RIDGE TYPE ESTIMATORS OF MULTINOMIAL CELL PROBABILITIES[1]

Ayodele Ighodaro and Thomas Santner

School of Operations Research and Industrial Engineering
Cornell University
Ithaca, New York, U.S.A.

I. INTRODUCTION AND SUMMARY

Let $\underset{\sim}{W} = (W_1,\ldots,W_t)$ have a multinomial distribution based on N(known) trials and having unknown vector of cell probabilities $\underset{\sim}{p} \in S \equiv \{\underset{\sim}{q} \in R^t: \ q_i \geq 0 \ \forall i \text{ and } \sum q_i = 1\}$. When the range of a summation or product is from 1 to t it will be suppressed for ease of notation. This paper develops analogues of normal theory ridge regression estimators for the problem of simultaneously estimating $\underset{\sim}{p}$. Sections 1 to 4 will study the case of squared error loss (SEL), $L_S(\underset{\sim}{p},\underset{\sim}{a}) \equiv N\sum (p_i-a_i)^2 = N||\underset{\sim}{p} - \underset{\sim}{a}||^2$, because of its wide use (see Bishop, Fienberg, and Holland [5], Chapter 12 and the references therein), but other loss functions will be mentioned in Section 5 because of their ability to penalize zero guesses of positive p_i.

The maximum likelihood estimator (mle) $\hat{\underset{\sim}{p}} \equiv \underset{\sim}{W}/N$ is known to be admissible under SEL for all t (Johnson [21]; Alam [3]; Brown [6]); hence there is no Stein-effect for this problem. Furthermore, there is a unique admissible minimax estimator (Steinhaus [25]; Trybula [29]). In the absence of any other information about $\underset{\sim}{p}$ one of these estimators might be satisfactory. However, $\hat{\underset{\sim}{p}}$ has risk $R_S(\underset{\sim}{p},\hat{\underset{\sim}{p}}) \equiv E_{\underset{\sim}{p}}[L_S(\underset{\sim}{p},\hat{\underset{\sim}{p}})] = 1-||\underset{\sim}{p}||^2$ which is near zero only when $\underset{\sim}{p}$ is near a vertex. This led Good [13], [14],

[1]This research was supported by the National Science Foundation under Grant No. ENG-7906914.

Sutherland, Fienberg and Holland [27], Albert [4] and others to consider Bayesian motivated estimators of $\underset{\sim}{p}$ which dominate $\hat{\underset{\sim}{p}}$ over a "large" portion of S. In particular, it is well known that the unique Bayes estimator of $\underset{\sim}{p}$ versus the conjugate Dirichlet prior with mean $E[\underset{\sim}{p}] = \underset{\sim}{\lambda}$ and variance-covariance matrix $E[(\underset{\sim}{p}-\underset{\sim}{\lambda})(\underset{\sim}{p}-\underset{\sim}{\lambda})'] = (D(\underset{\sim}{\lambda})-\underset{\sim}{\lambda}\underset{\sim}{\lambda}')/(K+1)$ is

$$(1.1) \qquad \hat{\underset{\sim}{p}}_K = (\underset{\sim}{W}+K\underset{\sim}{\lambda})/(N+K) = \omega\,\hat{\underset{\sim}{p}} + (1-\omega)\underset{\sim}{\lambda}$$

where $K > 0$ and $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_t) \in S$ are known, $D(\underset{\sim}{\lambda})$ is diagonal with the elements of $\underset{\sim}{\lambda}$ and $\omega = N/(N+K)$. Here and throughout the paper vectors are column vectors and prime denotes transpose. Formally $K = 0$ make sense in (1.1) and yields the mle $\hat{\underset{\sim}{p}}_0 \equiv \hat{\underset{\sim}{p}}$ of $\underset{\sim}{p}$.

Section 2 shows that the class of estimators $\{\hat{\underset{\sim}{p}}_K\colon\ K \geq 0\}$ is the analogue of the class of ridge estimators in a number of frequentist senses as well as in the well-known Bayesian sense. Section 3 uses the properties of $\hat{\underset{\sim}{p}}_K$ developed in Section 2 to illustrate the construction of two ridge estimators; it is possible to construct other ridge analogues. Small sample simulation studies are presented in Section 4 which compare the current estimators with those previously proposed; it concentrates on the large sparse multinomial framework introduced by Fienberg and Holland [11]. Section 5 critiques two other loss functions; it summarizes some possible approaches for estimating $\underset{\sim}{p}$ under various model assumptions in these cases.

II. RIDGE ESTIMATION

It is well known that the class of ridge estimators can be developed from a Bayesian viewpoint by postulating $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ for an $n\times 1$ vector of data $\underset{\sim}{Y}$ where $\underset{\sim}{X}$ is a known $n\times p$ matrix of rank p, $\underset{\sim}{\beta}$ is a $p\times 1$ vector of unknown parameters and $\underset{\sim}{\varepsilon}$ is an $n\times 1$ vector of

experimental errors satisfying $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 I_n)$ and by assuming $\underset{\sim}{\beta} \sim N_p(\underset{\sim}{0}, \frac{\sigma^2}{K} I_p)$. The Bayes estimator of $\underset{\sim}{\beta}$ with respect to squared error loss is well known to be

$$(2.1) \qquad \hat{\underset{\sim}{\beta}}_K = (\underset{\sim}{X}'\underset{\sim}{X} + KI_p)^{-1}\underset{\sim}{X}'\underset{\sim}{Y}.$$

Formally $\hat{\underset{\sim}{\beta}}_0 = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ also makes sense in (2.1); $\hat{\underset{\sim}{\beta}}_0$ is, of course, the mle and BLUE of $\underset{\sim}{\beta}$. $\hat{\underset{\sim}{\beta}}_0$ is inadmissible under squared error loss when $p \geq 3$; its (summed) mean squared length satisfies

$$(2.2) \qquad E_{\underset{\sim}{\beta},\sigma^2}[||\hat{\underset{\sim}{\beta}}_0||^2] - ||\underset{\sim}{\beta}||^2 = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$$

where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues of $\underset{\sim}{X}'\underset{\sim}{X}$. For any $K > 0$, $\hat{\underset{\sim}{\beta}}_K$ is biased; however, its length, $||\hat{\underset{\sim}{\beta}}_K||^2$, is shorter than that of $\hat{\underset{\sim}{\beta}}_0$ (see P1 below).

Hoerl and Kennard [18] developed the following properties and characterizations of $\hat{\underset{\sim}{\beta}}_K$.

P1: The distance $||\hat{\underset{\sim}{\beta}}_K|| = (\hat{\underset{\sim}{\beta}}_K' \hat{\underset{\sim}{\beta}}_K)^{1/2}$ of $\hat{\underset{\sim}{\beta}}_K$ to (the prior mean) $\underset{\sim}{0}$ (= $E[\underset{\sim}{\beta}]$) is a continuous monotone decreasing function of K such that $||\hat{\underset{\sim}{\beta}}_K|| \downarrow 0$ as $K \uparrow \infty$.

P2: $\hat{\underset{\sim}{\beta}}_K$ is a restricted maximum likelihood (least squares) estimator of $\underset{\sim}{\beta}$. Denote the residual sum of squares of $\underset{\sim}{\beta}$ by $\psi(\underset{\sim}{\beta}) = ||\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}||^2$; for fixed $K > 0$, $\hat{\underset{\sim}{\beta}}_K$ minimizes $\psi(\underset{\sim}{\beta})$ among $\underset{\sim}{\beta}$ in the sphere $B_K \equiv \{\underset{\sim}{\beta} \in R^p: ||\underset{\sim}{\beta}|| \leq ||\hat{\underset{\sim}{\beta}}_K||\}$. Hence $\psi(\hat{\underset{\sim}{\beta}}_K)$ is increasing in K by P1.

P3: If $||\underset{\sim}{\beta}||^2$ is bounded then there exists a $K_o > 0$ such that $E_{\underset{\sim}{\beta}}[||\hat{\underset{\sim}{\beta}}_K - \underset{\sim}{\beta}||^2] \leq E_{\underset{\sim}{\beta}}[||\hat{\underset{\sim}{\beta}}_0 - \underset{\sim}{\beta}||]$ for $0 < K < K_o$.

P2 characterizes $\hat{\underset{\sim}{\beta}}_K$ while P3 suggests the possibility of constructing minimax ridge estimators by using adaptive (stochastic)

is concave in δ_i and is therefore maximized when

$$(2.5) \qquad \frac{1}{K}\frac{W_i}{\delta_i} + \frac{\lambda_i}{\delta_i} = \frac{N+K}{K} \ (1 \leq i \leq t) \Leftrightarrow \delta_i = \hat{p}_{iK} \quad (1 \leq i \leq t).$$

Substituting (2.5) into the right hand side of (2.4) gives

$$\frac{1}{K}\sum W_i \ell n \delta_i \leq \frac{1}{K}\sum W_i \ell n \hat{p}_{iK}$$

for all $\underset{\sim}{\delta} \in B_K^e$ and completes the proof.

The analogue of a more severe form of P3 is immediate.

THEOREM 2.3. *For any* $\underset{\sim}{p} \in S$ *with at least two non-zero components, the mean square entropy distance of* $\hat{\underset{\sim}{p}}_K$ *to* $\underset{\sim}{p}$ *for any* $K > 0$ *is less than that of the* mle $\hat{\underset{\sim}{p}}$.

Remark 2.1. Unlike the regression analog P3, there is no upper bound on the value of K which "improves" $\hat{\underset{\sim}{p}}_K$ over $\hat{\underset{\sim}{p}}$.

Remark 2.2 If $\underset{\sim}{\lambda}$ has one or more zero components then Theorems 2.1 to 2.3 still hold though their proofs must be modified.

Remark 2.3. Marquardt [22] proves that $\hat{\underset{\sim}{\beta}}_K$ is the maximum likelihood estimator of $\underset{\sim}{\beta}$ when the data $\underset{\sim}{Y}$ are supplemented by fictitious data $\underset{\sim}{Y}^* \equiv 0$ drawn from an experiment orthogonal to the original design. An analogous statement is true for $\hat{\underset{\sim}{p}}_K$; $\hat{\underset{\sim}{p}}_K$ is the maximum likelihood estimator of $\underset{\sim}{p}$ when $\underset{\sim}{W}$ is supplemented by $\underset{\sim}{W}^* = (K\lambda_1, \ldots, K\lambda_t)$.

III. RIDGE ESTIMATORS OF $\underset{\sim}{p}$

The admissibility of $\hat{\underset{\sim}{p}}$ insures that unlike the regression case it is impossible to find an estimator $\hat{K}$ of K so that $\hat{\underset{\sim}{p}}_{\hat{K}}$ dominates $\hat{\underset{\sim}{p}}$. Yet it is still reasonable to study analogues of ridge estimators in the hope they have reasonable mean squared

error properties relative to $\hat{\underset{\sim}{p}}$ for $\underset{\sim}{p}$ "away" from the vertices. So now assume $K > 0$ but unknown while λ is known. A differentiation shows (see [5], page 407-8) that the K which minimizes $R_S(\underset{\sim}{p},\hat{\underset{\sim}{p}}_K)$ is

$$(3.1) \qquad K^* = K^*(\underset{\sim}{p},\underset{\sim}{\lambda}) \equiv (1-||\underset{\sim}{p}||^2)/||\underset{\sim}{p}-\underset{\sim}{\lambda}||^2.$$

We consider two iterative estimators of K^* or equivalently of $\omega^* = N/(N+K^*)$. The idea is to use the current estimate of ω^* to obtain an improved estimate of $\underset{\sim}{p}$ which in turn can be used to re-estimate ω^*.

Define the sequence $\{\omega_i\}_{i=1}^{\infty}$ by the algorithm:

1. Make an initial estimate ω_0 of ω^*, set i=1 and go to Step 2.

2. Set $\underset{\sim}{\delta}_i = \omega_{i-1}\hat{\underset{\sim}{p}} + (1-\omega_{i-1})\underset{\sim}{\lambda}$ and go to Step 3.

3. Set $\omega_i = \dfrac{N||\underset{\sim}{\delta}_i-\underset{\sim}{\lambda}||^2}{N||\underset{\sim}{\delta}_i-\underset{\sim}{\lambda}||^2 + 1-||\underset{\sim}{\delta}_i||^2}$, increment i by 1, and go to Step 2.

Set $\omega_{ST} = \lim_{i\to\infty} \omega_i$ and $\hat{\underset{\sim}{p}}_{ST} = \omega^*_{ST}\,\hat{\underset{\sim}{p}} + (1-\omega^*_{ST})\underset{\sim}{\lambda}$ when the limit exists.

When $\underset{\sim}{W} = N\underset{\sim}{\lambda}$ then it is easy to check that $\omega^*_{ST} = 0$ and $\hat{\underset{\sim}{p}}_{ST} = \underset{\sim}{\lambda}$. When $\underset{\sim}{W} = Ne_i$ where $e_i = (0,\ldots,0,1,0,\ldots,0)$ is the i^{th} unit vector then it is also obvious that $\omega^*_{ST} = 1$ and $\hat{\underset{\sim}{p}}_{ST} = e_i$ for $1 \le i \le t$. The value of ω^*_{ST} for arbitrary $\underset{\sim}{W}$ is given in Theorem 3.1. First an elementary lemma will be stated.

LEMMA 3.1. *Suppose* $\underset{\sim}{W} \neq \underset{\sim}{\lambda}$ *and* α *and* ν *are defined as* $\alpha \equiv \underset{\sim}{\lambda}'(\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda})/N||\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda}||^2$ *and* $\nu \equiv (1-\underset{\sim}{\lambda}'\underset{\sim}{\lambda})/N||\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda}||^2$, *respectively then* $\nu \ge 2\alpha$.

Proof. $\nu \ge 2a \Leftrightarrow 1 + \underset{\sim}{\lambda}'\underset{\sim}{\lambda} - 2\underset{\sim}{\lambda}'\underset{\sim}{p} \ge 0$

$\Leftrightarrow (\underset{\sim}{\lambda}-\hat{\underset{\sim}{p}})'(\underset{\sim}{\lambda}-\hat{\underset{\sim}{p}}) + (1-\hat{\underset{\sim}{p}}'\hat{\underset{\sim}{p}}) \ge 0$

which is obvious.

THEOREM 3.1. *Define* $\gamma = (N-1)/N$ *and* α, ν *as in Lemma 3.1. When the iterative estimator of the optimal constant* ω^* *begins at Step 1 with* $\omega_0 = 1$ *then the sequence* $\{\omega_i\}_{i=1}^{\infty}$ *converges to the limit*

$$\omega_{ST}^* = \begin{cases} 0 & \text{when } \underline{\text{either}}\ (2\alpha+1)^2 < 4\gamma\nu \text{ or} \\ & \quad (2\alpha+1) < 0 \quad \text{or} \\ & \quad (2\gamma < (2\alpha+1) \text{ and } \underset{\sim}{p} \notin \{e_1,\ldots,e_t\}) \\ \dfrac{(2\alpha+1) + [(2\alpha+1)^2 - 4\gamma\nu]^{1/2}}{2\gamma} & \text{when } ((2\alpha+1)^2 \geq 4\gamma\nu \\ & \quad \text{and } 2\gamma \geq (2\alpha+1) \geq 0) \\ 1 & \text{otherwise.} \end{cases}$$

Proof. After substitution and some algebra it can be shown that

$$\omega_i = \frac{N||\underset{\sim}{\delta}_i - \underset{\sim}{\lambda}||^2}{N||\underset{\sim}{\delta}_i - \underset{\sim}{\lambda}||^2 + 1 - ||\underset{\sim}{\delta}_i||^2} = f(\omega_{i-1}),$$

say, where $f(\omega) = \omega^2/(\gamma\omega^2 - 2\alpha\omega + \nu)$. Here $f\colon [0,1] \to [0,1]$; differentiation gives $f'(\omega) = 2\omega(\nu - \alpha\omega)/(\gamma\omega^2 - 2\alpha\omega + \nu)^2 > 0$ for $1 \geq \omega > 0$ by Lemma 3.1. Thus $\{\omega_i\}_{i=1}^{\infty}$ is a nonincreasing sequence bounded below and therefore must have a limit. Furthermore $\omega_{i+1} < \omega_i$ whenever $\omega_i > 0$. The limit points of $\{\omega_i\}_{i=1}^{\infty}$ must solve the fixed point equation

$$\begin{aligned} (3.2) \qquad & f(\omega) = \omega \\ & \Leftrightarrow \omega(\gamma\omega^2 - (2\alpha+1)\omega + \nu) = 0 \\ & \Leftrightarrow \omega \in \{\omega^1, \omega^2, \omega^3\}, \end{aligned}$$

$$\omega^1 \equiv 0,\ \omega^2 \equiv [(2\alpha+1) + [(2\alpha+1)^2 - 4\gamma\nu]^{1/2}]/2\gamma, \text{ and}$$

$$\omega^3 \quad [(2\alpha+1) - [(2\alpha+1)^2 - 4\gamma\nu]^{1/2}]/2\gamma.$$

If $(2\alpha+1)^2 < 4\gamma\nu$ then ω^1 is the only fixed point of $f(\omega)$ and

hence $\omega^*_{ST} = 0$. Suppose $(2\alpha+1)^2 \geq 4\gamma\nu$; it is easy to check that $g(\omega) = \gamma\omega^2 - (2\alpha+1)\omega + \nu$ has either both roots positive or both roots negative since $g(0) = \nu > 0$. Straightforward algebra shows that $\omega^2 \geq 0$, $\omega^3 \geq 0 \Leftrightarrow (2\alpha+1) \geq 0$. Hence if $(2\alpha+1) < 0$ then $\omega^2 \leq 0$ and $\omega^3 \leq 0 \Rightarrow \omega^1 = 0$ is again the only fixed point of $f(\omega)$ to which $\{\omega_i\}_{i=1}^{\infty}$ can converge since for every i, $\omega_i \in [0,1] \Rightarrow \omega_{i+1} = f(\omega_i) \geq f(0) = 0$. If $(2\alpha+1) \geq 0$ and $2\gamma - (2\alpha+1) \geq 0$ then straightforward but tedious algebra shows

$$\omega^2 \leq 1 \Leftrightarrow \hat{\underset{\sim}{p}}'\hat{\underset{\sim}{p}} \leq 1$$

which is always true. Hence $1 \geq \omega^2 \geq \omega^3 \geq \omega^1 = 0$ and since $\omega_0 = 1$ then for every i, $\omega_i \in [\omega^2, 1] \Rightarrow \omega_{i+1} = f(\omega_i) \geq f(\omega^2) = \omega^2$ and thus $\omega^*_{ST} = \omega^2$. The remaining subcases proceed along similar lines and the proof is completed.

Remark 3.1. $2\alpha+1 \geq 0 \Leftrightarrow \dfrac{\underset{\sim}{\lambda}'(\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda})}{N\|\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda}\|^2} \geq -\dfrac{1}{2}$

$$\Leftrightarrow (\hat{\underset{\sim}{p}}-\underset{\sim}{\lambda})'(\hat{\underset{\sim}{p}} - \frac{N-2}{N}\underset{\sim}{\lambda}) \geq 0$$

and hence for most $\underset{\sim}{W}$ is will be the case that $(2\alpha+1) \geq 0$.

Remark 3.2. The proof shows that while ω^*_{ST} always exists, its value can depend on the starting value ω_0. For example, if $(2\alpha+1) \geq 0$ and $2\gamma \geq (2\alpha+1) \geq 0$ then $1 \geq \omega^2 \geq \omega^3 \geq \omega^1 = 0$ hence if $\omega_0 \in [\omega^3, \omega^2)$ then $\omega^*_{ST} = \omega^3$ rather than ω^2. A similar phenomenon was noted in the ridge regression case by Hocking, Speed and Lynn [17].

It is instructive to consider the special case $\underset{\sim}{\lambda} = \underset{\sim}{c} \equiv (1/t,\ldots,1/t)$, it is easy to see $\alpha = 0$ and $\nu = (t-1)/\chi^2$ where $\chi^2 = \frac{t}{N}\sum_{i=1}^{t}(X_i - N/t)^2$. When $N > 2$ Theorem 3.1 simplifies to

$$\omega^*_{ST} = \begin{cases} 0, & \chi^2 < 4\,\frac{(N-1)}{N}\,(t-1) \\ \dfrac{N\left[1 + \left\{1-4\left(\frac{N-1}{N}\right)\frac{t-1}{\chi^2}\right\}^{1/2}\right]}{2(N-1)}, & \text{otherwise} \end{cases}$$

since $(2\alpha+1) = 1 > 0$ and $2\gamma - (2\alpha+1) = 2\gamma-1 = (N-2)/N > 0$; hence $\hat{p}_{ST} = 0\cdot\hat{p} + (1-0)\cdot c = c$ when $\chi^2 \leq 4(t-1)(N-1)/N$. We conclude that it requires an overwhelming amount of evidence against cell homogeneity to estimate the vector p by anything other than c. Furthermore when $\chi^2 \geq 4(t-1)(N-1)/N$ then $\omega^*_{ST} > N/2(N-1) > 1/2$ thus p_{ST} has a large discontinuity. Again a similar problem is encountered in the normal means case; see Hemmerle [16] and Hocking, Speed and Lynn [17].

The preceeding discussion suggests that ω^*_{ST} may iterate too far; one possible remedy is to use a finite step estimator of ω^*. If $\omega_0 = 1$ then the one step estimator of ω^* is ω_1 which is the maximum likelihood estimator of ω^*; denote the corresponding estimator of p by $\hat{p}_{ML}$. This estimator has been studied extensively beginning with Fienberg and Holland [10].

This paper will compare $\hat{p}_{ML}$ with the estimator which iterates one additional step; denote the two step estimator of ω^* by $\omega^*_{TS} \equiv \omega_2$ (initialized at $\omega_0 = 1$). After some algebra it can be shown that

$$\omega^*_{TS} = f(\omega_2) = f(f(1))$$

$$= [\gamma-2\alpha(\gamma-2\alpha+\nu) + \nu(\gamma-2\alpha+\nu)^2]^{-1}$$

where α, ν and γ are defined above. When $\lambda = c$ then ω^*_{TS} simplifies to

$$\omega^*_{TS} = \left[\frac{N-1}{N} + \frac{t-1}{\chi^2}\left(\frac{N-1}{N} + \frac{t-1}{\chi^2}\right)^2\right]^{-1} \text{ and}$$

$$\hat{p}_{TS} = \omega^*_{TS}(\hat{p}-c) + c.$$

Remark 3.3. In contrast, when $\underset{\sim}{\lambda} = \underset{\sim}{c}$ then the one step estimator is

$$\hat{\underset{\sim}{p}}_{ML} = [\frac{N-1}{N} + \frac{t-1}{\chi^2}]^{-1}(\hat{\underset{\sim}{p}}-\underset{\sim}{c}) + \underset{\sim}{c}.$$

When N is large $\hat{\underset{\sim}{p}}_{ML}$ is approximately the same as

$$\hat{\underset{\sim}{p}}_{G} = [1 + \frac{t-1}{\chi^2}]^{-1}(\hat{\underset{\sim}{p}}-\underset{\sim}{c}) + \underset{\sim}{c}.$$

Good [13] found $\hat{\underset{\sim}{p}}_{G}$ to be a reasonable approximation in several numerical examples to the so called "type II" maximum likelihood (or empirical Bayes) estimator of $\underset{\sim}{p}$.

An analogue of the McDonald and Galarneau [23] method for estimating K will now be developed. Let $\widehat{||\underset{\sim}{p}||}_e^2$ be an estimator of $||\underset{\sim}{p}||_e^2$; choose K so that

$$(3.3) \qquad ||\hat{\underset{\sim}{p}}_K||_e^2 = \widehat{||\underset{\sim}{p}||}_e^2.$$

Unlike the regression case, an unbiased estimate of $||\underset{\sim}{p}||_e^2$ does not exist; we make use of a concept due to Haldane [15].

Definition 3.1. The estimator $\hat{q}$ of the real function $q = q(\underset{\sim}{\theta})$ of the parameter $\underset{\sim}{\theta}$ is said to be *almost unbiased if*

$$E_{\theta}[\hat{q}] = q(\underset{\sim}{\theta}) + o(n^{-\ell})$$

for all $\underset{\sim}{\theta}$ and for all $\ell \in \{1,2,\ldots\}$.

Using a result of Cook, Kerridge, and Pryce [8] an almost unbiased estimator of $||\underset{\sim}{p}||_e^2$ is

$$(3.4) \qquad \widehat{||\underset{\sim}{p}||}_e^2 = D(N) + \sum \lambda_i(\ell n\lambda_i - D(W_i))$$

where $D(y) = \frac{d\ell n\Gamma(y+1)}{dy}$ is the digamma function. Define the fitted length estimator to be

$$\hat{\underset{\sim}{p}}_{FL} = \omega^*_{FL}\hat{\underset{\sim}{p}} + (1-\omega^*_{FL})\underset{\sim}{\lambda}$$

where $\omega^*_{FL} = N/(N+K^*_{FL})$ and

$$K^*_{FL} = \begin{cases} 0, & \|\hat{\underset{\sim}{p}}\|^2_e \geq \|\hat{\underset{\sim}{p}}\|^2_e \\ \infty, & \|\hat{\underset{\sim}{p}}\|^2_e \leq 0 \\ \text{solution of (3.3), otherwise.} \end{cases}$$

Remark 3.4. $\hat{\underset{\sim}{p}}_{FL} > 0$ for all $\underset{\sim}{W}$ since if $W_j = 0$ for some j then $\|\hat{\underset{\sim}{p}}\|^2_e = \infty > \|\hat{\underset{\sim}{p}}\|^2_e$ from (3.4).

Remark 3.5. When $\underset{\sim}{\lambda} = \underset{\sim}{c}$ then K^*_{FL} simplifies to

$$K^*_{FL} = \begin{cases} 0, & [\frac{1}{t}\sum D(W_i)]-D(N) \leq \frac{1}{t}\sum \ell n(W_i/N) \\ \infty, & [\frac{1}{t}\sum D(W_i)]-D(N) \geq \frac{1}{t}\sum \ell n(1/t) \\ \text{solution of (3.5), otherwise} \end{cases}$$

where

$$(3.5) \qquad [\frac{1}{t}\sum D(W_i)] - D(N) = \frac{1}{t}\sum \ell n(\hat{p}_{ik})$$

and $\hat{\underset{\sim}{p}}_K = (\hat{p}_{1K},\ldots,\hat{p}_{tK})$.

The next section compares the risk properties of the estimators proposed above with other estimators from the literature.

IV. SMALL SAMPLE SITUATIONS

First, some estimators proposed by Sutherland et al. [27] will be reviewed, their theoretical large sparse multinomial asymptotic behavior will be stated and finally the results of a simulation experiment to compare their small sample risks with those of Section 3 will be presented.

4.1 *Review of Previous Estimators*

Sutherland et al. proposed two methods of estimating the optimal constant K* in equation (3.1). The first method was mentioned in Section 3; it estimates K* by $K^*_{ML} = (N^2-\sum W_i^2)/(\sum(W_i-N\lambda_i)^2$, the mle of K* based on the conditional distribution of $\underset{\sim}{W}$ given $\underset{\sim}{p}$. The second ("ratio unbiased") method uses unbiased estimators of the numerator and denominator of K*,

$$K^*_{RU} = (N^2 - \sum W_i^2)/(\sum W_i^2 - 2(N-1) \sum W_i\lambda_i + N(N-1) \sum \lambda_i^2 - N).$$

Denote the estimators of $\underset{\sim}{p}$ corresponding to K^*_{ML} and K^*_{RU} by $\hat{\underset{\sim}{p}}_{ML}$ and $\hat{\underset{\sim}{p}}_{RU}$, respectively.

Sutherland [26] analyzes the asymptotic behavior of the risk of a large class of estimators, $C(\underset{\sim}{\lambda})$, as $N \to \infty$, $t \to \infty$ and $\rho \equiv N/t$ fixed (large, sparse tables); $C(\underset{\sim}{\lambda})$ includes $\hat{\underset{\sim}{p}}_{ML}$, $\hat{\underset{\sim}{p}}_{RU}$ and $\hat{\underset{\sim}{p}}$. Since the dimension of the true probability vector $\underset{\sim}{p}$ and of $\underset{\sim}{\lambda}$ increases to infinity he fixes two densities p(u) and λ(u) on [0,1] and defines

$$p_i = \frac{1}{t}\, p\left(\frac{i-.5}{t}\right), \quad i = 1,\ldots,t \text{ and}$$

$$\lambda_i = \frac{1}{t}\, \lambda\left(\frac{i-.5}{t}\right), \quad i = 1,\ldots,t.$$

for integer t. Sutherland proves that to o(1) the leading term of $R(\underset{\sim}{p},\hat{\underset{\sim}{p}}_{RU})$ is at least as small as the leading term of any $\underset{\sim}{p}^* \in C(\lambda)$; in particular it is smaller than the leading term of $R(\underset{\sim}{p},\hat{\underset{\sim}{p}}_{ML})$ which is smaller than the leading term of $R(\underset{\sim}{p},\hat{\underset{\sim}{p}})$. In addition he computes the asymptotic risk to $o(t^{-1})$ for arbitrary $\underset{\sim}{p}^* \in C(\underset{\sim}{\lambda})$. See also Bishop, Fienberg and Holland [5], Chapter 12.

4.2 *Small Sample Studies*

Our simulations attempt to answer three questions:

(1) How large must N and t be before the large sparse asymptotics are correct?

(2) Does the large sample improvement of $\hat{\underset{\sim}{p}}_{RU}$ over $\hat{\underset{\sim}{p}}_{ML}$ and/or $\hat{\underset{\sim}{p}}_{ML}$ over $\hat{\underset{\sim}{p}}$ carry over to small samples?

(3) How do the risk characteristics of the estimators of Sutherland et al. [27] compare with those of Section 3?

Remark 4.1. Bishop, Fienberg, and Holland [5] suggest improving $\hat{\underset{\sim}{p}}_{RU}$ by setting K^*_{RU} equal to zero whenever it would be negative; we denote the corresponding estimator of $\underset{\sim}{p}$ by $\hat{\underset{\sim}{p}}_{RUB}$. A modification that appears more reasonable is to set K^*_{RU} equal to $+\infty$ whenever it would be negative. To see this, note that K^*_{RU} is negative iff the unbiased estimate of $||\underset{\sim}{p}-\underset{\sim}{\lambda}||^2$ is negative. Thus setting the estimate of $||\underset{\sim}{p}-\underset{\sim}{\lambda}||^2$ equal to zero will result in our modification. Denote the corresponding estimator of $\underset{\sim}{p}$ by $\hat{\underset{\sim}{p}}_{RUI}$. Thus the simulation study examines the following estimators: (1) $\hat{\underset{\sim}{p}}$, (2) $\hat{\underset{\sim}{p}}_{RUB}$, (3) $\hat{\underset{\sim}{p}}_{RUI}$, (4) $\hat{\underset{\sim}{p}}_{ML}$, (5) $\hat{\underset{\sim}{p}}_{ST}$, (6) $\hat{\underset{\sim}{p}}_{TS}$, and (7) $\hat{\underset{\sim}{p}}_{FL}$.

Throughout this study $\lambda(u)$ was fixed to be a uniform distribution over [0,1], i.e. $\underset{\sim}{\lambda} = c = (t^{-1},\ldots,t^{-1})$ for each t. In this case all seven estimators have invariant risk with respect to permutations of $\underset{\sim}{p}$ and hence it suffices to study densities $p(u)$ which are monotone increasing. The beta densities

$$p(u|\eta) \equiv \begin{cases} \eta u^{\eta-1}, & 0 \le u \le 1 \\ 0, & \text{otherwise} \end{cases}$$

for $\eta = 1, 1.4, 3.5$ and 10 were selected to generate the $\underset{\sim}{p}$ sequences; for each (t,η), p_i was set equal to $p_i^*/\sum p_i^*$ where $p_i^* \equiv t^{-1}p((i-.5)t^{-1}|\eta)$ for $i = 1,\ldots,t$. The vector $\underset{\sim}{p}$ moves from the center of the simplex at $\eta = 1$ to a point near the vertex at $\eta = 10$. The simulation studied $t = 3, 5(5)20, 50, 100$ cells and $\rho = N/t = 1, 3, 5, 10$; one hundred replications were performed for most (η,t,ρ) combinations (not all combinations were used for $t = 50$ and 100 because of prohibitive computer expense). Among the checks made on the simulation were comparisons of the

estimated risks of $\hat{\underset{\sim}{p}}$ with their known values.

Extensive tables of the estimated risk values with their standard errors can be found in Ighodaro [19]. Some of the estimated risk values have been plotted in various ways together with the theoretical risk curves (to $o(t^{-1})$) for $\hat{\underset{\sim}{p}}_{RU}$ and $\hat{\underset{\sim}{p}}_{ML}$ in Figures 1 through 6. The theoretical risks were computed from the general expressions in Sutherland [26]. The standard errors of the estimated risk values ranged approximately from .003 to .08 although the majority were in the range .03 to .04.

1. When $(\eta,\rho) = (1,5)$ then $\underset{\sim}{p} = \underset{\sim}{c}$ and

$$R(\underset{\sim}{c},\hat{\underset{\sim}{p}}_{RU}) = 2/t + o(t^{-1}) \text{ and} \tag{4.1}$$

$$R(\underset{\sim}{c},\hat{\underset{\sim}{p}}_{ML}) = 2.5 + 0.175/t + o(t^{-1}). \tag{4.2}$$

2. When $(\eta,\rho) = (1.4,5)$ then

$$R(\hat{\underset{\sim}{p}},\hat{\underset{\sim}{p}}_{RU}) = .3076 + 1.4435/t + o(t^{-1}) \text{ and} \tag{4.3}$$

$$R(\underset{\sim}{p},\hat{\underset{\sim}{p}}_{ML}) = .4236 + .1408/t + o(t^{-1}). \tag{4.4}$$

3. When $(\eta,\rho) = (10,3)$ then

$$R(\underset{\sim}{p},\hat{\underset{\sim}{p}}_{RU}) = .9275 - 3.9517/t + o(t^{-1}) \text{ and} \tag{4.5}$$

$$R(\underset{\sim}{p},\hat{\underset{\sim}{p}}_{ML}) = .9278 - 4.0349/t + o(t^{-1}). \tag{4.6}$$

Figures 1, 3, and 5 plot equations (4.1) to (4.6) to $o(t^{-1})$ together with the simulated risk curves for $\hat{\underset{\sim}{p}}_{RUI}$ and $\hat{\underset{\sim}{p}}_{ML}$. Figure 1 also includes the simulated risk of $\hat{p}_{RUB}$; it is clear from this plot that $\hat{\underset{\sim}{p}}_{RUI}$ is the correct modification to $\hat{\underset{\sim}{p}}_{RU}$. Hence $\hat{\underset{\sim}{p}}_{RUB}$ is not considered in Figures 2-6. It can also be concluded that the asymptotic risk expressions to $o(t^{-1})$ are approximately correct for $t \geq 30$.

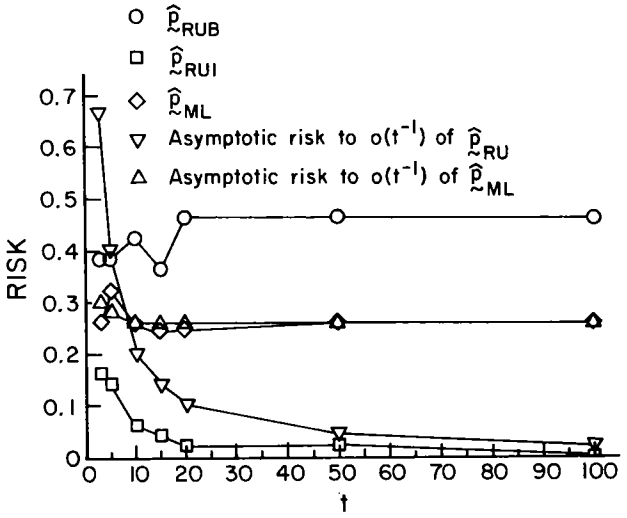

Figure 1. Simulated and asymptotic risks for $(\eta,\rho) = (1,5)$.

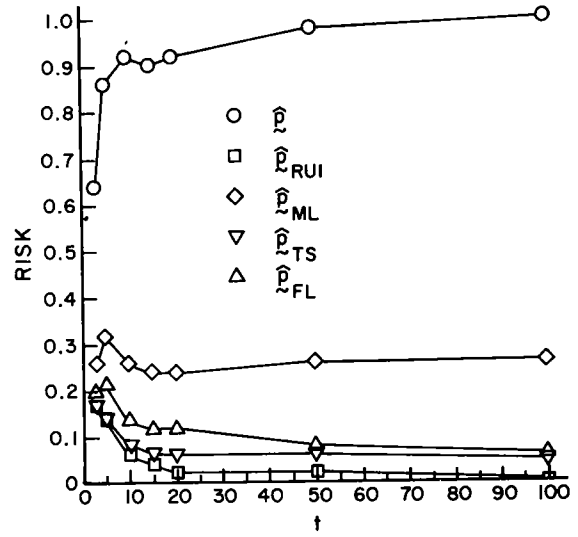

Figure 2. Simulated risk values for $(\eta,\rho) = (1,5)$.

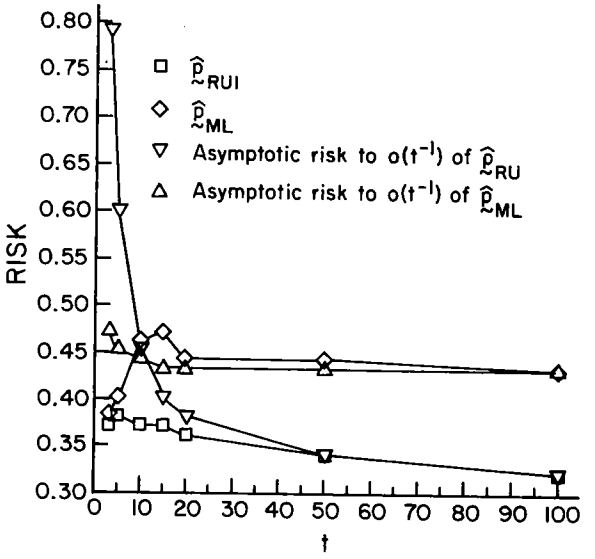

Figure 3. Simulated and asymptotic risks for $(\eta,\rho) = (1.4,5)$.

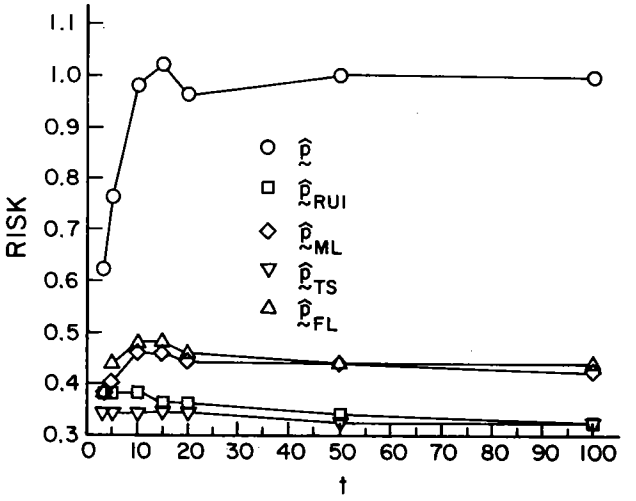

Figure 4. Simulated risk values for $(\eta,\rho) = (1.4,5)$.

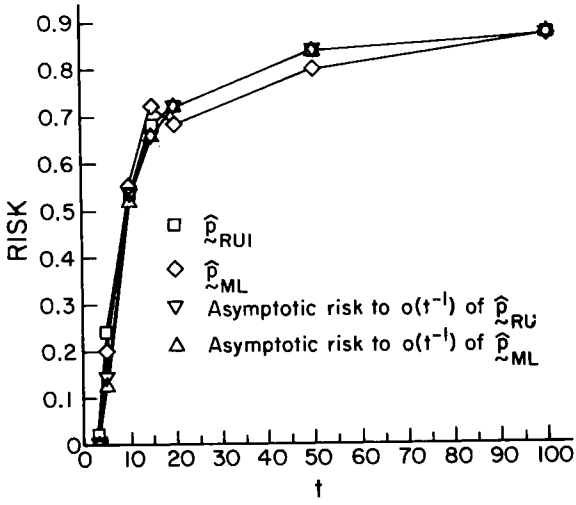

Figure 5. Simulated and asymptotic risks for $(\eta,\rho) = 10,3)$.

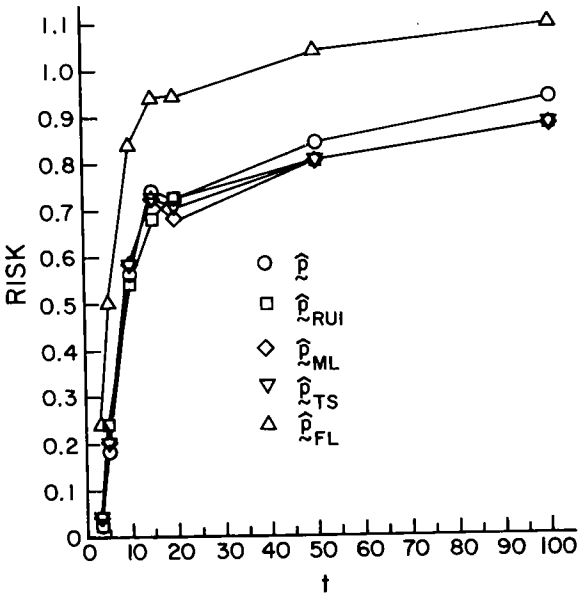

Figure 6. Simulated risk values for $(\eta,\rho) = (10,3)$.

Figures 2, 4 and 6 are plots of the simulated risk curves for $\hat{\underset{\sim}{p}}_{RUI}$, $\hat{\underset{\sim}{p}}_{ML}$, $\hat{\underset{\sim}{p}}$, $\hat{\underset{\sim}{p}}_{TS}$ and $\hat{\underset{\sim}{p}}_{FL}$. The curve for $\hat{\underset{\sim}{p}}_{ST}$ is not plotted because, as expected, it performs badly over a wide portion of the parameter space. For the remaining estimators we conclude that: (a) $\hat{\underset{\sim}{p}}_{RUI}$ and $\hat{\underset{\sim}{p}}_{TS}$ are comparable and better overall than any of the other estimators, (b) $\hat{\underset{\sim}{p}}_{FL}$ is superior to the mle $\hat{\underset{\sim}{p}}$ except for sparse data ($\rho = 3$) near vertices ($\eta = 10$), (c) all other estimators significantly outperform $\hat{\underset{\sim}{p}}$ except near the vertices ($\eta = 10$) where they are comparable to it.

V. DISCUSSION

Despite the favorable risk comparisons of most of the estimators in Section 4 relative to $\hat{\underset{\sim}{p}} = \underset{\sim}{W}/N$, it is unknown whether any of them are admissible for N and t fixed. Unique Bayes (hence admissible) estimators with data dependent K do arise from compound Dirichlet priors (Good [13], [14]), but they appear to be analytically intractible. More directly Brown's [6] characterization of the admissible rules might be useful for this problem.

SEL has been adopted throughout this paper because of its widespread use. However, under L_S the locus of equivalent actions $\underset{\sim}{a}$ versus a given $\underset{\sim}{p} > 0$ is spherical; hence L_S does not differentiate between positive and zero guesses of $p_i > 0$. One alternative, relative squared error loss (RSEL), $L_R(\underset{\sim}{p},a) \equiv \sum(p_i-a_i)^2 p_i$, deals with a related question. RSEL does not allow positive guesses of zero p_i however it still does allow zero guesses of positive p_i. Olkin and Sobel [24] apply the divergence theorem to show that the mle $\hat{\underset{\sim}{p}}$ is admissible unique minimax. Alternatively this can be proved by noting the unique Bayes estimator of $\underset{\sim}{p}$ relative to the Dirichlet prior

$$h(\underset{\sim}{p}) = \Gamma(K)\Pi \{p_i^{K\lambda_i-1}/\Gamma(K\lambda_i)\}, \quad \underset{\sim}{p} \in S$$

is for $1 \leq i \leq t$:

$$\hat{p}_i^B = \begin{cases} 0, & W_i = 0 \\ (W_i + K\lambda_i - 1)/(N+k-t), & W_i \geq 1; \end{cases}$$

in particular $\hat{\underset{\sim}{p}}$ is unique Bayes versus the prior with $\underset{\sim}{\lambda} = (1/t,\ldots,1/t)$ and $K = t$. Minimaxity follows from the fact that $\hat{\underset{\sim}{p}}$ has constant risk. Ighodaro [19] develops an asymptotic ($N \to \infty$, t fixed) application of the James-Stein estimator to the problem of estimating $\underset{\sim}{p}$ under the model assumption $\underset{\sim}{p} \in T \subset S$ where T is a smooth surface (e.g. $\ell n\ \underset{\sim}{p}$ satisfies a log linear model). The small sample risk performance of these estimators are unknown.

Entropy loss (EL), $L_E(\underset{\sim}{p},\underset{\sim}{a}) \equiv N \sum p_i \ell n(p_i/a_i)$ where $0\ \ell n\ 0 = 0$, is an alternative to SEL which effectively differentiates between positive and zero guesses of $p_i > 0$ (see Akaike [1], [2]); if $a_i = 0$ for any $p_i > 0$ then $L_E(\underset{\sim}{p},\underset{\sim}{a}) = +\infty$. The mle has entropy risk $R_E(\underset{\sim}{p},\hat{\underset{\sim}{p}}) = 0$ or $+\infty$ according as $\underset{\sim}{p}$ is a vertex or not. The mle is more dramatically unsatisfactory under L_E than under L_S; $\hat{\underset{\sim}{p}}$ is admissible when the problem has parameter space S but it is inadmissible when the problem has parameter space $S' \equiv \{\underset{\sim}{p} \in S:\ \underset{\sim}{p}$ not a vertex$\}$ (Ighodaro, Santner and Brown [20]). A theory of empirical or pseudo Bayes rules could be developed for L_E along the lines of Bishop et al. [5].

ACKNOWLEDGMENT

The authors would like to thank Professor M. Todd for suggesting a simplified proof of Theorem 2.1.

REFERENCES

[1] Akaike, H. (1971). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, B. Petrov and F. Csaki (eds.). Akademiai, Budapest.

[2] Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika 65*, 53-59.

[3] Alam, K. (1979). Estimation of multinomial probabilities. *Ann. Statist. 7*, 282-283.

[4] Albert, J. (1979). Robust Bayes estimation. Technical Report 79-9, Department of Statistics, Purdue University, W. Lafayette.

[5] Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge.

[6] Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *Ann. Statist. 9*, 1289-1300.

[7] Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist. 8*, 1036-1056.

[8] Cook, G., Kerridge, D., and Pryce, J. (1974). Estimation of functions of a binomial parameter. *Sankhyā Ser. A 36*, 443-448.

[9] Draper, N. and Van Nostrand, C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics 21*, 451-466.

[10] Fienberg, S., and Holland, P. (1970). Methods for eliminating zero counts in contingency tables. *Random Counts on Models and Structures*, G. P. Patel (ed.). Pennsylvania State Univ. Press, University Park.

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

[12] Goel, P. K. and Casella, G. (1976). A note on an explicit solution for generalized ridge regression. Technical Report No. 448, Department of Statistics, Purdue University, West Lafayette.

[13] Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge.

[14] Good, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. Ser. B 29*, 399-431.

[15] Haldane, J.B.S. (1957). Almost unbiased estimates of functions of sequences. *Sankhyā 17*, 201-8.

[16] Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. *Technometrics 17*, 309-314.

[17] Hocking, R., Speed, F., and Lynn, M. (1976). A class of biased estimators in linear regression. *Technometrics 18*, 425-437.

[18] Hoerl, A., and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*, 55-67.

[19] Ighodaro, A. O. (1980). Ridge and James-Stein methods for contingency tables. Ph.D. Thesis, School of OR/IE, Cornell University, Ithaca.

[20] Ighodaro, A., Santner, T., and Brown, L. (1980). Some admissibility and complete class results for the multinomial problem under entropy and squared error loss. To appear in *J. Mult. Anal.*

[21] Johnson, B. M. (1981). On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist. 42*, 1579-1587.

[22] Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics 12*, 591-612.

[23] McDonald, G., and Galarneau, D. (1975). A Monte-Carlo evaluation of some ridge-type estimators. *J. Amer. Statist. 70*, 407-416.

[24] Olkin, I. and Sobel, M. (1979). Admissible and minimax estimation for the multinomial distribution and for K independent binomial distributions. *Ann. Statist. 7*, 284-290.

[25] Steinhaus, H. (1957). The problem of estimation. *Ann. Math. Statist. 28*, 633-648.

[26] Sutherland, M. (1974). Estimation in large sparse multinomials. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge.

[27] Sutherland, M., Holland, P., and Fienberg, S. (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter. *Studies in Bayesian Econometrics and Statistics*, S. Fienberg and A. Zellner (eds.). North Holland, Amsterdam.

[28] Thisted, R. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Ph.D. Thesis, Technical Report No. 28, Division of Biostatistics, Stanford University, Stanford.

[29] Trybula, S. (1958). Some problems of simultaneous minimax estimation. *Ann. Math. Statist.* 39, 245-253.