The Annals of Statistics
1983, Vol. 11, No. 1, 208–218

A PENALTY FUNCTION APPROACH TO SMOOTHING LARGE SPARSE CONTINGENCY TABLES

BY JEFFREY S. SIMONOFF

New York University

Probabilities in a large sparse contingency table are estimated by maximizing the likelihood modified by a roughness penalty. It is shown that if certain smoothness criteria on the underlying probability vector are met, the estimator proposed is consistent in a one-dimensional table under a sparse asymptotic framework. Suggestions are made for techniques to apply the estimator in practice, and generalization to higher dimensional tables is considered.

1. Introduction. When the number of observations in a large table is close to the number of cells, standard estimation techniques are inappropriate. In this paper maximum penalized likelihood is used to estimate probabilities in large sparse tables. It is assumed that the true underlying probability vector π satisfies certain smoothness criteria, so that information in nearby cells is useful in estimation. The most obvious example of a contingency table that might satisfy the criteria is one with ordered categories.

Previous work in the field can be divided into three major areas: large sparse tables, tables with ordered categories and nonparametric density estimation. Fienberg and Holland (1973) examine estimators of the form

$$\hat{p}_i = (n_i + \alpha_i)/(N + \textstyle\sum_1^K \alpha_i) \tag{1.1}$$

(this is the Bayes estimator using a squared-error loss function if the prior is Dirichlet with parameter $\boldsymbol{\alpha}$). Here K is the number of cells in the table, and N is the number of observations. In order to examine the behavior of these estimators in large sparse multinomials, they develop a "sparse asymptotics," by letting $N \to \infty$ and $K \to \infty$ such that $N/K = \delta$, a constant. By putting a smoothness constraint on the probabilities, specifically $\pi_i = K^{-1}f((i - ½)/K)$ where f has a continuous second derivative, they show that their choice of $\boldsymbol{\alpha}$ results in an estimator with asymptotic mean squared error smaller than the mean squared errors of the frequency estimator $\hat{\pi}_i = n_i/N$ and the common estimators of the form (1.1). However, the estimator is not consistent (consistency in the sparse asymptotic framework being defined as the property $\sup_i|(\hat{p}_i/\pi_i) - 1| \to_p 0$).

Leonard (1973) develops a Bayesian framework for a one-dimensional table with ordered categories. He puts a multivariate normal prior distribution on $\boldsymbol{\gamma} = \log \mathbf{p}$, and incorporates ideas of smoothing by assuming that the prior covariance matrix has the form $\operatorname{Cov}(\gamma_i, \gamma_j) = \sigma^2\rho^{|i-j|}$. This is equivalent to using the prior proportional to

$$\textstyle\sum_{i=1}^{K-1} (\rho \log p_i - \log p_{i+1})^2. \tag{1.2}$$

Leonard does not examine the sparse asymptotics when using this prior. Leonard (1975) generalizes this work to two-dimensional tables by putting the prior on $x_{ij} = p_{ij}/p_{i\cdot}p_{\cdot j}$; again, ordering of categories can be reflected in the form of the prior covariance matrix. Laird (1978) provides a slight modification of this technique. Leonard (1978) proposes a general framework where $f(t)$ (the density being estimated) is a random process; his proposed prior then involves second differences of log probabilities, rather than first differences.

Received July 1981; revised July 1982.

AMS 1970 *subject classifications.* Primary 62G05; secondary 62E20.

Key words and phrases. Large sparse contingency tables, smoothing of probability estimates, maximum penalized likelihood, sparse asymptotics.

Techniques for smoothing probabilities have a close relationship to nonparametric density estimation techniques, since both require smoothness assumptions about the underlying probabilities. The most commonly used estimator is the Parzen-Rosenblatt kernel estimator (Rosenblatt, 1956; Parzen, 1962); if probability estimates are restricted to be nonnegative, the optimal convergence rate of the estimator is $\int \{\hat{f}_N(x) - f(x)\}^2\, dx = O(N^{-4/5})$.

Good and Gaskins (1971) introduced an important density estimation technique which they call maximum penalized likelihood. They propose estimating probabilities by maximizing the function $L = \text{log likelihood} - \Phi(f)$, where Φ is a roughness penalty. They propose two penalties (where f is the density being estimated):

$$\Phi(f) = \alpha \int_{-\infty}^{\infty} \frac{\{f'(x)\}^2}{f(x)}\, dx + \frac{\beta}{4} \int_{-\infty}^{\infty} \left[\frac{\{f''(x)\}^2}{f(x)} + \frac{\{f'(x)\}^2}{4f^3(x)} - \frac{f'(x)f''(x)}{f^2(x)} \right] dx \tag{1.3}$$

and

$$\Phi(f) = (\alpha/4) \int_{-\infty}^{\infty} \{f'(x)/f(x)\}^2\, dx. \tag{1.4}$$

Note that this is equivalent to Leonard's approach with a more complicated prior. They also give guidelines for how to choose the parameters of the penalties from simulations. Good and Gaskins (1980) look at penalty (1.3) above and set $\alpha = 0$; they then recommend estimating β from the data using goodness-of-fit tests. The existence and uniqueness of the two distinct estimators using these penalties is shown by deMontricher, Tapia and Thompson (1975).

Tapia and Thompson (1978, Chapter 5) examine the method of maximum penalized likelihood in the case where the data is presented in bins, rather than continuously (this corresponds to a one-dimensional contingency table). The roughness penalty which they propose is

$$\Phi = \alpha K^3 \textstyle\sum_{i=1}^{K-1} (p_i - p_{i+1})^2. \tag{1.5}$$

They show that the estimator $\hat{p}_i$ is consistent if $K = N^q$, $0 < q < 1/4$; Monte Carlo simulations indicate that it has properties similar to kernel estimators.

The penalty function proposed here is a special case of Leonard's prior (1.2) with $\rho = 1$ and can be viewed as a discrete version of (1.4). The novelty of the present work is the sparse asymptotics that will be presented in the next section. These asymptotics are a generalization of Fienberg and Holland's model of $N/K = \delta$, and provide a framework where large sparse multinomials can be studied.

2. Probability estimation.

2.1 *The model.* Consider a one-dimensional table $\{n_i\}$, $i = 1, \cdots, K$, generated from $\{p_i\}$ by a multinomial likelihood:

$$\ell(\mathbf{n} \mid \mathbf{p}) \propto \textstyle\prod_{i=1}^{K} p_i^{n_i}, \sum_{i=1}^{K} p_i = 1.$$

The prior distribution (or penalty function) is defined, as outlined above, such that the log of the posterior is

$$L(\mathbf{p} \mid \mathbf{n}) = \textstyle\sum_{i=1}^{K} n_i \log p_i - \beta \sum_{i=1}^{K-1} \{\log(p_i/p_{i+1})\}^2, \quad \sum_{i=1}^{K} p_i = 1, \beta \geq 0$$

(omitting constants that do not depend on $\mathbf{p}$). The "true" value of $\mathbf{p}$, which we are trying to estimate, is denoted by π.

The MLE $\hat{\pi}$ is defined as the value of $\mathbf{p}$ that maximizes $\ell(\mathbf{n} \mid \mathbf{p})(\hat{\pi}_i = n_i/N)$. The MPE ("maximum posterior estimator") is defined as the value of $\mathbf{p}$ that maximizes $L(\mathbf{p} \mid \mathbf{n})$. The

MPE will be a "smoothed" estimator because the maximization process forces the ratio of adjacent probabilities towards one. It is shown in Theorem 2.1 that a unique maximizing value exists.

THEOREM 2.1. *Let $L(\mathbf{p} \mid \mathbf{n})$ be defined as above. A unique value $\hat{\mathbf{p}}$ maximizing $L(\mathbf{p} \mid \mathbf{n})$ exists and is given by the solution to the set of equations* (2.1):

$$\begin{aligned} & n_1 + 2\beta(x_2 - x_1) && -N\exp(x_1) = 0, \\ & \qquad\vdots && \qquad\vdots \\ & n_1 + 2\beta(x_{i+1} - 2x_i + x_{i-1}) && -N\exp(x_i) = 0, \quad i = 2, \cdots, K-1, \\ & \qquad\vdots && \qquad\vdots \\ & n_K + 2\beta(x_{K-1} - x_K) && -N\exp(x_K) = 0, \end{aligned} \tag{2.1}$$

where $x_i = \log p_i$.

PROOF. Let $x_i = \log p_i$. Then

$$L(\mathbf{p} \mid \mathbf{n}) = \textstyle\sum_{i=1}^{K} n_i x_i - \beta \sum_{i=1}^{K-1} (x_i - x_{i+1})^2, \sum_{i=1}^{K} \exp(x_i) = 1.$$

By the theory of Lagrange multipliers, any local maximum of L will satisfy equations (2.1). This is equivalent to the unconstrained maximization of $L' = \sum_1^K n_i x_i - \beta \sum_1^{K-1} (x_i - x_{i+1})^2 - N \sum_1^K \exp(x_i)$. But L' is strictly concave, and approaches $-\infty$ as any $x_i \to \pm\infty$; hence, there is one and only one stationary point (a global maximum), and it solves (2.1).

2.2 *Sparse asymptotics.*

THEOREM 2.2. *Let $\hat{\mathbf{p}}$ be the* MPE. *Let N, K, $\beta \to \infty$ with $0 < \gamma_1 < N/K < \gamma_2 < \infty$; assume $\beta K^{-2} \to 0$ and $K^{4/3}(\log K)^{2/3}\beta^{-1} \to 0$. Assume $\exists\, M \in (1, \infty)$ such that $0 < (MK)^{-1} < \pi_1 < M/K < 1$ for all i; also, assume the smoothness constraint*

$$\sup_i \left| \log(\pi_{i-1}\pi_{i+1}/\pi_i^2) \right| = O(K^{-2}).$$

Then

$$\sup_i \left| (\hat{p}_i/\pi_i) - 1 \right| = O_p(\beta^{-1/4}(\log K)^{1/2} + \beta K^{-2}).$$

(Note this implies that taking β of the order $K^{8/5}(\log K)^{2/5}$ results in the rate of convergence

$$\sup_i \left| (\hat{p}_i/\pi_i) - 1 \right| = O_p(K^{-2/5}(\log K)^{2/5})).$$

It is worth noting that the smoothness constraint assumed in the theorem is implied by the usual density estimation assumption that the log of the density being estimated has a bounded second derivative (which means the density itself has bounded second derivative and is bounded away from zero in the region of interest).

In order to prove this result, we need a number of lemmas; several are presented with only sketch proofs.

LEMMA 2.3. *Consider the linear equations*

$$\begin{aligned} & y_1 - a_1 x_1 && + B(x_2 - x_1) = 0, \\ & && \quad\vdots \\ & y_i - a_i x_i && + B(x_{i+1} - 2x_i + x_{i-1}) = 0, \quad i = 2, \cdots, K-1, \\ & && \quad\vdots \\ & y_K - a_K x_K && + B(x_{K-1} - x_K) = 0, \end{aligned} \tag{2.2}$$

with $a_i > 0$ *all* i *and* $B > 0$. *Then* $y_i \geq 0\ \forall i \Rightarrow x_i \geq c_i x_{i+1},\ 0 < c_i < 1$, *and* $y_i \geq 0\ \forall i \Rightarrow x_i \geq \tilde{c}_i x_{i-1},\ 0 < \tilde{c}_i < 1$.

Proof. This is proven by induction on i; if $y_i \geq 0$, the equations (2.2) become inequalities. Successive summation of the equations gives the result. □

Lemma 2.4. *For* $a_i > 0$, $B > 0$, *the tridiagonal matrix*

$$F = \begin{pmatrix} B + a_1 & -B & & & 0 \\ -B & 2B + a_2 & -B & & \\ & \ddots & & \ddots & \\ & -B & 2B + a_{K-1} & -B & \\ 0 & & \ddots & & \\ & & -B & B + a_K \end{pmatrix}$$

is nonsingular. Furthermore, if we call its inverse H, then each row of H has the property that the elements increase monotonically to the diagonal, and then decrease.

Proof.

(a) The equations (2.2) may be written as $F\mathbf{x} = \mathbf{y}$. Setting $\mathbf{y} = \mathbf{0}$ and summing the equations (2.2) implies that F is nonsingular. (b) Since the row $\mathbf{h}_i (h_{ij}, j = 1, \cdots, K)$ is the solution $\{x_j\}$ to the equations (2.2) when $y_j = 0, j \neq i$ and $y_j = 1, j = i$, we have $x_{j+1} - x_j \geq 0$ for $j < i$ and $x_{j+1} - x_j \leq 0$ for $j \geq i$. This implies the result given. □

Lemma 2.5. *Write the solution to the equations* (2.2) *as* $\hat{x} = H\mathbf{y}$. *Let* $K \to \infty$ *and assume* $B \to \infty$ *with* $BK^{-2} \to 0$. *Assume* $\exists\, M \in (1, \infty)$ *such that* $M^{-1} < a_i < M\ \forall i$ *as* $K \to \infty$. *Then* (a) $h_{ij} \geq 0\ \forall ij$, (b) $\sup_i (\sum_{j=1}^{K} h_{ij}) \leq M$, (c) $\sup_{ij}(h_{ij}) = O(B^{-1/2})$, (d) $\{\inf_i (\sum_{j=1}^{K} h_{ij}^2)\}^{-1} = O(B^{1/2})$.

Proof.

(a) This is equivalent to showing $y_i \geq 0\quad \forall i \Rightarrow x_i \geq 0\ \forall i$, which is a direct consequence of Lemma 2.3.

(b) A direct substitution into (2.2) yields that $y_i = 1\ \forall i \Rightarrow \sup_i(x_i) \leq \sup_i(a_i^{-1}) \leq M$ which was the result to be proven.

(c) Using the parameterization of Lemma 2.4 (b), we have

$$\textstyle\sum_{p=1}^{j} a_p x_p = \sum_{p=1}^{j} y_p - B(x_j - x_{j+1}).$$

By part (b), this means

$$\left| \textstyle\sum_{p=1}^{j} y_p - \sum_{p=1}^{j} a_p x_p \right| \leq M^2$$

which implies that adjacent entries in $\mathbf{h}_i$ differ by at most $M^2 B^{-1}$. Since $\sum_{j=1}^{i+B^{1/2}} h_{ij} \leq M$ (by part (b)), this implies

$$h_{ii} \leq (M^2 B^{1/2} + 2MB^{1/2} + M^2)/2B, \text{ or } \sup(h_{ij}) = O(B^{-1/2}).$$

(d) Consider again the ith row of H, $\mathbf{h}_i$(denoted $\mathbf{x}$). Then, by part (b),

$$\textstyle\sum_{j=1}^{i} a_j x_j \leq M x_i \sum_{j=1}^{i} \{1 + (i-j)^2/2MB\}^{-1} \leq M x_i \{1 + \tfrac{1}{2}(2MB\pi^2)^{1/2}\}. \tag{2.3}$$

By a similar argument,

$$1 = \textstyle\sum_{j=1}^{K} a_j x_j \leq M x_i \{1 + (2MB\pi^2)^{1/2}\} \tag{2.4}$$

implying

$$h_{ii} \equiv x_i \geq [M\{1 + (2MB\pi^2)^{1/2}\}]^{-1}. \tag{2.5}$$

Since h_{ii} differs from h_{ij} by at most $(j-i)M^2/B$, it can be shown that

$$\sum_{j=1}^{K} h_{ij}^2 \geq Q/[8M^2\{MB\pi^2 + (2MB\pi^2)^{1/2} + \tfrac{1}{2}\}], \tag{2.6}$$

where $Q = B/[2M^3\{1 + (2MB\pi^2)^{1/2}\}]$. This implies that $B^{1/2} \sum_{j=1}^{K} h_{ij}^2$ is bounded away from zero as $K \to \infty$; therefore, $(\inf_i \sum_j \mathrm{h}_{ij}^2)^{-1} = O(B^{1/2})$. □

A matrix that satisfies the conditions (a) – (d) above will be said to satisfy Condition 1. We now prove a lemma that gives the general asymptotic framework for a particular class of estimates; we then show that the MPE is a member of this class.

LEMMA 2.6. *Suppose the vector* $\mathbf{u}$ *satisfies* $\mathbf{u} = G\mathbf{z}$, $\sum_{i=1}^{K} \exp(u_j) = K$, *where* $\mathbf{z}$ *is defined by*

$$z_j = (n_j - N\pi_j) + N\pi_j\{1 + u_j - \exp(u_j)\} + r_j, \quad j = 1, \cdots, K.$$

The vector $\mathbf{n}$ *is a random vector distributed as Multinomial* (N, π). *Assume* $\exists\, M \in (1, \infty)$ *such that* $0 < (MK)^{-1} < \pi_i < M/K < 1\ \forall i$. *Let* $N, K, B \to \infty$ *with* $0 < \gamma_1 < N/K < \gamma_2 < \infty$, $K^{4/3}(\log K)^{2/3}B^{-1} \to 0$ *and* $BK^{-2} \to 0$. *Assume* $\tilde{r} \equiv \sup_j(r_j) = O(B^{1/2}K^{-1})$. *Let* $G = (g_{ij})$ *satisfy Condition* 1 (*from Lemma* 2.5). *Then* $\sup_i |u_i| = O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r})$.

PROOF. We will examine the behavior of u_i in two parts: Write $u_i = W_i + v_i$, where $w_i = \sum_{j=1}^{K} g_{ij}(n_j - N\pi_j)$, $v_i = \sum_{j=1}^{K} g_{ij}N\pi_j\{1 + u_j - \exp(u_j)\}$ and $W_i = \sum_{j=1}^{K} g_{ij}r_j + w_i$.

(1) Let $\mathbf{Y}_p$, $p = 1, \cdots, N$, be defined by

$$Y_{jp} = \begin{cases} 1 & \text{if } p\text{th observation falls in } j\text{th cell} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{Y}_1, \cdots, \mathbf{Y}_N$ are independent of each other. Define X_p by $X_p = \mathbf{G}_i'(\mathbf{Y}_p - \pi)$, where $\mathbf{G}_i$ is the ith column of G. Then the X_p's are independent of each other and

$$E(X_p) = 0 \tag{2.7}$$

$$V(X_p) = \sum_{j=1}^{K} g_{ij}^2\pi_j(1-\pi_j) - 2\sum_{\substack{m=1\\m\neq j}}^{K}\sum_{j=1}^{K} g_{im}g_{ij}\pi_m\pi_j \tag{2.8}$$

$$|X_p| \leq \sup_{ij}(g_{ij}) \tag{2.9}$$

and

$$w_i = \sum_{p=1}^{N} X_p.$$

Condition 1 implies $C_1 < KB^{1/2}V(X_p) < C_2$ for C_1 and C_2 constants greater than zero. Bennett (1962) showed that for iid random variables $S_1, \cdots, S_N$ with $E(S_i) = 0$, $|S_i| \leq b$, then for all x

$$P\left(\left|\sum_{i=1}^{N} S_i\right| \geq x\right) \leq 2\exp\left(-\frac{x}{b}\left[\left\{1 + \frac{NV(S_i)}{bx}\right\}\log\left\{1 + \frac{bx}{NV(S_i)}\right\} - 1\right]\right). \tag{2.10}$$

Applying this result to w_i gives

$$\begin{aligned} \sup_i P(|w_i| \geq t_K) &\leq 2\exp\left[-\frac{t_K}{\sup(g_{ij})}\left\{\left(1 + \frac{KV(X_p)}{\sup(g_{ij})t_K}\right)\right.\right. \\ &\qquad \left.\left.\cdot\log\left(1 + \frac{\sup(g_{ij})t_K}{KV(X_p)}\right) - 1\right\}\right] \\ &\leq 2\exp\left\{-t_K^2\left(\frac{B^{1/2}}{2C_2} - \frac{B^{1/2}C_3t_K}{2C_1^2}\right)\right\}, \end{aligned} \tag{2.11}$$

where $\sup_i(g_{ij}) \leq C_3B^{-1/2}$. Take $t_K = B^{-1/4}\lambda_K(\log K)^{1/2}$, where $\lambda_K \to \infty$, $t_K \to 0$ as $K \to \infty$; then

$$\sup_i P(|w_i| \geq t_K) \leq 2\exp\left[-\lambda_K^2(\log K)\left\{(2C_2)^{-1} - (C_3t_K)/2C_1^2\right\}\right].$$

It is easy to show that

$$(2.12)\qquad P(\sup_i |w_i| \geq t_K) \leq K \sup_i P(|w_i| \geq t_K)$$
$$\leq 2K^{1-\lambda_K^2\{1/2C_2-(C_3t_K)/2C_1^2\}} \rightarrow 0 \text{ as } K \rightarrow \infty.$$

This implies that $P(\sup_i |w_i| \geq B^{-1/4}\lambda_K(\log K)^{1/2}) \rightarrow 0$ as $K \rightarrow \infty$ with $\lambda_K \rightarrow \infty$, and hence that $\sup_i |w_i| = O_p(B^{-1/4}(\log K)^{1/2})$. But $\sup_i \sum_j g_{ij} r_j = O(\tilde{r})$, so

$$\sup_i |W_i| = O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r}).$$

(2) (a) Since $v_i \leq 0$ (with equality only at $u_j = 0\ \forall j$) and $u_i = W_i + v_i$, for all $u_i > 0$, $|v_i| \leq W_i$. Thus $\sup^* |v_i| = O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r})$, where $\sup^*(s_i)$ is defined as $\sup(s_i)$ over all cells i such that $u_i > 0$. Therefore $\sup^* |u_i| = O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r})$. That is, we have shown that the positive u_i's are $O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r})$.

(b) We now examine the negative u_i terms. We can divide all of the u_i's into three groups: there are q u_i's such that $u_i \leq -\varepsilon$ (Group G_1), $K - p - q$ u_i's such that $-\varepsilon < u_i \leq 0$ (Group G_2), and p u_i's such that $u_i > 0$ (Group G_3) for some constant $\varepsilon \in (0, 1)$. Then some algebra gives

$$\mathrm{K} = \textstyle\sum_{i=1}^K \exp(u_i) \leq qe^{-\varepsilon} + K - p - q + p\exp(O_p(\sup{}^* u_i))$$

which implies

$$q \leq p\{\exp(O_p(\sup{}^* u_i)) - 1\}/(1 - e^{-\varepsilon}).$$

This means that $q = O_p(p \sup^* u_i)$; so, even if $p \sim K$, the biggest q can be is $O_p(K \sup^* u_i)$.

(c) Let $u_{\min}$ be the smallest value of u_i, occurring at cell I; say $u_{\min} \leq -\varepsilon$. Then $u_{\min} = W_I + v_I$. We examine the three groups G_1, G_2, G_3 (of v_I) separately.

$$\textstyle\sum_{G_1} g_{Ij} N\pi_j\{1 + u_j - \exp(u_j)\} \geq \sum_{u_j \leq -\varepsilon}(\gamma_1/M)\sup(g_{Ij})\{1 + u_{\min} - \exp(u_{\min})\}$$

$$\text{(i)}\qquad \geq O_p(KB^{-1/2}\sup{}^* u_i)\{1 + u_{\min} - \exp(u_{\min})\}$$
$$\geq O_p(KB^{-1/2}\sup{}^* u_i)u_{\min}.$$

$$\text{(ii)}\qquad \textstyle\sum_{G_2} g_{Ij} N\pi_j\{1 + u_j - \exp(u_j)\} \geq \sum_{-\varepsilon < u_j \leq 0} g_{Ij} N\pi_j(-\varepsilon^2/2) \geq \varepsilon^2 O_p(1).$$

(iii) By part (a), it is easy to show

$$\textstyle\sum_{G_3} g_{Ij} N\pi_j\{1 + u_j - \exp(u_j)\} \geq O_p((\sup{}^* u_i)^2).$$

Parts (i), (ii) and (iii) above imply

$$(2.13)\qquad u_{\min} \geq W_I + \varepsilon^2 O_p(1) + O_p((\sup{}^* u_i)^2)$$

as $K \rightarrow \infty$. Since $W_I = o_p(1)$, $|u_{\min}| = o_p(1)$, and so $\sup_i |u_i| = o_p(1)$. Using a Taylor Series expansion of $\exp(u_j)$ then implies

$$u_i = W_i + O_p(\sup_i u_i^2).$$

This finally implies

$$(2.14)\qquad \sup_i |u_i| = O_p(B^{-1/4}(\log K)^{1/2}) + O(\tilde{r}). \square$$

We now use the previous lemmas to prove Theorem 2.2

Proof of Theorem 2.2. Let $u_i = \log(p_i/\pi_i)$. Then

$$L = \textstyle\sum_{i=1}^K n_i u_i - \beta \sum_{i=1}^{K-1} \{u_i - u_{i+1} + \log(\pi_i/\pi_{i+1})\}^2 + \sum_{i=1}^K n_i \log \pi_i, \quad \sum_{i=1}^K \exp(u_i) = K.$$

The optimal $\mathbf{\hat{u}}$ satisfies the equations (2.2) of Lemma (2.5) with

$$a_i = N\pi_i, \quad B = 2\beta, \quad y_j = (n_j - N\pi_j) + N\pi_j\{1 + u_j - \exp(u_j)\} + 2\beta\log(\pi_{j-1}\pi_{j+1}/\pi_j^2).$$

We have used the notation $\mathbf{u} = H\mathbf{y}$ to express these equations previously.

By the assumptions of Theorem 2.2 and the results of Lemma 2.5, all of the conditions of Lemma 2.6 are satisfied, so

$$\sup_i |u_i| = O_p(\beta^{-1/4}(\log K)^{1/2}) + O\left(2 \sup_i \textstyle\sum_j h_{ij}\beta \left| \log \frac{\pi_{j-1}\pi_{j+1}}{\pi_j^2} \right|\right) \tag{2.15}$$

or

$$\sup_i |u_i| = O_p(\beta^{-1/4}(\log K)^{1/2} + \beta K^{-2}).$$

This means

$$\sup_i |\log(\hat{p}_i/\pi_i)| = O_p(\beta^{-1/4}(\log K)^{1/2} + \beta K^{-2}) \tag{2.16}$$

which implies

$$\sup_i |(\hat{p}_i/\pi_i) - 1| = O_p(\beta^{-1/4}(\log K)^{1/2} + \beta K^{-2}).$$

Taking $\beta \sim (K^{8/5}(\log K)^{2/5})$ implies

$$\sup_i |(\hat{p}_i/\pi_i) - 1| = O_p(K^{-2/5}(\log K)^{2/5}). \square \tag{2.17}$$

As a result of the asymptotic analysis, we can make more specific statements about the approximate distribution of $\hat{\mathbf{p}}$. Theorem 2.2 shows that if $\hat{p}_i = \pi_i(1 + \delta_i)$, then $\boldsymbol{\delta}$ satisfies

$$\log(1 + \delta_i) = \textstyle\sum_{j=1}^K h_{ij}[(n_j - N\pi_j) + 2\beta \log(\pi_{j-1}\pi_{j+1}/\pi_j^2) + N\pi_j\{\log(1 + \delta_j) - \delta_j\}]. \tag{2.18}$$

Since $\delta_j = o_p(1)$, terms in δ_j^2 may be neglected. Therefore

$$\delta_i \doteq \textstyle\sum_{j=1}^K h_{ij}\{(n_j - N\pi_j) + 2\beta \log(\pi_{j-1}\pi_{j+1}/\pi_j^2)\}. \tag{2.19}$$

We can use this result to get the asymptotic form of $\hat{\mathbf{p}}$; namely

$$\begin{aligned} \hat{p}_i &= \pi_i + \pi_i \textstyle\sum_{j=1}^K h_{ij}\{(n_j - N\pi_j) + 2\beta \log(\pi_{j-1}\pi_{j+1}/\pi_j^2)\} \\ &= \pi_i(1 + O_p(K^{-2/5}(\log K)^{2/5})). \end{aligned} \tag{2.20}$$

This is the same order of magnitude as the optimal order of magnitude for the Parzen-Rosenblatt kernel estimators (which require the observations themselves, rather than a histogram of them); the proof is a substantial improvement over corresponding proofs for the Good and Gaskins version of the MPE, whose consistency results have only been shown for $K = cN^q$, $0 < q < 1/4$ (in our case, $q = 1$).

The approximate form in (2.20) also provides a way of establishing the asymptotic distribution of $\hat{p}_i$. Let $z_i = (\hat{p}_i/\pi_i) - 1$. Then the results of Morris (1975, Corollary 4.1) imply that z_i is asymptotically normally distributed, specifically, $\{z_i - \mu(z_i)\}/\sigma(z_i) \to_{\mathscr{D}} N(0, 1)$, where

$$\mu(z_i) = 2\beta \textstyle\sum_{j=1}^K h_{ij}\log(\pi_{j-1}\pi_{j+1}/\pi_j^2) \tag{2.21}$$

$$\sigma(z_i) = \{\textstyle\sum_{j=1}^K h_{ij}{}^2 N\pi_j(1 - \pi_j)\}^{1/2}. \tag{2.22}$$

3. Applications.

In this section we briefly address some of the issues involved in the implementation of the MPE procedure. The results here are basically preliminary; further investigation is continuing.

3.1 *The algorithm.* The MPE is defined as the value $\hat{p}$ that maximizes $L(\mathbf{p} \mid \mathbf{n})$ subject to $\sum_1^K p_i = 1$. The most efficient algorithm to get $\hat{\mathbf{p}}$ is the Newton-Raphson procedure; however, this requires calculation of the matrix H of Lemma 2.5 at each iteration (with the current probability estimates replacing π). This is not computationally desirable, since it requires $O(K^3)$ operations at each step. An approximate maximization procedure that requires far fewer operations is to use the inverse H (at all steps) with $\pi_1 = 1/\{2(K - 1)\}$; $\pi_i = 1/(K - 1)$, $i = 2, \cdots, K - 1$; $\pi_K = 1/\{2(K - 1)\}$; this can be calculated in closed form (we denote it by Q). This requires only $O(K^2)$ operations, and extensive computations have indicated that Q and H are close enough so that the convergence properties are similar.

3.2 *Determining* β. Although Theorem 2.2 indicates that taking β of the order $K^{8/5}(\log K)^{2/5}$ (or $NK^{3/5}(\log K)^{2/5}$) will give the convergence (2.20), this does not give any indication as to the proper choice of β for a particular problem. An algorithm can be formulated that chooses β based on the data itself through (2.20) (this follows a suggestion by James R. Thompson). Specifically, the mean squared error of the estimator $\sum_1^K(\hat{p}_i - \pi_i)^2$ is asymptotically a quadratic function of β:

$$\text{(3.1)} \qquad \text{MSE} = \textstyle\sum_{i=1}^K [\pi_i \sum_{j=1}^K h_{ij}\{(n_j - N\pi_j) + 2\beta \log(\pi_{j-1}\pi_{j+1}/\pi_j^2)\}]^2.$$

Using (3.1), the value of β that minimizes the MSE can be determined easily as a function of π. This can be used algorithmically by substituting the current estimate $\hat{\mathbf{p}}$ for the value of π and updating the value of β. Combining this with the modified Newton-Raphson procedure outlined above gives a computational procedure that provides a data-dependent technique to determine β and $\hat{\mathbf{p}}$. Although convergence of this procedure is not guaranteed, extensive computations have indicated that convergence is not a problem. It can be noted that Laird (1978) provides consistent estimation of prior parameters through the use of the EM algorithm of Dempster, Laird and Rubin (1977).

3.3 *Effectiveness of* MPE. Theorem 2.2 provides the theoretical justification for the MPE, but cannot be used to determine what value of K is large enough for the MPE to be effective. In this section we examine that question through the use of simulations. The MLE and MPE are compared with the Fienberg and Holland (1973) estimator (which is the minimum MSE estimator of the form (1.1) with $\alpha = \alpha\mathbf{1}$). In this limited study, N and K are set equal to each other for varying values ranging from 15 to 100, and the MSE of each estimator is calculated based on 200 simulations. The true probability vector π is uniform ($\pi_i = 1/K\ \forall i$).

The results are given below.

K = N	MLE	F & H	MPE
15	.05581	.03135	.01104
20	.04640	.02545	.00774
50	.01926	.00996	.00164
100	.00976	.00492	.00048

As can be seen, the F & H estimator provides the expected improvement over the MLE (approximately a 50% improvement for the range of $K = N$ given). However, for K as small as 15, the MPE is significantly better (more than an 80% reduction in the MSE over the MLE). This superiority becomes more pronounced as K increases; for $K = N = 100$, the reduction in MSE of the MPE over the MLE exceeds 95%.

3.4 *Pre-testing for smoothness.* The consistency result of Theorem 2.2 requires the smoothness constraint

$$\sup_i |\log(\pi_{i-1}\pi_{i+1}/\pi_i^2)| = O(K^{-2}).$$

As was pointed out earlier, this is implied by the usual smoothing assumptions of a bounded second derivative of the density being estimated. If the probability vector π is not smooth, there is no intuitive basis for the smoothing procedure proposed here.

The estimates in $\hat{\mathbf{p}}$ can themselves be used as a screening device to test the smoothness assumption. If π is not smooth, adjacent values in $\hat{\mathbf{p}}$ should be farther apart than if π is

TABLE 1
Percentage of silica in 22 chondrites

Percentage	Frequency	MLE	MPE
20.00-21.6	1	.0455	.0664
21.61-23.2	3	.1364	.0796
23.21-24.8	0	.0000	.0561
24.81-26.4	0	.0000	.0665
26.41-28.0	6	.2727	.1508
28.01-29.6	2	.0909	.1041
29.61-31.2	1	.0455	.0848
31.21-32.8	1	.0455	.0969
32.81-34.4	7	.3182	.1894
34.41-36.0	1	.0455	.1055

smooth. This suggests the statistic

$$S_K = \sum_{i=1}^{K} \{\log(\hat{p}_i/\hat{p}_{i+1})\}^2$$

to test smoothness. Note that this is proportional to the penalty function $\Phi(\mathbf{p})$. Large values of S_K are indicative of non-smoothness.

In order to use S_K as a testing tool, its null distribution must be determined. This can be accomplished through the use of simulations. Preliminary examination suggests that the power of the S_K test to detect a highly non-smooth π for $K = N = 100$ is quite high. More work is necessary to evaluate the test completely.

3.5 *Numerical examples.* In this section, two numerical examples are given to illustrate the effectiveness of the algorithm proposed here, and to compare it with other techniques.

The first data set is a small sparse table ($K = 10$) giving the percentage of silica in 22 chondrite meteors, originally from Ahrens (1965), and analyzed by Good and Gaskins (1980) and Leonard (1978). The data were originally in continuous form; they have been converted to a table for analysis.

Table 1 gives the original chondrite data (in a discretized form), the frequency estimates and the MPE probability estimates found using the algorithm of Sections 3.1 and 3.2.

It can be seen that the density appears to be trimodal with modes in the cells corresponding to (21.6, 23.2), (26.4, 28.0) and (32.8, 34.4); each succeeding mode is larger than the previous one. This is the conclusion reached by Good and Gaskins, and contrasts with Leonard's result of a bimodal distribution.

The other data set analyzed here is considerably larger ($K = 55$) and comes from Maguire, Pearson and Wynn (1952). It gives the time intervals between explosions in mines involving more than 10 men killed in Great Britain from December 8, 1875 to May 29, 1951. This data set was also analyzed by Leonard (1978).

Maguire *et. al.* fitted an exponential curve to this data, but as Leonard points out this is not necessarily the best choice (as can be seen in the estimates in Table 2). The true distribution has more of a peak than the exponential, and a thicker tail. Additionally, there is a bulge around the 250 day cell and a slight rise at the upper extreme. These are similar to the conclusions Leonard reached.

4. Higher dimensional tables. Consider an $R \times C$ two dimensional table $\{n_{ij}\}$. The likelihood function is multinomial:

$$\ell(\mathbf{n} \mid \mathbf{p}) \propto \prod_{i=1}^{R}\prod_{j=1}^{C} p_{ij}^{n_{ij}}, \quad \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} = 1. \tag{4.1}$$

The penalty function proposed in this case generalizes the use of the ratios p_i/p_{i+1} to the use of cross-product ratios, so that the function to be maximized to get the estimates is

TABLE 2
Time intervals between explosions in mines

Days	Frequency	MLE	MPE	Days	Frequency	MLE	MPE
0–30	18	.1651	.1144	841–870	0	.0000	.0040
31–60	14	.1284	.1028	871–900	1	.0092	.0038
61–90	9	.0826	.0874	901–930	0	.0000	.0036
91–120	8	.0734	.0751	931–960	0	.0000	.0034
121–150	6	.0550	.0647	961–990	0	.0000	.0033
151–180	4	.0367	.0569	991–1020	0	.0000	.0032
181–210	6	.0550	.0522	1021–1050	0	.0000	.0031
211–240	7	.0642	.0476	1051–1080	0	.0000	.0030
241–270	1	.0092	.0419	1081–1110	0	.0000	.00293
271–300	6	.0550	.0396	1111–1140	0	.0000	.00290
301–330	7	.0642	.0362	1141–1170	0	.0000	.00289
331–360	5	.0459	.0312	1171–1200	0	.0000	.00290
361–390	5	.0459	.0260	1201–1230	1	.0092	.00292
391–420	0	.0000	.0208	1231–1260	0	.0000	.00290
421–450	0	.0000	.0174	1261–1290	0	.0000	.00291
451–480	2	.0183	.0151	1291–1320	1	.0092	.00293
481–510	1	.0092	.0130	1321–1350	0	.0000	.00291
511–540	1	.0092	.0113	1351–1380	1	.0092	.00291
541–570	1	.0092	.0099	1381–1410	0	.0000	.00286
571–600	0	.0000	.0086	1411–1440	0	.0000	.00284
601–630	0	.0000	.0077	1441–1470	0	.0000	.00283
631–660	1	.0092	.0070	1471–1500	0	.0000	.00284
661–690	0	.0000	.0063	1501–1530	0	.0000	.00287
691–720	0	.0000	.0057	1531–1560	0	.0000	.00291
721–750	1	.0092	.0053	1561–1590	0	.0000	.00297
751–780	0	.0000	.0049	1591–1620	1	.0092	.00305
781–810	0	.0000	.0045	1621–1650	1	.0092	.00309
811–840	0	.0000	.0042				

$$
(4.2)\quad \begin{aligned} L(\mathbf{p}\,|\,\mathbf{n}) &= \textstyle\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\log p_{ij} \\ &\quad -\beta \textstyle\sum_{i=1}^{R-1}\sum_{j=1}^{C-1}[\log\{(p_{ij}p_{i+1,j+1})/(p_{i+1,j}p_{i,j+1})\}]^2,\ \sum_{i=1}^{R}\sum_{j=1}^{C}p_{ij}=1. \end{aligned}
$$

The effect of this penalty function is to smooth the estimates by pushing the inherent 2 × 2 tables of adjacent rows and columns towards independence (since the logs of the cross-product ratios are forced towards zero). If the underlying probability matrix π satisfies smoothness criteria, it is easy to show that

$$
(4.3)\qquad \log\{(\pi_{ij}\pi_{i+1,j+1})/(\pi_{i+1,j}\pi_{i,j+1})\} = O(R^{-1}C^{-1});
$$

so, as $R \to \infty$, $C \to \infty$, the logs of the true cross-product ratios approach zero. This means that in a large table, smoothness corresponds to local independence, so the penalty function proposed is a sensible one. Although a proof of sparse asymptotic consistency analogous to the proof in Section 2.2 has not been found, computer simulations indicate the estimator has similar consistency properties (Simonoff, 1980).

In order to generalize this to higher dimensions, an extension of the concept of cross-product ratios to multidimensional tables is necessary; the log-linear model provides this in the form of the highest interaction term. For instance, in the two-dimensional case, the saturated log-linear model has the form

$$
(4.4)\qquad \log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}.
$$

We use estimates that are moved toward

$$
(4.5)\qquad \log p_{ij} = u + u_{1(i)} + u_{2(j)},
$$

(the model of independence). This is done by making the penalty function of the form

$$-\beta \textstyle\sum_i \sum_j (u_{12(ij)})^2 = -\beta \sum_{i=1}^{R-1} \sum_{j=1}^{C-1} [\log\{(p_{ij}p_{i+1,j+1})/(p_{i+1,j}p_{i,j+1})\}]^2. \tag{4.6}$$

This clearly generalizes to multidimensional tables. For instance, the penalty function for a three-dimensional table would be

$$\begin{aligned} -\beta \textstyle\sum_i \sum_j \sum_m (u_{123(ijm)})^2 \\ = -\beta \textstyle\sum_{i=1}^{K_1-1} \sum_{j=1}^{K_2-1} \sum_{m=1}^{K_3-1} \left(\log \frac{p_{ijm} p_{i,j+1,m+1} p_{i+1,j,m+1} p_{i+1,j+1,m}}{p_{i,j,m+1} p_{i,j+1,m} p_{i+1,j,m} p_{i+1,j+1,m+1}} \right). \end{aligned} \tag{4.7}$$

5. Conclusions. This paper provides a technique for consistent estimation of all probabilities in a large sparse multinomial (one-dimensional contingency table). Further investigation of the applications-oriented problems of Section 3 is necessary, as is study of the behavior of higher dimensional analogues.

Acknowledgment. The author would like to thank Professor J. A. Hartigan of Yale University for his assistance in the preparation of this work which, in its original form, was part of the author's Ph.D. thesis. The suggestions of the referees and the associate editor also helped improve the presentation greatly.

REFERENCES

AHRENS, L. H. (1965). Observations on the Fe-Si-Mg relationship in chondrites. *Geochimica et Cosmochimica Acta* **29** 801-806.

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** 33-45.

DE MONTRICHER, G., TAPIA, R. and THOMPSON, J. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Ann. Statist.* **3** 1329-1348.

DEMPSTER, A. P., LAIRD, N., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1-38 (with discussion).

FIENBERG, S. and HOLLAND, P. (1973). Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68** 683-691.

GOOD, I. J. and GASKINS, R. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255-277.

GOOD, I. J. and GASKINS, R. (1980). Density estimation and bump-hunting by penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42-74 (with discussion).

LAIRD, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65** 581-590.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60** 297-308.

LEONARD, T. (1975). Bayesian estimation models for two-way tables. *J. Roy. Statist. Soc. Ser. B* **37** 23-37.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B* **40** 113-146 (with discussion).

MAGUIRE, B. A., PEARSON, E. S., and WYNN, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika* **39** 168-180.

MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165-188.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065-1076.

ROSENBLATT, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* **27** 832-835.

SIMONOFF, J. S. (1980). A penalty function approach to smoothing large sparse contingency tables. Ph.D. Thesis, Yale University.

TAPIA, R. and THOMPSON, J. (1978). *Nonparametric probability density estimation.* Johns Hopkins Press, Baltimore.

NEW YORK UNIVERSITY
GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
QUANTITATIVE ANALYSIS AREA
100 TRINITY PLACE
NEW YORK, NEW YORK 10006