mathbf{p}_i = 1$.

Several modifications of the prescription (2.1) are in order. First, there are no important qualitative or quantitative differences between the cases $t = 1$ and $t > 1$. Taking $t = 1$ in (2.1) greatly simplifies notation, and so we adopt this convention, defining

$$p_i = Cf(\omega^T \underline{i}), \qquad \underline{i} \in \mathscr{I}. \tag{2.2}$$

Second, it is of interest to incorporate an additional parameter into (2.2), so as to govern the smoothness of the binary distribution. Perhaps the simplest way of doing this is to define

$$p_i = Cf(\delta \omega^T \underline{i}), \qquad \underline{i} \in \mathscr{I}, \tag{2.3}$$

where $\delta \geqslant 0$ can be varied to adjust the distance between p_i's for neighbouring $\underline{i}$'s. The smaller the value of δ, the smoother the probability distribution; this suggests defining smoothness as δ^{-1}.

For the sake of definiteness we take ω in (2.3) to be an m-vector of 1's, $\omega = \mathbf{1} = (1, ..., 1)^T$. Other choices give rise to distributions with similar properties, but require the elements of ω to be carefully defined for each m. This is feasible, but cumbersome. Taking $\omega = \mathbf{1}$ in (2.3) is equivalent to defining

$$p_i = Cf(\delta\, |\underline{i}|). \tag{2.4}$$

Models such as (2.4) are anisotropic in character; they require all the cell probabilities to be the same function of a single projection of the cell index. Many binary data problems are intrinsically isotropic, with the relationship between $p_{\underline{i}}$ and $p_{\underline{i}'}$ depending principally on some measure of the

separation of $\underline{i}$ and $\underline{i}'$, without regard to a specific direction. An isotropic version of (2.4) is given by

$$p_{\underline{i}} = C\|\Omega\|^{-1} \int_{\Omega} f(\delta\omega^{\mathrm{T}}\underline{i})\, d\omega, \tag{2.5}$$

where Ω denotes the set of all unit vectors ω (i.e., the m-dimensional sphere of unit radius, centered at the origin) and $\|\Omega\|$ equals the surface content of Ω. Of course, the normalising constants C in (2.3)–(2.5) are generally quite different.

The probability distribution determined by the $p_{\underline{i}}$'s, defined by (2.4), is centered in a corner of the $2 \times \cdots \times 2$ "cube" $\mathscr{I}$. The center may be effectively translated to the centroid of the cube by re-defining

$$p_{\underline{i}} = Cf\{\delta(|\underline{i}| - \tfrac{1}{2}m)\}. \tag{2.6}$$

Asymptotic properties of estimators based on the distributions given by (2.4) and (2.6) are quite similar, and may be established by identical arguments. For the sake of brevity, our asymptotic analysis will focus on the models (2.5) and (2.6). We show that the smoothness, δ^{-1}, has greater impact for the isotropic distribution (2.5) than it does in the anisotropic case (2.6).

2.3. *Definition of Estimator*

Let $X_1, \ldots, X_n$ denote a random sample drawn from the distribution p. Thus, each X_j is an m-vector of 0's and 1's. Aitchison and Aitken's estimator is given by

$$\hat{p}_{\underline{i}} = n^{-1} \sum_{j=1}^{n} \theta^{|X_j - \underline{i}|}(1-\theta)^{m - |X_j - \underline{i}|},$$

where $\theta \in [0, \frac{1}{2}]$ is the smoothing parameter (not be confused with the smoothness of the distribution p). Taking $\theta \to 0$ produces the cell frequency estimator,

$$\tilde{p}_{\underline{i}} = n^{-1} \sum_{j=1}^{n} I(X_j = \underline{i}) = n^{-1}\,(\text{number of } X_j\text{'s in cell } \underline{i}).$$

Here, $I(\cdot)$ denotes the indicator function. Taking $\theta = \frac{1}{2}$ gives the other extreme; i.e., $\hat{p}_{\underline{i}} = 2^{-m}$ for each $\underline{i}$. We confine attention to cases where the optimal smoothing parameter, θ_{opt}, converges to zero as m and n diverge. Indeed, it turns out that the condition $m\theta_{\mathrm{opt}}^3 \to 0$ is necessary and sufficient for obtaining the variance formulae which we give.

If we define N_{ir} to equal the number of rth order neighbours of cell i, i.e.,

$$N_{ir} = \sum_{j=1}^{n} I(|X_j - i| = r),$$

then we may give the following alternative, equivalent definitions of $\hat{p}_i$ and $\tilde{p}_i$:

$$\hat{p}_i = n^{-1} \sum_{r=0}^{m} N_{ir} \theta^r (1-\theta)^{m-r}, \qquad \tilde{p}_i = n^{-1} N_{i0}.$$

2.4. *Properties of Mean Squared Error*

As indicated in Subsection 2.3, the probabilities p_i are governed by two variable parameters m and δ. Each of these may be regarded as a function of n. Thus, conditions such as $m\delta^2 \to 0$ in the theory below may be interpreted as $m(n)\,\delta(n)^2 \to 0$ as $n \to \infty$. According to this convention, the constant C in (2.5) and (2.6) is a function of n.

In much of our work we assume the following regularity conditions.

(a) The probabilities p_i are defined by (2.6), for Theorems 2.1 and 2.2, or by (2.5), for Theorems 2.3 and 2.4.

(b) Either (b1) f'' exists and is bounded on $(-\infty, \infty)$, $f \geqslant 0$, $f(0) \neq 0$, and $m\delta^2 \to 0$; or (b2) f is a positive semi-definite quadratic function, say $f(x) = a_0 + a_1 x + a_2 x^2$, with $a_0 a_2 \neq 0$, and either $m\delta^2 \to 0$, or $m\delta^2 \to l$ $(0 < l < \infty)$, or $m\delta^2 \to \infty$.

Mean summed squared error, $MSSE = \sum E(\hat{p}_i - p_i)^2$, is a measure of the performance of $\hat{p}$. In the case $\theta = 0$, where $\hat{p} = \tilde{p}$ (the cell proportion estimator), $MSSE$ admits the formula

$$MSSE = \sum_i E(\tilde{p}_i - p_i)^2 = n^{-1} \sum_i p_i(1-p_i) = n^{-1}\{1 + o(1)\}, \qquad (2.7)$$

assuming only that $\max p_i \to 0$ as m increases. Thus, any estimator whose mean summed squared error decreases at a faster rate than n^{-1} improves on the performance of $\tilde{p}$ by an order of magnitude.

Theorem 2.1 below provides simplified asymptotic formulae for $MSSE$. In those expressions, C denotes the normalizing constant in (2.6). Theorem 2.2 refines those formulae and shows that if the distribution p is sufficiently smooth then, for an appropriate choice of θ, the kernel estimator $\hat{p}$ can indeed provide a significant improvement over the cell frequency estimator.

THEOREM 2.1. *Let the p_i's be given by* (2.6), *and assume that $m \to \infty$ and $0 < \delta \leqslant 1$ as $n \to \infty$.*

(I) *If* (b1) *holds then*

$$\left\{\sum_i E(\hat{p}_i - p_i)^2\right\}\{C^2 2^m m\ \delta^2\theta^2 f'(0)^2 + n^{-1} e^{2m\theta}\}^{-1} = 1 + R_n(\theta) \tag{2.8}$$

where, for any sequence $\varepsilon = \varepsilon(n) > 0$ *satisfying* $\varepsilon = o(m^{-1/3})$,

$$\sup_{0 < \theta \leq \varepsilon} |R_n(\theta)| \to 0 \tag{2.9}$$

as $n \to \infty$.

(II) *If* $f(x) = a_0 + a_1 x + a_2 x^2$ *is a positive semi-definite quadratic function, and* $f(0) > 0$, *then*

$$\left\{\sum_i E(\hat{p}_i - p_i)^2\right\}\{C^2 2^m m\,\delta^2\theta^2(a_1^2 + 2m\,\delta^2\{(1-\theta)\,a_2\}^2) + n^{-1} e^{2m\theta}\}^{-1} = 1 + R_n(\theta), \tag{2.10}$$

where (2.9) *holds for any positive sequence* $\varepsilon = o(m^{-1/3})$.

Given $0 < \lambda_0 < \infty$, define x_0 to be the solution of the equation $x_0 e^{x_0} = \lambda_0$, and put $q = (1 + \frac{1}{2}x_0)\,e^{-x_0} < 1$. Let $D(\theta) = \sum E(\hat{p}_i - p_i)^2$.

THEOREM 2.2. *Let the* p_i*'s be given by* (2.6). *Assume condition* (b), *and that* $m \to \infty$ *and* $0 < \delta \leqslant 1$. *If* $m\delta^2 \to 0$, *define* $\rho = f(0)^{-2} f'(0)^2$ *and* $\lambda = 2^{m+1} m(\rho n \delta^2)^{-1}$; *if* $m\delta^2 \to l$ *with* $0 < l < \infty$, *define* $\rho = (a_0 + \frac{1}{4}a_2 l)^{-2}(a_1^2 + 2a_2^2 l)l$ *and* $\lambda = 2^{m+1} m^2 (\rho n)^{-1}$; *and if* $m\delta^2 \to \infty$, *define* $\rho = 16$ *and* $\lambda = 2^{m+1} m^2 (\rho n)^{-1}$. *Assume that either* $\lambda \to 0$, *or* $\lambda \to \lambda_0$ *with* $0 < \lambda_0 < \infty$, *or* $\lambda \to \infty$, *and put*

$$\theta_0 = \begin{cases} 0 & \text{if } \lambda \to 0 \\ (2m)^{-1} x_0 & \text{if } \lambda \to \lambda_0 \\ (2m)^{-1}(\log \lambda - \log\log \lambda) & \text{if } \lambda \to \infty. \end{cases} \tag{2.11}$$

In the case $\lambda \to \infty$, *assume in addition that*

$$m^{-2/3} \log(2^m / n\delta^2) \to 0. \tag{2.12}$$

Then

$$\inf_\theta D(\theta) \sim D(\theta_0) \sim \begin{cases} n^{-1} & \text{if } \lambda \to 0 \\ n^{-1} q & \text{if } \lambda \to \lambda_0 \\ n^{-1}(2\lambda)^{-1}(\log \lambda)^2 & \text{if } \lambda \to \infty. \end{cases} \tag{2.13}$$

The following remarks elucidate and expand on Theorems 2.1 and 2.2.

Remark 2.1. In the case where $m\delta^2 \to l$ and $0 \leqslant l < \infty$, the constant C appearing in (2.8) and (2.10) is asymptotic to a multiple of 2^{-m}. When $m\delta^2 \to \infty$ it is asymptotic to a multiple of $(m\delta^2)^{-1} 2^{-m}$.

Remark 2.2. In the terms forming the denominators on the left-hand sides of (2.8) and (2.10), the quantity $n^{-1} e^{-2m\theta}$ represents the contribution from variance to *MSSE*. That is,

$$\sum_i \operatorname{var}(\hat{p}_i) \sim n^{-1} e^{-2m\theta}. \tag{2.14}$$

The other quantities represent the squared bias contribution. The condition $m\theta^3 \to 0$ is necessary and sufficient for formula (2.14) to hold. Indeed, if $m\theta^3 \to k$, where $0 < k < \infty$, then the right-hand side of (2.14) must be multiplied by a constant factor $c(k) > 1$; and if $m\theta^3 \to \infty$, then $\sum \operatorname{var}(\hat{p}_i)$ is an order of magnitude larger than $n^{-1} e^{-2m\theta}$. This behaviour will be elucidated at the end of Part (ii) of our proof of Theorems 2.1 and 2.2, in the Appendix. Additional results, dealing with the case where $m\theta^3$ is bounded away from zero, may be derived by similar methods but are not given here.

Remark 2.3. Condition (2.12) is necessary and sufficient to ensure the vital condition $m\theta_0^3 \to 0$. It is needed only when $\lambda \to \infty$, since the condition $m\theta_0^3 \to 0$ is trivially satisfied if λ is bounded.

It may be shown by an argument based on convexity that there is a unique θ_{opt} which minimises $D(\theta)$. Our proof of Theorem 2.2 will show that if (2.12) holds (for $\lambda \to \infty$), then there exists a local minimum of $D(\theta)$ at a point θ which satisfies $m\theta^3 \to 0$. This must, by uniqueness, be the point which produces the global minimum.

Remark 2.4. Theorem 2.2 describes the behaviour of $\min D(\theta)$ in a total of nine different cases: three depending on asymptotic properties of $m\delta^2$, by three for the properties of λ.

Remark 2.5. The values of θ_0 given in (2.11) are of course not identical to the optimal θ's which minimise *MSSE*. Rather, they denote relatively simple quantities which asymptotically minimise *MSSE*, to first order. In the case where $\lambda \to \infty$, the asymptotic formula $D(\theta_0) \sim n^{-1}(2\lambda)^{-1}(\log \lambda)^2$ fails if the "log log λ" term is deleted from the definition $\theta_0 = (2m)^{-1}(\log \lambda - \log\log \lambda)$. This makes it clear that it is not sufficient to have simply $\theta_{\text{opt}}/\theta_0 \to 1$.

Remark 2.6. The quantity λ represents a particular combination of the notion of smoothness of the probability distribution p, and sparseness of the distribution of data among cells. That combination turns out to be crucial in determining the relative performances of the kernel estimator $\hat{p}$

and the cell frequency estimator $\tilde{p}$. Note that λ is an increasing function of the smoothness of p (represented by δ^{-1}) and of the sparseness of the data (represented by $2^m/n$). Furthermore, since mean summed squared error for the cell frequency estimator is asymptotic to n^{-1} (see (2.7)), then we may deduce from (2.13) that the extent of improvement which $\hat{p}$ offers over $\tilde{p}$ increases with increasing λ. Therefore, the performance of $\hat{p}$ relative to $\tilde{p}$ is an increasing function of both distribution smoothness and data sparseness. This conclusion is not unexpected, but it is particularly satisfying to be able to verify it rigorously within the confines of a mathematical model. The precise way in which λ, and hence the performance of $\hat{p}$ relative to $\tilde{p}$, depends on distribution smoothness and data sparseness was not expected, and seems difficult to explain intuitively.

We now state versions of Theorems 2.1 and 2.2 for the case where the isotropic model (2.5) is used to define the p_i's. For the sake of brevity our proof will be given in the case of quadratic f, and so we state the results only for that context. Theorem 2.4, the analogue of Theorem 2.2 in the case of isotropy, is particularly interesting. Note that we do not any longer need to isolate separate cases depending on the behaviour of $m\delta^2$, and that the definition of λ is different from what it was for Theorem 2.2, but that the form of the result is unchanged.

Recall that $D(\theta) = \sum E(\hat{p}_i - p_i)^2$.

THEOREM 2.3. *Let the p_i's be given by* (2.5). *Assume that* $f(x) = a_0 + a_1 x + a_2 x^2$ *is a positive semi-definite quadratic function with* $f(0) > 0$, *and that* $0 < \delta \leqslant 1$ *and* $m \to \infty$. *Then*

$$\left\{\sum_i E(\hat{p}_i - p_i)^2\right\} \{C^2 2^m m^{-1} \delta^4 a_2^2 + n^{-1} e^{-2m\theta}\}^{-1} = 1 + R_n(\theta), \quad (2.15)$$

where $R_n(\theta)$ *satisfies* (2.9) *for any sequence* $\varepsilon = \varepsilon(n) = o(m^{-1/3})$.

THEOREM 2.4. *Let f and the p_i's be as in Theorem* 2.3. *Assume that* $0 < \delta \leqslant 1$ *and* $m \to \infty$; *define* $\rho = (a_0 + \frac{1}{2}a_2\delta^2)^{-2} a_2^2$ *and* $\lambda = 2^{m+1} m^3 (\rho n \delta^4)^{-1}$; *and suppose that either* $\lambda \to 0$, *or* $\lambda \to \lambda_0$ *with* $0 < \lambda_0 < \infty$, *or* $\lambda \to \infty$. *Let* x_0 *and* q *be as in Theorem* 2.2, *and in the case* $\lambda \to \infty$, *assume in addition that*

$$m^{-2/3} \log(2^m/n\delta^4) \to 0.$$

Then with θ_0 *defined by* (2.11), *result* (2.13) *holds.*

Remark 2.7. The constant C in (2.15) is identical to that in (2.5). If $\delta \to \delta_0$, where $0 \leqslant \delta_0 \leqslant 1$, then C is asymptotic to a constant multiple of 2^{-m}.

Remark 2.8. Comparing Theorems 2.2 and 2.4 we see that the principal difference between the isotropic and anisotropic cases is that λ is of a larger order of magnitude in the former case. In view of (2.13), this means that $\hat{p}$ can offer greater improvement over $\tilde{p}$ in the case of isotropy than it can for anisotropic distributions. Of course, the larger size of λ is due to the fact that bias is small under isotropy.

3. Properties of Squared-Error Cross-Validation

Section 2 concentrated on mean summed squared error as a measure of the performance of the kernel estimator $\hat{p}$. In the present section we show that to first order, summed squared error provides an asymptotically equivalent description. That is, "loss" and "risk" are equivalent, to first order. We also show that the cross-validatory criterion is first-order equivalent to summed squared error, up to terms which do not depend on the smoothing parameter θ. This indicates that minimising the cross-validation criterion is asymptotically equivalent to minimising summed squared error, and also to minimising mean summed squared error.

For the sake of brevity we establish these relationships in a pointwise sense for θ's. For example, we prove that if

$$SSE = \sum_i (\hat{p}_i - p_i)^2$$

denotes summed squared error, then

$$\sup_{0 \leqslant \theta \leqslant \varepsilon} E(SSE - MSSE)^2 (MSSE)^{-2} \to 0 \tag{3.1}$$

provided $\varepsilon = o(m^{-1/3})$. Substantially longer proofs may be used to establish related results, such as

$$\sup_{0 \leqslant \theta \leqslant \varepsilon} |SSE - MSSE| (MSSE)^{-1} \to 0$$

in probability. Also for the sake of brevity, we treat only the case where f is positive semi-definite quadratic.

Theorem 3.1. *Let $f(x) = a_0 + a_1 x + a_2 x^2$ be a positive semi-definite quadratic function with $f(0) > 0$, and define the cell probabilities p_i by either* (2.5) *or* (2.6), *with $m \to \infty$ and $0 < \delta \leqslant 1$. Let $\varepsilon = \varepsilon(n)$ denote a sequence of positive numbers satisfying $\varepsilon = o(m^{-1/3})$. Then* (3.1) *holds.*

Our final result is an analogue of Theorem 3.1 for the difference between

SSE and the cross-validation criterion, *CV*. To introduce the latter, observe that minimising *SSE* is equivalent to minimising

$$SSE' = \sum_i \hat{p}_i^2 - 2 \sum_i p_i \hat{p}_i.$$

The first term on the right-hand side is of course known. An approximation to the second term is given by

$$-2n^{-1} \sum_{j=1}^{n} \hat{p}_{j,X_j},$$

where

$$\hat{p}_{j,i} = (n-1)^{-1} \sum_{k \neq j} \theta^{|X_k - i|}(1-\theta)^{m - |X_k - i|}$$

is the version of $\hat{p}_i$ constructed from the $(n-1)$-sample obtained by deleting X_j. (Note that the value taken by X_j is one of the m-variate indices $\underline{i}$.) Now, $\sum p_i \hat{p}_i$ and $n^{-1}\hat{p}_{j,X_j}$ have the same expected values. This suggests that

$$CV = CV(\theta) = \sum_i \hat{p}_i^2 - 2n^{-1} \sum_{j=1}^{n} \hat{p}_{j,X_j}$$

be taken as an approximation to *SSE'*, and that θ be chosen to minimise *CV*. Theorem 3.2 shows that, up to terms which do not depend on θ, *CV* provides a good approximation to *SSE*.

Define

$$T = \sum_i p_i \hat{p}_i = n^{-1} \sum_{j=1}^{n} \sum_i p_i I(X_j = \underline{i}),$$

which does not depend on θ.

THEOREM 3.2. *Assume the conditions of Theorem* 3.1. *Then*

$$\sup_{0 \leq \theta \leq \varepsilon} E(SSE - CV - T)^2 (MSSE)^{-2} \to 0$$

as $n \to \infty$.

A substantially longer proof may be used to show that

$$\sup_{0 \leq \theta \leq \varepsilon} |SSE - CV - T| (MSSE)^{-1} \to 0$$

in probability, which together with (3.1) may be employed to show that minimising the cross-validatory criterion, in the range $0 \leqslant \theta \leqslant \varepsilon$, is asymptotically equivalent to minimising *MSSE* in the same range.

4. Discussion of the Anisotropic Model

At first sight, the investigated models give the impression of being rather restrictive, as they inevitably assign equal probability to all cells with the cell indices consisting of the same number of 0's and 1's. In the following we show that the results obtained under the anisotropic model really provide us a feeling for the *general* behaviour of kernel estimators, although they may seem at first sight to be confined to a subset of "nice" parameters, likely to suit Aitchison and Aitken's estimator.

We prove that for any probability distribution on $\mathscr{I}$ we can find a $\delta > 0$ and a function $f: \mathbb{R} \to \mathbb{R}^+$, such that the p_i's defined by (2.4) or (2.6) give the same *MSSE* as the original distribution—for all smoothing parameters $\theta \in [0, \frac{1}{2}]$. Moreover, these equal-risk cell probabilities are "rougher" than the original distribution. Thus, we decompose the space of probability distributions in equivalence classes of equal risk (independent of θ), ensuring that the anisotropic model contains *reasonable* representatives of each class.

In this section, each distribution on $\mathscr{I}$ is described by the corresponding cell probability vector $p = (p_i)_{i \in \mathscr{I}}$, with the cell indices following the lexicographical order; and the set of all cell probability vectors is given by $\mathscr{S}_m = \{ p \in \mathbb{R}^{2^m} \mid p_i \geqslant 0, \sum_i p_i = 1 \}$, with the center $c_m = 2^{-m}(1, ..., 1)^T$.

Obviously, the equivalence classes of equal risk have to depend on the special structure of Aitchison and Aitken's estimator. We use particularly the eigenvalue decomposition of the corresponding kernel matrix, provided by Lemma 4.1. The following corollaries describe the equal risk classes, whereas Theorem 4.1 ensures that the anisotropic model indeed covers an outstanding representative of each class.

For the sake of brevity we denote by

$$\bigotimes_{j=1}^{m} B_j = B_1 \otimes \cdots \otimes B_m \tag{4.1}$$

the Kronecker product of matrices or vectors $B_1, ..., B_m$ and, letting $J \subseteq \{1, ..., m\}$ denote an index set, we write $\{ \bigotimes_{j=1}^{m} B_j \rightleftharpoons_J H \}$ for the Kronecker product obtained by replacing in (4.1) each of the matrices B_j, $j \in J$, by H.

LEMMA 4.1. *The kernel estimator of Aitchison and Aitken satisfies* $\hat{p} = A\tilde{p}$, *where the kernel matrix*

$$A = \bigotimes_{j=1}^{m} U(\theta) \tag{4.2}$$

is the m-fold Kronecker product of the 2×2*-matrix* $U(\theta) = (1-2\theta)\mathbf{I} + \theta\mathbf{1}\mathbf{1}^{\mathrm{T}}$, *and where* $\mathbf{I}$ *is the identity matrix and* $\mathbf{1} = (1, 1)^{\mathrm{T}}$. *The matrix A has* $(m+1)$ *distinct eigenvalues*

$$\lambda_k = (1-2\theta)^k, \qquad k = 0, \ldots, m. \tag{4.3}$$

The $\binom{m}{k}$*-dimensional space of eigenvectors belonging to the eigenvalue* λ_k *is spanned by the basis*

$$h_{k;J(k)} = 2^{-m/2} \left\{ \bigotimes_{j=1}^{m} \mathbf{1} \rightleftharpoons_{J(k)} h \right\}, \tag{4.4}$$

where $h = (1, -1)^{\mathrm{T}}$ *and* $J(k) \subseteq \{1, \ldots, m\}$ *varies over all subsets of size k.*

COROLLARY 4.1. *Each cell probability vector* $p \in \mathscr{S}_m$ *has a unique representation*

$$p = c_m + \alpha \sum_{k=1}^{m} \sum_{J(k)} \beta_{k;J(k)} h_{k;J(k)}, \tag{4.5}$$

with $\alpha > 0$ *and* $\sum_k \sum_{J(k)} \beta_{k;J(k)}^2 = 1$. *In the double series* $\sum_k \sum_{J(k)}$, *the inner series is taken over all subsets of* $\{1, \ldots, m\}$ *which are of size k.*

COROLLARY 4.2. *For each* $\alpha > 0$, $\gamma_1, \ldots, \gamma_m > 0$ *the MSSE of* $\hat{p}$ *is constant within the class*

$$\mathscr{S}_m(\alpha, \gamma_1, \ldots, \gamma_m) = \Bigg\{ p \in \mathscr{S}_m \mid p \text{ satisfies (4.5) with } \sum_{J(k)} \beta_{k;J(k)}^2 = \gamma_k,\ k = 1, \ldots, m \Bigg\}, \tag{4.6}$$

independently of the smoothing parameter θ.

Remark. Formula (4.2) generalises the Kronecker product representation of Aitchison and Aitken's estimator in Brown and Rundall [4]. The proofs of Lemma 4.1 and the two corollaries are straightforward and are therefore omitted.

Formula (4.5) suggests that we regard the parameter space $\mathscr{S}_m$ as

consisting of concentric $(2^m - 1)$-dimensional spheres around c_m. The sets of equal risk, given in (4.6), contain cuts of such spheres with hyperplanes that are perpendicular to the eigenvector spaces defined by (4.4). Motivating the anisotropic model, we consider in Theorem 4.1 the decomposition of $\mathscr{S}_m$ into classes of equal risk

$$\mathscr{S}_m = \bigcup_{\alpha, \gamma_1, \ldots, \gamma_m > 0} \mathscr{S}_m(\alpha, \gamma_1, \ldots, \gamma_m) \cap \mathscr{S}_m \tag{4.7}$$

and show that each class is appropriately represented by the model.

THEOREM 4.1. (I) *For each $\alpha > 0$, $\gamma_1, \ldots, \gamma_m > 0$ and $\delta > 0$ we can find a function $f: \mathbb{R} \to \mathbb{R}^+$, such that the cell probability vector p defined by* (2.4) *or* (2.6) *is in $\mathscr{S}_m(\alpha, \gamma_1, \ldots, \gamma_m)$.*

(II) *The vector p is the roughest distribution in $\mathscr{S}_m(\alpha, \gamma_1, \ldots, \gamma_m)$, maximising the average difference of probabilities of first-order neighbour cells.*

Remark. Theorem 4.1 shows, together with (4.7), that to any cell probability vector $p \in \mathscr{S}_m$ there corresponds a vector belonging to the anisotropic model, in such a way that Aitchison and Aitken's kernel estimator has for all $\theta \in [0, \frac{1}{2}]$ the same *MSSE* under both distributions. Therefore, restricting our considerations in the previous sections to the anisotropic model, we nevertheless obtain representative results on the general behaviour of the kernel estimator for large sparse data sets.

APPENDIX

We replace the constant C in (2.5) and (2.6) by $2^{-m}c$, where c is a new normalizing constant. Throughout our proofs our estimates of remainder terms are of the stated orders of magnitude uniformly in values of θ such that $0 \leqslant \theta \leqslant \varepsilon$, where $\varepsilon = \varepsilon(n)$ converges to zero so fast that $m\varepsilon^3 \to 0$.

Proof of Theorems 3.1 *and* 3.2

Our proof is in three parts.

Part (i): *Bias Contribution*

In this part we prove that if (b1) holds,

$$\sum_i (E\hat{p}_i - p_i)^2 = 2^{-m} c^2 m \delta^2 \theta^2 \{f'(0)^2 + o(1)\} \tag{A.1}$$

as $m \to \infty$; and if $f(x) = a_0 + a_1 x + a_2 x^2$ is a positive semi-definite quadratic then, for general m and θ,

$$\sum_i (E\hat{p}_i - p_i)^2 = 2^{-m} c^2 m\delta^2\theta^2[a_1^2 + 2m\delta^2\{(1-\theta)a_2\}^2(1-m^{-1})]. \quad \text{(A.2)}$$

Of course, c is defined by

$$c^{-1} = 2^{-m} \sum_i f\{\delta(|\underline{i}| - \tfrac{1}{2}m)\}. \quad \text{(A.3)}$$

Assuming (b1), we have $c = f(0) + o(1)$. Therefore, by (A.1),

$$\sum_i (E\hat{p}_i - p_i)^2 = 2^{-m} m\delta^2\theta^2[\{f(0)^{-1} f'(0)\}^2 + o(1)]. \quad \text{(A.4)}$$

In the case where f is a quadratic, formula (A.3) reduces to

$$c^{-1} = 2^{-m} \sum_i \{a_0 + a_2\delta^2(|\underline{i}| - \tfrac{1}{2}m)^2\} = a_0 + \tfrac{1}{4}a_2 m\delta^2. \quad \text{(A.5)}$$

Therefore, by (A.2),

$$\sum_i (E\hat{p}_i - p_i)^2 = 2^{-m}(a_0 + \tfrac{1}{4}a_2 m\delta^2)^{-2} m\delta^2 \times \theta^2[a_1^2 + m\delta^2\{(1-\theta)a_2\}^2(1-m^{-1})] \quad \text{(A.6)}$$

The remainder of this part of the proof is devoted to checking (A.1) and (A.2). First we state a lemma, whose proof follows by simple combinatorial identities for series of integers, and by Taylor expansion.

LEMMA. *Let C_1 denote an upper bound to $|f''|$, and define*

$$R_{1,m}(\underline{i}, \theta) = \tfrac{1}{2}C_1\delta^2\{m^2\theta^2(1 - 2m^{-1}|\underline{i}|)^2 + m\theta(1-\theta)\}.$$

Then

$$S_m(\underline{i}, \theta) \equiv \sum_{r=0}^{m} \theta^r(1-\theta)^{m-r} \sum_{k:|k-i|=r} f\{\delta(|\underline{k}| - \tfrac{1}{2}m)\}$$
$$= f\{\delta(|\underline{i}| - \tfrac{1}{2}m)\} + (1 - 2m^{-1}|\underline{i}|) f'\{\delta(|\underline{i}| - \tfrac{1}{2}m)\}\,\delta m\theta + R_{2,m}(\underline{i}, \theta),$$

where $|R_{2,m}(\underline{i}, \theta)| \leqslant R_{1,m}(\underline{i}, \theta)$. If f is a positive semi-definite quadratic, and $C_1 \equiv f''$, then $R_{2,m} = R_{1,m}$.

In the notation of the lemma, $E(\hat{p}_i) = 2^{-m} c S_m(\underline{i}, \theta)$ and

$$E(\hat{p}_i) - p_i = 2^{-m} c[(1 - 2m^{-1}|\underline{i}|) f'\{\delta(|\underline{i}| - \tfrac{1}{2}m)\}\, \delta m\theta + R_{2,m}(\underline{i}, \theta)].$$

Hence, writing B for a binomial $(m, \frac{1}{2})$ random variable, we have

$$\sum_{\underline{i}} (E\hat{p}_i - p_i)^2 = 2^{-m} c^2 E[(1 - 2m^{-1}B)^2 f'\{\delta(B - \tfrac{1}{2}m)\}^2](\delta m\theta)^2 + R_{3,m}(\theta), \tag{A.7}$$

where, for constants C_2 and C_3 depending only on f,

$$\begin{aligned} 2^m |R_{3,m}(\theta)| &\leqslant C_2[\delta^3 m\theta\{(m\theta)^2 E(|1 - 2m^{-1}B|^3) + m\theta E|1 - 2m^{-1}B|\} \\ &\quad + \delta^4\{(m\theta)^4 E(1 - 2m^{-1}B)^4 + (m\theta)^2\}] \\ &\leqslant C_3\theta^2(\delta^3 m^{3/2} + \delta^4 m^2). \end{aligned} \tag{A.8}$$

If $m^{1/2}\delta \to 0$ then

$$E[(1 - 2m^{-1}B)^2 f'\{\delta(B - \tfrac{1}{2}m)\}^2] = m^{-1} f'(0)^2 + o(m^{-1}).$$

Result (A.1) follows from this result, (A.7) and (A.8).

Assume next that $f(x) = a_0 + a_1 x + a_2 x^2$. Then $C_1 = 2a_2$, and by the lemma,

$$\begin{aligned} E(\hat{p}_i) - p_i = 2^{-m} c(1 - 2m^{-1}|\underline{i}|)\{a_1 - m\delta(1 - 2m^{-1}|\underline{i}|)\, a_2\}\, \delta m\theta \\ + \{(m\theta)^2(1 - 2m^{-1}|\underline{i}|)^2 + m\theta(1 - \theta)\}\, \delta^2 a_2, \end{aligned}$$

whence it may be proved that

$$\sum_{\underline{i}} (E\hat{p}_i - p_i)^2 = 2^{-m}(c\delta m\theta)^2[m^{-1}a_1^2 + 2\{\delta(1 - \theta)\, a_2\}^2(1 - m^{-1})].$$

This establishes (A.2).

Part (ii): *Variance Contribution*

Here we prove that, if $m\theta^3 \to 0$ and f, f', f'' are bounded, then

$$\sum_{\underline{i}} \operatorname{var}(\hat{p}_i) = n^{-1} e^{-2m\theta} + o\left\{\sum E(\hat{p}_i - p_i\right\}. \tag{A.9}$$

From the formula

$$\hat{p}_i = n^{-1} \sum_{r=0}^{m} N_{ir}\theta^r(1 - \theta)^{m-r},$$

where the variables $N_{i0}, ..., N_{im}$ are distributed as a multinomial with parameters n, $(q_{i0}, ..., q_{im})$, it follows after some algebra that

$$\begin{aligned}\sum_i \operatorname{var}(\hat{p}_i) &= \sum_i n^{-1}\left\{\sum_r q_{ir}\theta^{2r}(1-\theta)^{2m-2r} - (E\hat{p}_i)^2\right\} \\ &= n^{-1}\left\{\sum_{r=0}^{m}\binom{m}{r}\theta^{2r}(1-\theta)^{2m-2r} - \sum_i (E\hat{p}_i)^2\right\} \\ &= n^{-1}\{1+O(m\theta^3)\}e^{-2m\theta} - n^{-1}\sum_i (E\hat{p}_i)^2. \qquad \text{(A.10)}\end{aligned}$$

Next, observe that

$$\sum_i (E\hat{p}_i)^2 \leqslant 2\left\{\sum_i p_i^2 + \sum_i (E\hat{p}_i - p_i)^2\right\}.$$

If f'' is bounded, and $f(0) > 0$, then p_i is bounded by a constant multiple of $m2^{-m}$, which, if $\theta = \theta(m) \to 0$, is of smaller order than $e^{-2m\theta}$. In this case, $\sum p_i^2 = O(\max p_i) = o(e^{-2m\theta})$. In the case where $f(x) = a_0 + a_1 x + a_2 x^2$,

$$\begin{aligned}\sum_i p_i^2 &= 2^{-m}c^2\{a_0^2 + (a_1^2 + 2a_0a_2)\tfrac{1}{4}m\delta^2 + a_2^2(\tfrac{3}{10}m^2 - \tfrac{1}{8}m)\delta^4\} \\ &\leqslant 2^{-m}c^2C(1 + m\delta^2 + m^2\delta^4),\end{aligned}$$

where C' depends only on a_0, a_1, a_2. We may deduce from (A.5), and the fact that the quadratic $f(x)$ is positive semi-definite, that c is bounded by a constant multiple of $(1+m\delta^2)^{-1}$. Therefore, $\sum p_i^2 = O(2^{-m}) = o(e^{-2m\theta})$. Hence for both types of f, $\sum p_i^2 = o(e^{-2m\theta})$. The desired result (A.9) follows on substituting this formula into (A.10).

We should comment on the necessity of the assumption $m\theta^3 \to 0$, for results such as (A.9). It may be shown that if $m\theta^3$ does not converge to zero, then the variance contribution to mean summed squared error is greater than the asymptotic amount $n^{-1}e^{-2m\theta}$, claimed at (A.9), by at least a constant factor. If $m\theta^3 \to \infty$ then the variance exceeds $n^{-1}e^{-2m\theta}$ by a factor which diverges to $+\infty$. This observation makes it clear that the assumption $m\theta^3 \to 0$, which entails conditions such as (2.13), is necessary and sufficient for our formulae to hold as stated.

Part (iii): *Mean Squared Error*

Here we combine results from Parts (i) anf (ii) to obtain mean squared error formulae. Three distinct cases are identified: Case A, where $m\delta^2 \to 0$; Case B, where $m\delta^2 \to l$ and $0 < l < \infty$; and Case C, where $m\delta^2 \to \infty$.

Case A. $m\delta^2 \to 0$. Define $D_1(\theta) = 2^{-m} m\delta^2\theta^2\rho_1 + n^{-1}e^{2m\theta}$, where $\rho_1 = f(0)^{-2}f'(0)^2$. By (A.4) and (A.9),

$$D(\theta) \equiv \sum_i E(\hat{p}_i - p_i)^2 = D_1(\theta) + o\{D_1(\theta)\}.$$

Now, $D_1'(\theta) = 0$ when $2m\theta e^{2m\theta} = \lambda$, where $\lambda = 2^{m+1}m(\rho_1 n\delta^2)^{-1}$. We identify three different subcases, depending on the behaviour of λ. In subcases A.1 and A.2, $m\theta_0^3 \to 0$.

Subcase A1. $\lambda \to 0$. Here, $\min D(\theta) \sim D_1(\theta_0) \sim n^{-1}$, where $\theta_0 = (2m)^{-1}\lambda = 2^m(\rho_1 n\delta^2)^{-1}$.

Subcase A2. $\lambda \to \lambda_0$, $0 < \lambda_0 < \infty$. Let x_0 denote the solution of the equation $xe^x = \lambda_0$. Here, $\min D(\theta) \sim D_1(\theta_0) \sim n^{-1}(1 + \frac{1}{2}x_0)e^{-x_0} = n^{-1}q$, where $\theta_0 = (2m)^{-1}x_0$.

Subcase A3. $\lambda \to \infty$. Here, $\min D(\theta) \sim D_1(\theta_0) \sim n^{-1}(2\lambda)^{-1}(\log\lambda)^2$, where $\theta_0 = (2m)^{-1}(\log\lambda - \log\log\lambda)$. The condition $m\theta_0^3 \to 0$ is equivalent to (2.12).

Case B. $f(x) = a_0 + a_1x + a_2x^2$ and $m\delta^2 \to l$, $0 < l < \infty$. Define $D_1(\theta) = 2^{-m}\theta^2\rho_2 + n^{-1}e^{-2m\theta}$, where $\rho_2 = (a_0 + \frac{1}{4}a_2l)^2(a_1^2 + a_2^2l)l$. Let $q = q(\lambda_0)$ be as in Subcase A2, and let $\theta_0 = 0$ if $\lambda \to 0$, $(2m)^{-1}x_0$ if $\lambda \to \lambda_0$, $(2m)^{-1}(\log\lambda - \log\log\lambda)$ if $\lambda \to \infty$. Then by (A.6) and (A.9), $\min D(\theta) \sim D_1(\theta_0) \sim n^{-1}$ if $\lambda \to 0$, $n^{-1}q$ if $\lambda \to \lambda_0$, $n^{-1}(2\lambda)^{-1}(\log\lambda)^2$ if $\lambda \to \infty$. The condition $m\theta_0^3 \to 0$ is equivalent to (2.12).

Case C. $f(x) = a_0 + a_1x + a_2x^2$ and $m\delta^2 \to \infty$. The argument and conclusions in case B are valid as before, except that ρ_2 should be replaced by $\rho_3 = 16$ throughout.

Proof of Theorems 2.3 *and* 2.4

The proof follows closely the three part argument used to establish Theorems 2.1 and 2.2, and so is given only in bare outline. We may prove that

$$\sum_i (E\hat{p}_i - p_i)^2 = 2^{-m}(a_0 + \tfrac{1}{2}a_2\delta^2)^2 a_2^2 m^{-1}\delta^4\theta^2,$$

and that (A.9) continues to hold. From these results it may be shown that, with

$$D_1(\theta) = 2^{-m}(a_0 + \tfrac{1}{2}a_2\delta^2)^2 a_2^2 m^{-1}\delta^4\theta^2 + n^{-1}e^{2m\theta},$$

we have $D(\theta)/D_1(\theta) = 1 + o(1)$; and that the equation $D_1'(\theta) = 0$ is equivalent to $2m\theta e^{2m\theta} = \lambda$, where $\lambda = 2^{m+1}m^3(\rho n\delta^4)^{-1}$ and $\rho = (a_0 + \frac{1}{2}a_2\delta^2)^2 a_2^2$.

Proof of Theorem 3.1

We consider only the case where cell probabilities are given by (2.6), since the isotropic case (2.5) is similar. Our proof is based on the identity

$$SSE - MSSE = 2(1-\theta)^{2m}\left\{n^{-2}\sum\sum_{j_1<j_2} U_{j_1 j_2} + n^{-1}(1-n^{-1})\sum_{j=1}^{n} V_j\right\} - 2(1-\theta)^m n^{-1}\sum_{j=1}^{n} W_j, \tag{A.11}$$

where we use the notation $\theta_1 = \theta(1-\theta)$, $Y_{j_1 j_2} = \sum_i \theta_1^{|X_{j_1} - i| + |X_{j_2} - i|}$,

$$q_{ir} = \sum_{k:|i-k|=r} p_k, \qquad \mu_i(\theta) = \sum_{r=0}^{m} \theta_1^r q_{ir}, \qquad \mu(\theta) = \sum_i \mu_i(\theta)^2,$$

$$Z_j = E(Y_{j j_1} \mid X_j) = \sum_i \theta_1^{|X_j - i|}\mu_i(\theta) \qquad (\text{for } j_1 \neq j),$$

$$U_{j_1 j_2} = Y_{j_1 j_2} - Z_{j_1} - Z_{j_2} + \mu(\theta), \qquad V_j = Z_j - \mu(\theta),$$

$$V(\theta) = \sum_i p_i \mu_i(\theta), \qquad W_j = \sum_i p_i \theta_1^{|X_j - i|} - V(\theta).$$

Part (i): *Contribution of* $\sum V_j$ *and* $\sum W_j$

Define

$$T_1 = (1-\theta)^{2m} n^{-1}(1-n^{-1})\sum_{j=1}^{n} V_j - (1-\theta)^m n^{-1}\sum_{j=1}^{n} W_j.$$

We sketch a proof that

$$E(T_1^2) = o\left[\left\{\sum_i (E\hat{p}_i - p_i)^2 + n^{-1}e^{-2m\theta}\right\}^2\right]. \tag{A.12}$$

Put

$$\Delta_{1ij} = \theta^{|X_j - i|}, \qquad \Delta_{ij} = \Delta_{1ij} - E(\Delta_{1ij}),$$

$$T_{11} = (1-n^{-1})n^{-1}\sum_{j=1}^{n}\sum_i (E\hat{p}_i - p_i)\Delta_{ij},$$

$$T_{12} = -(1-\theta)^m n^{-2}\sum_{j=1}^{n} W_j.$$

In this notation, $T_1 = T_{11} + T_{12}$, and so

$$E(T_1^2) \leqslant 2E(T_{11}^2) + 2E(T_{12}^2).$$

It may be show that for constants $C_1, C_2 > 0$,

$$\begin{aligned} nE(T_{11}^2) &\leqslant C_1 2^{-m} m^6 \sum_i (E\hat{p}_i - p_i)^2 \\ &= o\left[\left\{\sum_i (E\hat{p}_i - p_i)^2\right\} n^{-1} 2^{-2m\theta}\right] \\ &= o\left[\left\{\sum_i (E\hat{p}_i - p_i)^2 + n^{-1} e^{-2m\theta}\right\}^2\right], \\ n^3 E(T_{12}^2) &\leqslant 2\left\{\sum_i p_i (E\hat{p}_i - p_i)^2 + \sum_i p_i^3\right\} \\ &\leqslant 2\left\{C_2 2^{-m} m^2 \sum_i (E\hat{p}_i - p_i)^2 + C_2^2 2^{-2m} m^4\right\} \\ &= o\left\{2^{-2m\theta} \sum_i (E\hat{p}_i - p_i)^2 + 2^{-4m\theta}\right\}. \end{aligned}$$

These results imply (A.12).

Part (ii): *Contribution of* $\sum\sum U_{j_1 j_2}$

Define

$$T_2 = (1-\theta)^{2m} n^{-2} \sum_{j_1 < j_2} \sum U_{j_1 j_2}, \qquad T_3 = \sum_{j_1 < j_2} \sum U_{j_1 j_2}.$$

Observe that

$$\begin{aligned} E(T_3^2) &= \sum_{j_1 < j_2} \sum E(U_{j_1 j_2}^2) = \tfrac{1}{2} n(n-1)\, E(U_{12}^2) \\ &\leqslant 4n^2 E(Y_{12}^2) = 4n^2 \sum_{i'} \sum_{i''} p_{i'} p_{i''} \left(\sum_i \theta_1^{|i-i'|+|i-i''|}\right)^2. \end{aligned}$$

Since $p_i \leqslant C_3 2^{-m} m^2$, then

$$\begin{aligned} E(T_3^2) &= O\left\{n^2 2^{-2m} m^4 \sum_{i'} \sum_{r=0}^{m} \sum_{i'':\,|i-i'|+|i-i''|=r} \left(\sum_i \theta_1^{|i-i'|+|i-i''|}\right)^2\right\} \\ &= O\left\{n^2 2^{-2m} m^4 \sum_{i'} \sum_{r=0}^{m} \sum_{i'':\,|i'-i''|=r} (2\theta_1)^{2r} (1+\theta_1^2)^{2m-2r}\right\} \\ &= O\left\{n^2 2^{-m} m^4 \sum_{r=0}^{m} \binom{m}{r} (2\theta_1)^{2r} (1+\theta_1^2)^{2m-2r}\right\} \\ &= O\left[n^2 2^{-m} m^4 \{(2\theta_1)^2 + (1+\theta_1^2)^2\}^m\right]. \end{aligned}$$

It follows that $E(T_2^2) = O(n^{-2} 2^{-m}) = o\{(n^{-1} e^{-2m\theta})^2\}$, whence by (A.11) and (A.12),

$$E(SSE - MSSE)^2 = o\left[\left\{\sum_i (E\hat{p}_i - p_i)^2 + n^{-1}\right\}^2\right].$$

Theorem 3.1 follows from this result and (A.9).

The proof of Theorem 3.2 is broadly similar to that of Theorem 3.1, and so will not be given here.

Proof of Theorem 4.1

We derive only Part (1). Let $\alpha > 0$, $\gamma_1, ..., \gamma_m > 0$ and $\delta > 0$ be arbitrary, provided $\mathscr{S}_m(\alpha, \gamma_1, ..., \gamma_m) \cap \mathscr{S}_m \neq \varnothing$. It is sufficient to prove the existence of a cell probability vector $p \in \mathscr{S}_m(\alpha, \gamma_1, ..., \gamma_m)$ such that

$$|\underline{i}| = |\underline{i}'| \text{ implies } p_{\underline{i}} = p_{\underline{i}'} \qquad \text{for all} \quad \underline{i}, \underline{i}' \in \mathscr{I}. \tag{A.13}$$

We prove that (A.13) is valid for all vectors p that are given by (4.5) with the parameters

$$\beta_{k;J(k)} = \binom{m}{k}^{-1/2} \gamma_k^{1/2} d_k \qquad \text{for all} \quad J(k), k = 1, ..., m, \tag{A.14}$$

and arbitrary $d_k \in \{1, -1\}$.

Therefore, let us consider an arbitrary cell $\underline{i}$ with $|\underline{i}| = v \geqslant 1$ and any of its first-order neighbours, say, cell $\underline{i}'$ with $|\underline{i}'| = v - 1$, that differs from $\underline{i}$ only in variable l, so that

$$i_j = i_j' \qquad \text{for all} \quad j \neq l, \qquad \text{and} \qquad i_l = 1, i_l' = 0. \tag{A.15}$$

Due to the lexicographical ordering of the cells, the unit vector $e_{\underline{i}}$ with 1 in cell $\underline{i}$ and 0's else has the representation $e_{\underline{i}} = \bigotimes_{j=1}^{m} e_{i_j}$, where $e_{i_j} = (1, 0)^{\mathrm{T}}$ if $i_j = 0$, and $e_{i_j} = (0, 1)^{\mathrm{T}}$ else. Thus, we obtain

$$p_{\underline{i}} - p_{\underline{i}'} = p^{\mathrm{T}}(e_{\underline{i}} - e_{\underline{i}'}) = p^{\mathrm{T}}\left\{\bigotimes_{j=1}^{m} e_{i_j} \rightleftharpoons (-1, 1)^{\mathrm{T}}\right\}. \tag{A.16}$$

The special structure of the eigenvectors $h_{k;J(k)}$, see (4.4), and (A.15) provide

$$h_{k;J(k)}^{\mathrm{T}}\left\{\bigotimes_{j=1}^{m} e_{i_j} \rightleftharpoons (-1, 1)^{\mathrm{T}}\right\} = \begin{cases} 2^{-m/2}(-1)^{\mu} & \text{if} \quad l \in J(k) \\ 0 & \text{else,} \end{cases} \tag{A.17}$$

where $\mu = |\{j \mid j \in J(k) \text{ and } i_j = 1\}|$.

Combining (A.16), (4.5), (A.14), and (A.17), we obtain

$$p_i - p_{i'} = \alpha 2^{1-m/2} \sum_{k=1}^{m} \binom{m}{k}^{1/2} \gamma_k^{1/2} d_k \times \left\{ \sum_{\mu=1}^{k} \binom{v-1}{\mu-1} \binom{m-v}{k-\mu} (-1)^{\mu} \right\}. \tag{A.18}$$

Obviously, (A.18) depends on $\underline{i}$ only via $|\underline{i}| = v$. Formula (A.13) follows immediately.

ACKNOWLEDGMENTS

Part of the work of the first author was supported by a Research Fellowship at CORE, Université Catholique de Louvain, Belgium. Work of the second author was supported by a CORE Research Fellowship. We are grateful to a referee for detailed and helpful suggestions, which led to this substantially shorter, more succinct version of the paper.

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. BROWN, P. J., AND RUNDALL, P. W. K. (1985). Kernel estimates for categorical data. *Technometrics* **27** 293–299.
5. BURMAN, P. (1987). Smoothing sparse contingency tables, *Sankhya Ser. A* **49** 24–36.
6. BURMAN, P. (1987). Central limit theorems for quadratic forms for sparse tables. *J. Multivariate Anal.* **22** 258–277.
7. GRUND, B. (1987). *Schätzungen für die Zellwahrscheinlichkeiten in multinomialverteilten Kontingenztafeln*. Doctoral thesis, Humboldt-Universität, Berlin.
8. GRUND, B. (1991). Kernel estimators for cell probabilities. To appear in *J. Multivariate Anal.*
9. HALL, P. (1981). On parametric multivariate binary discrimination. *Biometrika* **69** 287–294.
10. HALL, P., AND TITTERINGTON, D. M. (1987). On smoothing sparse multinomial data. *Austral. J. Statist.* **29** 19–37.
11. IGHODARO, A., AND SANTNER, T. J. (1982). Ridge type estimators of multinomial cell probabilities. In *Statistical Decision Theory and Related Topics III*, Vol. 2. Academic Press, New York.
12. SANTNER, T. J., AND DUFFY, D. E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

13. SIMONOFF, J. S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Ann. Statist.* **11** 208–218.
14. SIMONOFF, J. S. (1987). Probability estimation via smoothing in sparse contingency tables with ordered categories. *Statist. Probab. Lett.* **5** 55–63.
15. SUTHERLAND, M., HOLLAND, P. W., 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.
16. TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22** 259–268.