The Annals of Statistics
1996, Vol. 24, No. 6, 2431–2461

USING SPECIALLY DESIGNED EXPONENTIAL FAMILIES FOR DENSITY ESTIMATION

BY BRADLEY EFRON[1] AND ROBERT TIBSHIRANI

Stanford University

We wish to estimate the probability density $g(y)$ that produced an observed random sample of vectors $y_1, y_2, \ldots, y_n$. Estimates of $g(y)$ are traditionally constructed in two quite different ways: by maximum likelihood fitting within some parametric family such as the normal or by nonparametric methods such as kernel density estimation. These two methods can be combined by putting an exponential family "through" a kernel estimator. These are the specially designed exponential families mentioned in the title. Poisson regression methods play a major role in calculations concerning such families.

1. Introduction. Suppose that we wish to estimate the probability density $g(y)$ that produced an observed random sample of vectors $y_1, y_2, \ldots, y_n$,

$$y_i \overset{\text{i.i.d.}}{\sim} g(y) \quad \text{for } i = 1, 2, \ldots, n. \tag{1.1}$$

The vectors y_i take values in a sample space $\mathcal{Y}$. The numerical examples in this paper have $\mathcal{Y}$ being portions of the real line or of the plane, but the methodology applies just as well to higher dimensionalities and to more complicated spaces.

Estimates of $g(y)$ are traditionally constructed in two quite different ways: by maximum likelihood fitting within some parametric family such as the normal or by nonparametric methods such as kernel density estimation. These two methods can be combined by putting an exponential family "through" a nonparametric estimator. The resulting hybrid estimators are the specially designed exponential families of the title.

Figure 1 shows a simple example of this methodology. The y_i are pain scores for $n = 67$ women, each obtained by averaging the results from a questionnaire administered after an operation. The scale runs from $y = 0 =$ no pain to $y = 4 =$ worst pain, so the sample space $\mathcal{Y}$ is the interval $[0, 4]$. The 67 scores y_i, indicated by the histogram, run from 0.02 to 3.08. The dashed curve $\hat{g}_0(y)$ is a normal kernel density estimator with window width $\lambda = 1$, described more carefully in Section 2. Also shown are two special exponential family estimates, $\hat{g}_1(y)$ and $\hat{g}_2(y)$, described below.

Received January 1995; revised October 1995.
[1]Supported by NSF Grant DMS-95-04379 and Public Health Service Grant 5 ROI CA59039-20.
AMS 1991 *subject classifications*. 62F05, 62G05.
Key words and phrases. Poisson regression, degrees of freedom, expected deviance, local and global smoothing, moment-matching.

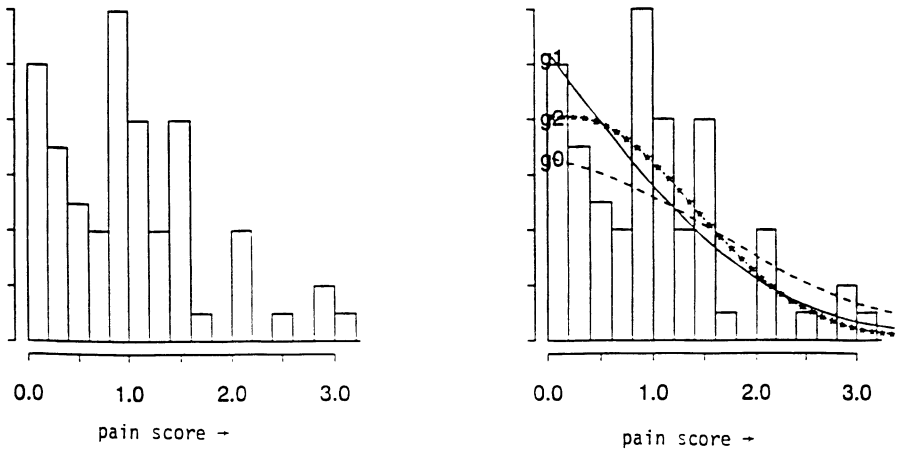

FIG. 1. *Pain-score data. Left*: *Histogram of pain scores for* 67 *women following an operation*: 0 = *no pain*; 4 = *worst pain. Right*: $\hat{g}_0(y)$ *is normal kernel density estimator*, *window width* 1; $\hat{g}_1(y)$ *is the special exponential family through* $\hat{g}_0(y)$ *with sufficient statistic* $t(y) = y$; $\hat{g}_2(y)$ *uses sufficient statistics* $t(y) = (y, y^2)$. *Density* $\hat{g}_2(y)$ *matches the empirical mean and variance of the* 67 *data points*.

An exponential family of densities on $\mathcal{Y}$, $\{g_\beta(y)\}$ is given by

$$g_\beta(y) = g_0(y)\exp\bigl(\beta_0 + t(y)\beta_1\bigr), \tag{1.2}$$

which is called the exponential family through $g_0(y)$ with sufficient statistics $t(y)$. Here $g_0(y)$ is a *carrier density*, $t(y)$ is a $1 \times p$ vector of sufficient statistics, β_1 is a $p \times 1$ parameter vector and β_0 is a normalizing parameter that makes $g_\beta(y)$ integrate to 1 over $\mathcal{Y}$. For example, the one-dimensional normal family, with all possible choices of expectation and variance, can be obtained using the standard normal carrier $g_0(y) = \varphi(y) = \exp\{-0.5y^2\}/\sqrt{2\pi}$, with the sufficient statistics $t(y) = (y, y^2)$. The densities in (1.2) are defined with respect to some background measure, which we will take to be Lebesgue measure. See Section 1.4 of Lehmann (1983).

The estimates $\hat{g}_1(y)$ and $\hat{g}_2(y)$ in Figure 1 are of the form

$$g_{\hat{\beta}}(y) = \hat{g}_0(y)\exp\bigl(\hat{\beta}_0 + t(y)\hat{\beta}_1\bigr), \tag{1.3}$$

where $\hat{g}_0(y)$ is the kernel density estimate indicated by the dashed curve. Estimate $\hat{g}_1(y)$ uses the single sufficient statistic $t(y) = y$, so $p = 1$, while $\hat{g}_2(y)$ uses $t(y) = (y, y^2)$, $p = 2$. The parameter values $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ were chosen by maximum likelihood, that is, by maximizing $\prod_{i=1}^n g_{\hat{\beta}}(y_i)$, ignoring the fact that the carrier $\hat{g}_0(y)$ is itself data-dependent. This choice of β matches the $t(y)$ moments of $g_{\hat{\beta}}(y)$ to their empirical averages:

$$\int_{\mathcal{Y}} t(y)g_{\hat{\beta}}(y)\,dy = \frac{1}{n}\sum_{i=1}^{n} t(y_i). \tag{1.4}$$

Thus $\hat{g}_2$ matches the first two empirical moments of the 67 pain scores or, equivalently, it matches the empirical mean and variance. Property (1.4) implies that the special exponential family estimate of $g(y)$ is unbiased for the moments of $t(y)$. Linear transformations to "post-repair" the mean and variance of a kernel estimate are familiar in the density-estimation literature; see, for example, Jones (1993). The two-dimensional example of Section 4 shows a more ambitious example of moment-matching.

We can think of a special exponential family estimator in two complementary ways: (1) as being a standard exponential family estimator, except one that is preceded by an adaptive choice of the carrier, or (2) as being a standard nonparametric smoothing estimator, except one that is followed by a correction to match certain sample moments.

Either way, we will argue that special exponential families can be a favorable compromise between parametric and nonparametric density estimation. From the first point of view, we will be able to use exponential family theory more flexibly than when restricted to the usual small catalogue of normals, gammas, betas and so forth. An approach more in the spirit of a regression analysis than a density estimate is possible, including exploratory choices of the sufficient statistics $t(y)$.

From the second point of view, the moment-matching correction will usually reduce the bias of a nonparametric smoother, since the moments $t(y)$ are estimated unbiasedly. This has an important practical consequence shown in the numerical examples: *it allows the nonparametric smoother to use a substantially greater window width without badly degrading the overall fit to the data*. The result is a substantially smoother estimate of the density $g(y)$. (This phenomenon is illustrated in the bivariate example of Figures 3 and 4.)

To put things another way, the special exponential family estimate (1.3) works at two different scales. The nonparametric smoother allows local adaptation to the data, while the exponential term matches some of the data's global properties.

Sections 2–5 describe how to compute and interpret special exponential family (SEF) estimates such as those in Figure 1. Most of our computations are done using a Poisson regression model for density estimation (Section 2), originally introduced in Lindsey (1974a, b). Section 3 gives a delta-method formula for the covariance of $\hat{\beta}_1$ in (1.3), which takes into account the data-based choice of the carrier $\hat{g}_0(y)$. We can use this formula in the usual way to select among possible choices of the sufficient statistic. Section 6 concerns formulas for choosing the window width of the smoother that produces $\hat{g}_0(y)$. This choice involves "degrees of freedom" calculations like those introduced by Hastie and Tibshirani (1990). Multisample SEF estimates are discussed in Section 7. Remarks appear in Section 9.

Special exponential families are an example of what Green and Silverman (1994) call *semiparametric methods*. Many other semiparametric methods have been proposed for density estimation. Hjort and Glad (1995) propose reversing the SEF order: first fit a parametric family to the data and then fit a nonparametric smoother to the residuals from the parametric estimator.

Hastie and Tibshirani's *backfitting* approach can also be applied to semiparametric density estimation. It amounts to an iteration to convergence between the smoother and the exponential family. The principal advantage of the SEF estimator is its theoretical tractability. The relationship between these ideas is discussed in Section 8.

Stone (1994) and Kooperberg and Stone (1991) use adaptive exponential families based on spline functions for density estimation. The carrier $g_0(y)$ is effectively the uniform density in this work, but adaptation is still possible because of the local character of the spline basis. Olkin and Spiegelman (1987) combine parametric and nonparametric density estimates by mixtures, rather than by exponential family methods as in this paper.

2. Poisson regression for density estimation. Density estimation problems can be rephrased in terms of Poisson regression models. This technique was introduced by Lindsey (1974a, b) as a way of using generalized linear model software to fit difficult exponential family models. See also Aitken (1993), Lindsey and Mersch (1992) and Efron (1988). Our results here will be phrased in Poisson regression terms, mainly for reasons of conceptual clarity, though there are also some computational advantages.

Lindsey's construction begins with a discretization of the problem: the sample space $\mathcal{Y}$ is partitioned into K disjoint cells $\mathcal{Y}_k$,

$$\mathcal{Y} = \bigcup_{k=1}^{K} \mathcal{Y}_k, \tag{2.1}$$

and the data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are reduced to the cell counts

$$s_k = \#\{y_i \in \mathcal{Y}_k\} \quad \text{for } k = 1, 2, \ldots, K. \tag{2.2}$$

The vector of counts $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ has sum $s_+ = n$. Table 1 shows the counts for the pain-score data of Figure 1, where $\mathcal{Y} = [0, 4]$ has been partitioned into $K = 40$ cells $\mathcal{Y}_k$ each of length 0.1.

Suppose that $\{g_\theta(y),\ \theta \in \Theta\}$ is a family of probability densities on $\mathcal{Y}$. The probability of observing y in the kth cell is

$$\pi_k(\theta) = \int_{\mathcal{Y}_k} g_\theta(y)\, dy, \tag{2.3}$$

TABLE 1

Discretized version of the pain-score data of Figure 1; $\mathcal{Y} = [0, 4]$ *partitioned into* $K = 40$ *cells of length* 0.1; $s_1 = 3$ *of the* 67 *scores occurred in* $\mathcal{Y}_1 = [0, 0.1)$, 7 *in* $\mathcal{Y}_2 = [0.1, 0.2)$ *and so forth*

3	7	6	1	2	3	3	1	7	5
4	4	1	3	3	5	0	1	0	0
2	2	0	0	0	1	0	0	1	1
1	0	0	0	0	0	0	0	0	0

with sum $\pi_+(\theta) = 1$. Then $\mathbf{s}$ has a multinomial distribution on K categories, with n draws and probability vector $\boldsymbol{\pi}(\theta) = (\pi_1(\theta), \pi_2(\theta), \ldots, \pi_K(\theta))$,

$$\mathbf{s} \sim \text{Mult}_K(n, \boldsymbol{\pi}(\theta)). \tag{2.4}$$

We could find the maximum likelihood estimate (MLE) of θ, based on $\mathbf{s}$, by maximizing the multinomial probability of $\mathbf{s}$.

Instead we consider the s_k to be *independent* Poisson observations

$$s_k \overset{\text{ind}}{\sim} \text{Po}(\mu_k(\gamma, \theta)), \qquad k = 1, 2, \ldots, K, \tag{2.5}$$

with expectations

$$\mu_k(\gamma, \theta) = \gamma \pi_k(\theta). \tag{2.6}$$

Here γ is a free parameter, restricted only to be positive. Standard Poisson properties allow (2.5) and (2.6) to be expressed as

$$s_+ \sim \text{Po}(\gamma) \quad \text{and} \quad s|s_+ \sim \text{Mult}_K(s_+, \boldsymbol{\pi}(\theta)). \tag{2.7}$$

This means that the maximum likelihood estimates from (2.5) and (2.6) are

$$\hat{\gamma} = s_+ = n, \tag{2.8}$$

and $\hat{\theta}$ is equal to the MLE for θ in (2.4). Lindsey's method is to (approximately) maximize the original likelihood $\prod_{i=1}^n g_\theta(y_i)$ by finding the Poisson MLE in (2.5) and (2.6). *Note*: The parameters γ and θ are orthogonal in Cox and Reid's (1987) sense, so that the information for estimating θ is the same in (2.4) and (2.5).

This method of finding the MLE is particularly convenient when the original densities are of the exponential family form (1.2). We will consider the density estimation problem (1.1) and (1.2) in the Poisson regression form (2.5) and (2.6):

$$s_k \overset{\text{ind}}{\sim} \text{Po}(\mu_k(\beta)) \quad \text{for } k = 1, 2, \ldots, K, \tag{2.9a}$$

with

$$\mu_k(\beta) = \mu_k^o \exp(\beta_0 + t_k \beta_1). \tag{2.9b}$$

Here μ_k^o is proportional to $\pi_k^o = \int_{\mathcal{Y}_k} g_0(y)\, dy$, a discretized version of the carrier, and

$$t_k = t(y_{(k)}), \tag{2.10}$$

the sufficient vector $t(y)$ evaluated at a convenient point $y_{(k)}$ in $\mathcal{Y}_k$. The free parameter β_0 corresponds to $\gamma = e^{\beta_0}$ in (2.6).

Define X to be the $K \times (p + 1)$ matrix whose kth row equals

$$x_k = (1, t_k). \tag{2.11}$$

The maximum likelihood equations for $\beta = (\beta_0, \beta_1)$ in the generalized linear model (2.9) are

$$X'\left[\mathbf{s} - \boldsymbol{\mu}(\hat{\beta})\right] = 0, \tag{2.12}$$

where $\boldsymbol{\mu}(\hat{\beta})$ indicates the vector with kth component $\mu_k^o \exp(x_k \hat{\beta})$. Standard generalized linear model software easily solves for $\hat{\beta}$ in (2.12), even for difficult nonstandard forms of the exponential family (1.2). This was Lindsey's principal point.

Here is how Figure 1 was constructed. The problem was discretized into $K = 40$ cells as in Table 1. The carrier vector $\boldsymbol{\mu}^o = (\mu_1^o, \mu_2^o, \ldots, \mu_K^o)$ in (2.9b) was estimated using a $K \times K$ smoothing matrix $M(\lambda)$,

$$\hat{\boldsymbol{\mu}}^o = M(\lambda)\mathbf{s}. \tag{2.13}$$

Matrix $M(\lambda)$ was taken to be a normal kernel smoother, having kjth element

$$M_{kj}(\lambda) = \frac{c_k}{\lambda}\varphi\left(\frac{y_{(k)} - y_{(j)}}{\lambda}\right), \tag{2.14}$$

with $y_{(k)} = (k - 0.5)/10$, the midpoint of cell $\mathscr{Y}_k$. The constants c_k were chosen to make $M_{k+} = 1$. The starred curve labelled $\hat{g}_0$ in Figure 1 is actually $\hat{\mu}_k^o$ plotted as a function of $y_{(k)}$, with the window width λ in (2.14) set equal to 1.

The curves labelled $\hat{g}_1(y)$ and $\hat{g}_2(y)$ are really the discrete analogs of the special exponential family (1.3), say $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K)$,

$$\hat{\mu}_k = \hat{\mu}_k^o \exp\left(\hat{\beta}_0 + t_k \hat{\beta}_1\right), \tag{2.15}$$

plotted versus $y_{(k)}$: $\hat{g}_1$ uses $t_k = y_{(k)}$, while $\hat{g}_2$ is based on the quadratic vector $t_k = (y_{(k)}, y_{(k)}^2)$. The MLE estimates $\hat{\beta}$ were obtained by iterative solution of (2.12), so

$$X'[\mathbf{s} - \hat{\boldsymbol{\mu}}] = 0, \tag{2.16}$$

where X is a 40×2 matrix for $\hat{g}_1$ and a 40×3 matrix for $\hat{g}_2$. [The estimates were actually computed using a centered version of y, $\tilde{y}_{(k)} = (y_{(k)} - 2)/4$.] To fit this model in the GLIM language or the `glm` function in SPlus, one simply includes $\log \hat{\mu}^o$ as an offset in a Poisson generalized linear model.

Equation (2.16) shows that

$$\hat{\mu}_+ = s_+ = n \tag{2.17a}$$

and

$$\sum_{k=1}^{K} \frac{\hat{\mu}_k}{n} t_k = \sum_{k=1}^{K} \frac{s_k}{n} t_k, \tag{2.17b}$$

the discrete analog of the moment-matching property (1.4). Notice that because of (2.17a) and the fact that the cells are of length 0.1, the curves in Figure 1 integrate over $\mathscr{Y}$ to 6.7 rather than to 1.

Changing K to 20 or to 80 made very little difference in Figure 1. The numerical calculations in this paper were insensitive to the form of discretization. In fact, discretization is not really necessary for any of our results, as discussed in Remark E of Section 9.

Nevertheless, it is conceptually easier to discuss special exponential family density estimation in terms of the discrete Poisson model (2.9). The problem becomes one of fitting a smooth regression curve to the independent observations s_k, and this lets us make use of the arsenal of regression tools. It also emphasizes the important point that density estimation is equivalent to Poisson regression, and not to ordinary least squares regression. The Poisson nature of the problem will be evident in the formulas developed below.

3. Estimating the covariance of $\hat{\boldsymbol{\beta}}$. This section derives an approximate covariance matrix for $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, the estimated parameters in a special exponential family such as (1.3) or (2.15). The formula for the covariance takes into account the data-based choice of the carrier. We will use the estimated standard errors of the components of $\hat{\beta}$ for model building, checking the significance of the corresponding components of the sufficient statistic $t(y)$ in the usual way.

We consider the Poisson form (2.9) of the SEF model,

$$\mathbf{s} \sim \mathrm{Po}_K(\boldsymbol{\mu}(\beta)) \quad \text{with } \boldsymbol{\mu}(\beta) = \boldsymbol{\mu}^o e^{X\beta}, \tag{3.1}$$

where $\mathrm{Po}_K(\boldsymbol{\mu})$ indicates a vector of K independent Poisson variates having expectations $(\mu_1, \mu_2, \ldots, \mu_K) = \boldsymbol{\mu}$. The notation $\boldsymbol{\mu}^o e^{X\beta}$ indicates the vector with kth component $\mu_k^o e^{x_k\beta}$, $x_k = (1, t_k)$, as in (2.11). Generalizing (2.13), we first estimate $\boldsymbol{\mu}^o$ by some function of $\mathbf{s}$, say $m(\mathbf{s})$, and then solve for $\hat{\beta}$ in the MLE equations (2.16):

$$\hat{\boldsymbol{\mu}}^o = m(\mathbf{s}) \quad \text{and} \quad \hat{\beta}\colon X'\left[\mathbf{s} - \hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}\right] = 0. \tag{3.2a}$$

The special exponential family estimate of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}. \tag{3.2b}$$

LEMMA 1. *Let $\hat{D}$ be the $K \times K$ diagonal matrix with kth diagonal element $\hat{\mu}_k = \hat{\mu}_k^o e^{x_k\hat{\beta}}$ and let $\hat{H}$ be the $K \times K$ derivative matrix of $\log(\hat{\boldsymbol{\mu}}^o) = (\log(\hat{\mu}_1^o), \log(\hat{\mu}_2^o), \ldots, \log(\hat{\mu}_K^o))$ with respect to $\mathbf{s}$, with "j" indexing columns,*

$$\hat{H} = \frac{d\log\hat{\boldsymbol{\mu}}^o}{d\mathbf{s}} = \left(\frac{\partial\log(\hat{\mu}_k^o)}{\partial s_j}\right). \tag{3.3}$$

Then the $(p+1) \times K$ derivative matrix of $\hat{\beta}$ with respect to $\mathbf{s}$ is

$$\frac{d\hat{\beta}}{d\mathbf{s}} = [X'\hat{D}X]^{-1}Z', \tag{3.4a}$$

where

$$Z' = X'(I - \hat{D}\hat{H}). \tag{3.4b}$$

In case (2.13), $\hat{\boldsymbol{\mu}}^o = M\mathbf{s}$, *this becomes*

$$Z' = X'\left(I - D(e^{X\hat{\beta}})M\right), \tag{3.4c}$$

where $D(e^{X\hat{\beta}})$ is the diagonal matrix with kth diagonal element $e^{x_k\hat{\beta}}$. (*The proof appears below.*)

By the usual delta-method argument, an approximate covariance matrix estimate for $\hat{\beta}$ is given by

$$\left(\frac{d\hat{\beta}}{d\mathbf{s}}\right)\widehat{\mathrm{Cov}}(\mathbf{s})\left(\frac{d\hat{\beta}}{d\mathbf{s}}\right)'. \tag{3.5}$$

COROLLARY 1. *The* SEF *vector* $\hat{\beta}$ *obtained from* (3.1) *and* (3.2) *has approximate covariance matrix*

$$\overline{\mathrm{Cov}}(\hat{\beta}) = [X'\hat{D}X]^{-1}[Z'\overline{D}Z][X'\hat{D}X]^{-1}, \tag{3.6a}$$

where $\overline{D}$ *is the diagonal matrix with kth diagonal element* s_k. *An alternative estimate is*

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = [X'\hat{D}X]^{-1}[Z'\hat{D}Z][X'\hat{D}X]^{-1}. \tag{3.6b}$$

For any K-vector $\mathbf{v}$, we let $D(\mathbf{v})$ be the $K \times K$ diagonal matrix with kth diagonal element v_k. The true covariance of $\mathbf{s} \sim \mathrm{Po}_k(\boldsymbol{\mu})$ is $D(\boldsymbol{\mu})$. Approximation (3.6a) estimates $\mathrm{Cov}(\mathbf{s})$ by $D(\mathbf{s}) = \overline{D}$ in (3.5), while (3.6b) uses $D(\hat{\boldsymbol{\mu}}) = \hat{D}$. The former may be preferred if model (3.1) is suspect. In our numerical examples the two formulas gave nearly the same results.

Table 2 applies the corollary to the pain-score data. The quadratic model, described in (2.13)–(2.16) has $\hat{\beta}_1 = (-2.74, -3.80)$, with standard errors $(0.93, 2.45)$ according to (3.6a). [$\hat{\beta}_0$ serves only to normalize $\hat{\boldsymbol{\mu}}$ to $\hat{\mu}_+ = n$, (2.17a), so its value is of no statistical interest.] The coefficient of $\tilde{y} = (y - 2)/4$, the centered version of y, is nearly 3 standard errors below zero. The coefficient of $\tilde{y}^2$ is -1.55 standard errors below zero, so it is not so clear that the quadratic term is significant. The cubic model, in which $t(y) = (\tilde{y}, \tilde{y}^2, \tilde{y}^3)$, has the cubic coefficient only 0.07 standard errors above zero. Either the linear or the quadratic model seem reasonable here; the cubic model is definitely excessive. Section 6 discusses model selection in more detail.

TABLE 2

Parameter estimates $\hat{\beta}$ *and estimated standard errors for the pain-score data of Table* 1. *The quadratic model is described in* (2.13)–(2.16); $\tilde{y} = (y - 2)/4$, *centered version of* y; $\overline{\mathrm{se}}$ *square root of diagonal elements* (3.6a); $\widehat{\mathrm{se}}$ *from* (3.6b); "*jack*" *is jackknife standard error*; "*naive*" *is the usual exponential family* sterr *estimated ignoring data-based choice of carrier*. *The cubic model uses the same* M, *but* $t(y) = (\tilde{y}, \tilde{y}^2, \tilde{y}^3)$; *the cubic term is not at all significant*

	Quadratic model						**Cubic model**	
	$\hat{\boldsymbol{\beta}}_1$	$\overline{\mathbf{se}}$	**(ratio)**	$\widehat{\mathbf{se}}$	**jack**	**naive**	$\hat{\boldsymbol{\beta}}_1$	$\overline{\mathbf{se}}$
$\tilde{y}$	−2.74	0.93	(−2.95)	0.96	1.07	1.18	−2.78	1.24
$\tilde{y}^2$	−3.80	2.45	(−1.55)	2.50	2.66	2.85	−3.59	2.70
$\tilde{y}^3$							0.59	8.30

The jackknife standard errors for the components of $\hat{\beta}_1$ [formula (11.5) of Efron and Tibshirani (1993)] were computed as a check on the corollary. They came out a little larger than the delta-method estimates from (3.6a) or (3.6b), as is often the case; see Section 2 of Efron (1992).

Suppose that we ignore the fact that the carrier $\hat{\boldsymbol{\mu}}^o$ in (3.2) is a function of the data $\mathbf{s}$. This amounts to taking $\hat{H} = 0$ in (3.3), so $Z' = X'$ in (3.4b). Then (3.6b) reduces to the usual covariance estimate for exponential families,

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = (X'\hat{D}X)^{-1}, \tag{3.7}$$

while (3.6a) becomes the "sandwich" estimate $(X'\hat{D}X)^{-1}(X'\overline{D}X)(X'\hat{D}X)^{-1}$. The standard error estimates from (3.7), labelled "naive" in Table 2, are considerably larger than the standard errors that take into account the adaptive choice of the carrier.

The naive standard errors will usually exceed those from (3.6). The reason is that the adaptive choice of $\hat{\boldsymbol{\mu}}^o$ absorbs some of the variability in $\hat{\beta}$. Here is a simple normal-theory version of the same phenomenon: suppose we observe $z \sim N(\mu^o + \beta, 1)$ and we wish to estimate β, the distance of the expectation $\mu = \mu^o + \beta$ from some origin of measurement μ^o. Then $\hat{\beta} = z - \mu^o$ has standard error 1. However, if the origin is chosen adaptively, say by $\hat{\mu}^o = 0.75 \cdot z$, then $\hat{\beta} = z - \hat{\mu}^o$ has standard error 0.25. Observing $z = 4$, for example, gives $\hat{\beta} = 1$, which in the adaptive case is 4 standard errors above 0. See Remark B in Section 9.

PROOF OF LEMMA 1. Using the D notation for diagonal matrices, (3.2) for $\hat{\beta}$ can be expressed as

$$X'\left[\mathbf{s} - D(\hat{\boldsymbol{\mu}}^o)e^{X\hat{\beta}}\right] = 0, \tag{3.8}$$

where $e^{x\hat{\beta}}$ is the vector with components $e^{x_k\hat{\beta}}$. A small change $d\mathbf{s}$ in $\mathbf{s}$ produces change $d\hat{\beta}$ in the MLE vector and change

$$d\hat{\boldsymbol{\mu}}^o \doteq D(\hat{\boldsymbol{\mu}}^o)\hat{H}\,d\mathbf{s} \tag{3.9}$$

in $\hat{\boldsymbol{\mu}}^o$. Then (3.8) gives

$$\begin{aligned} \mathbf{0} &= X'\left[\mathbf{s} + d\mathbf{s} - D(\hat{\boldsymbol{\mu}}^o + d\hat{\boldsymbol{\mu}}^o)\exp\left(X(\hat{\beta} + d\hat{\beta})\right)\right] \\ &\doteq X'\left[\mathbf{s} + D(\hat{\boldsymbol{\mu}}^o)\exp\left(X\hat{\beta}\right) + d\mathbf{s} - D(d\hat{\boldsymbol{\mu}}^o)\exp\left(X\hat{\beta}\right)\right. \\ &\qquad \left. - D\left(\hat{\boldsymbol{\mu}}^o \exp\left(X\hat{\beta}\right)\right)X\,d\hat{\beta}\right]. \end{aligned} \tag{3.10}$$

Using (3.8), (3.9) and the fact that $D(d\hat{\boldsymbol{\mu}}^o)e^{X\hat{\beta}} = D(e^{X\hat{\beta}})\,d\hat{\boldsymbol{\mu}}^o$, (3.10) becomes

$$\begin{aligned} 0 &= X'\,d\mathbf{s} - X'D\left(\hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}\right)\hat{H}\,d\mathbf{s} - X'D\left(\hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}\right)X\,d\hat{\beta} \\ &= X'\left[I - \hat{D}\hat{H}\right]d\mathbf{s} - X'\hat{D}X\,d\hat{\beta}. \end{aligned} \tag{3.11}$$

This verifies (3.4). □

4. A bivariate example. Density estimation becomes more interesting, and more difficult, when the sample space $\mathcal{Y}$ is of higher dimension. This section introduces an example when $\mathcal{Y}$ is a portion of the plane. We will use this example to illustrate some of the advantages of special exponential family density estimation.

Figure 2 shows $l = \log(\text{redshift})$ and m equal to the apparent magnitude for 486 galaxies taken from Loh and Spillar's (1988) redshift survey,

$$y_i = (l_i, m_i) \quad \text{for } i = 1, 2, 3, \ldots, n = 486. \tag{4.1}$$

Hubble's law, that larger redshift implies greater distance from Earth, is apparent in the figure. The galaxies with larger values of l tend to appear dimmer, that is, to have larger apparent magnitudes, leaving the lower right corner nearly empty.

The data in Figure 2 are a truncated subsample of the 879 galaxies in the Loh–Spillar catalog. It is all of the catalog entries falling into the rectangle

$$\log(0.2) \le l \le \log(1.2) \quad \text{and} \quad 17.2 \le m \le 21.5. \tag{4.2}$$

We will take this rectangle to be the sample space $\mathcal{Y}$. Some of the scientific reasons for truncation are discussed in Efron and Petrosian (1992).

Figure 2 shows $\mathcal{Y}$ partitioned into $K = 285$ rectangular cells $\mathcal{Y}_k$, by dividing the l axis into 15 equal strips and dividing the m axis into 19 equal strips. The corresponding counts s_k, (3.2), are shown on the right side of the figure. It is convenient to index the cells by

$$k = (i, j), \quad i = 1, 2, \ldots, 15, \; j = 1, 2, \ldots, 19. \tag{4.3}$$

The midpoint of rectangle $\mathcal{Y}_k$ is $y_{(k)} \equiv (l_{(i)}, m_{(j)})$.

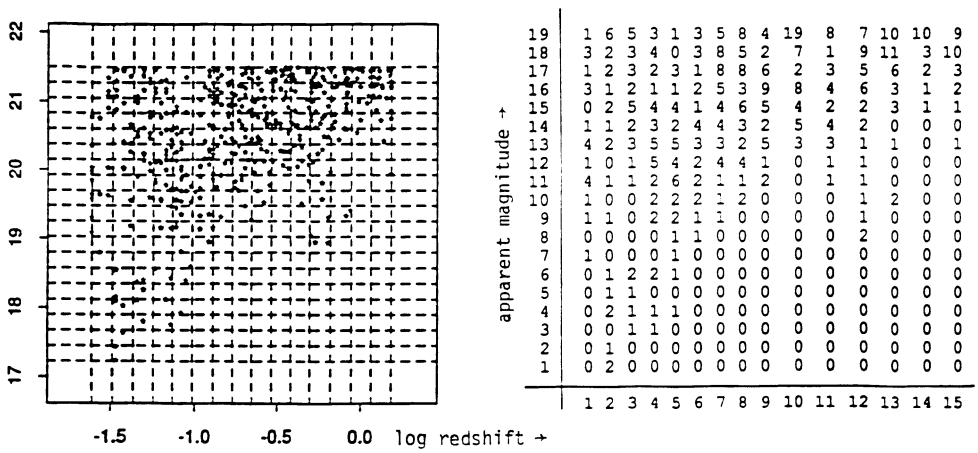

FIG. 2. *The galaxy data*: 486 *galaxies from Loh and Spillar's* (1988) *redshift survey* (log *redshift* l *and apparent magnitude* m) *discretized into* $285 = 15 \times 19$ *equal cells. The counts* $\mathbf{s}$ *for the* 285 *cells are shown at right.*

The right side of Figure 3 shows the results of applying a linear smoother $\hat{\boldsymbol{\mu}} = M(\lambda)\mathbf{s}$ to the 285-dimensional count vector $\mathbf{s}$ given in Figure 2. The $K \times K$ matrix $M(\lambda)$ is a two-dimensional version of (2.14), with kk'th element

$$M_{kk'}(\lambda) = \frac{c_k}{\lambda^2}\exp\left(-\frac{1}{2\lambda^2}\left[(i-i')^2 + (j-j')^2\right]\right) \tag{4.4}$$

for $k = (i, j)$ and $k' = (i', j')$. The c_k are chosen to make $\sum_{k'} M_{kk'}(\lambda) = 1$. The choice $\lambda = 1.5$, suggested by the expected deviance calculations of Section 6, gave the smoothed estimate $\hat{\boldsymbol{\mu}}^o = M(1.5) \cdot \mathbf{s}$ plotted versus $y_{(k)}$ on the right side of Figure 3.

The left side of Figure 3 shows an SEF estimate $\hat{\boldsymbol{\mu}}$ of form(3.2), with

$$\hat{\boldsymbol{\mu}}^o = M(2)\mathbf{s} \quad \text{and} \quad X = \left(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{i}^2, \mathbf{j}^2, \mathbf{ij}\right). \tag{4.5}$$

Here $\mathbf{i}$ is the K vector with kth element $i - 8$, and $\mathbf{j}$ is the K vector with kth element $j - 9$. This choice amounts to using the usual sufficient statistics for a bivariate normal in (1.2), $t(y) = (l, m, l^2, m^2, lm)$. The fitted density matches the empirical means, variances and correlation of the galaxy data.

The variance calculations of Section 5 and the expected deviance calculations of Section 6 suggest that these two estimates are roughly equal in their overall ability to predict the true density. However, the SEF estimate is much smoother than the smoothing-only choice. This is obvious in Figure 4, which shows contour plots of the two density estimates.

Suppose that the carrier $\hat{\boldsymbol{\mu}}^o$ in (4.5) was taken to be the constant vector $\hat{\mu}^o_k \equiv 1$ instead of $M(2)\mathbf{s}$. Then the SEF $\hat{\boldsymbol{\mu}}$ would be the (discretized) truncated bivariate normal MLE for the galaxy data, the truncation being to the rectangle (4.2). In fact, the SEF in Figure 3 looks like the lower corner of a bivariate normal, though there are some discrepancies due to the adaptation of $\hat{\boldsymbol{\mu}}^o$ to the galaxy data.

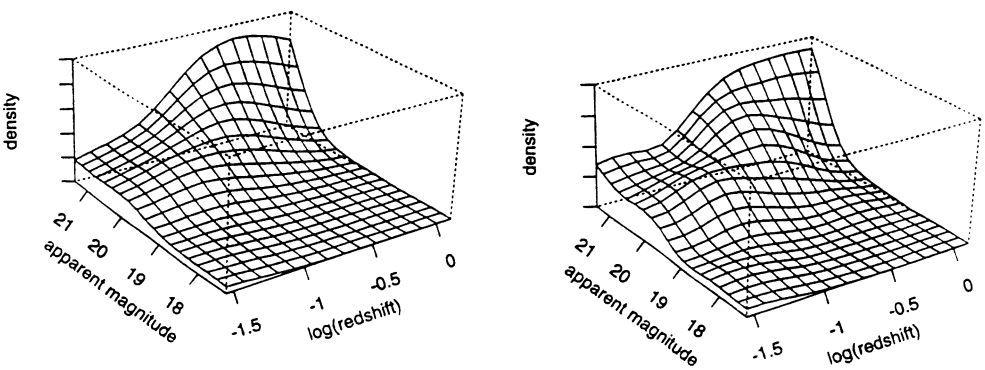

FIG. 3. *Left panel*: *SEF estimate* (3.2) *for the galaxy data as discretized in Figure* 2; $\hat{\mu}^o = M(2) \cdot \mathbf{s}$, *quadratic matrix* X, (4.5). *Right panel*: *The smoothing-only estimate* $\hat{\mu}^o = M(1.5) \cdot \mathbf{s}$. *The calculations of Sections* 5 *and* 6 *suggest that the two estimates are roughly equal in accuracy. However, the SEF estimate is much smoother.*

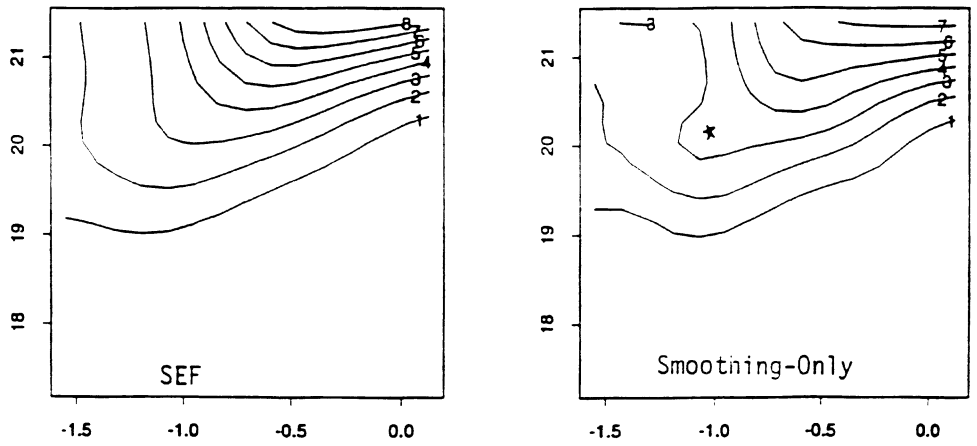

FIG. 4. *Contour plots of the density estimates in Figure* 3. *The SEF contours* (*left panel*) *are much smoother than those from the smoothing-only estimate* (*right panel*). *The smoothing-only estimate has a weak second mode at the starred point*.

Parametric models such as the bivariate normal are fierce data smoothers. This can be a big advantage if the statistician is interested in global properties of the density, like the general shape of its contours, and especially if the data-sampling process is suspect. In Figure 2 we can see that the galaxy data are quite patchy and clumpy, which is not surprising since the Loh–Spillar catalog was a census of the brighter galaxies in a few degrees of sky, and not a random sample of the full sky.

If those few degrees of sky are of great individual interest, then a narrow-band smoother like that in the right side of Figure 3 may be appropriate. [Notice that it estimates a weak second mode near $(l, m) = (-1, 20)$.] However if we really want to know the density for the whole sky, then it pays to *oversmooth* the estimate from a flawed sample like that in Figure 2. The SEF methodology allows us to oversmooth without losing much estimating efficiency compared to smoothing-only estimates and without making the drastic assumptions of the usual parametric models.

Various SEF models besides the quadratic choice of X in (4.5) were tried. One of these added a further cross-term to (4.5),

$$\hat{\boldsymbol{\mu}}^o = M(2)\mathbf{s} \quad \text{and} \quad X = \left(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{i}^2, \mathbf{j}^2, \mathbf{ij}, (\mathbf{i}^2\mathbf{j}^2)_{\text{orthog}}\right), \tag{4.6}$$

where $(\mathbf{i}^2\mathbf{j}^2)_{\text{orthog}}$ is the component of $\mathbf{i}^2\mathbf{j}^2$ orthogonal to $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{i}^2, \mathbf{j}^2, \mathbf{ij})$. Orthogonalization makes the first six components of $\hat{\beta}$ have roughly the same values and standard errors as in (4.5). The term $(\mathbf{i}^2\mathbf{j}^2)_{\text{orthog}}$ in (4.6) allows the regression surface to turn more quickly near the corners of the rectangle $\mathscr{Y}$.

The MLE vector $\hat{\beta}_1$ for model (4.6) appears in Table 3, along with the standard error estimates $\overline{\text{se}}$ from (3.6a) and the t-values $\hat{\beta}/\overline{\text{se}}$. All of the coefficients except the one for m^2 are significantly different than zero [the

TABLE 3

Parameter estimates and standard errors for the components of $\hat{\beta}_1$ *in the* SEF *model* (4.6). *All of the components of* $\hat{\beta}_1$ *are significantly nonzero, except for the coefficient of* m^2. *l indicates the coefficient corresponding to* **i**, *lm to* **ij** *and so forth. Standard errors are from* (3.6a). *Model* (4.5) *gave similar estimates, standard errors and t-values for* l, m, l^2, m^2, lm

	l	m	l^2	m^2	lm	$(l^2m^2)_{\text{orthog}}$
$\hat{\beta}_1$	−0.105	0.084	−0.0115	−0.0018	0.0158	0.000267
se	0.019	0.015	0.0021	0.0014	0.0027	0.000071
t-Value	−5.6	5.8	−5.5	−1.4	5.8	3.75

same is true for model (4.5)]. This includes the coefficient for the term $(l^2m^2)_{\text{orthog}}$, which takes us beyond the normal-theory SEF analogue (4.5). Compared to the left side of Figure 3, the SEF density estimate from (4.6) has its highest point at the upper right corner of the rectangle $\mathscr{Y}$.

The next two sections discuss several criteria for choosing among possible SEF models: observed deviance, expected deviance, degrees of freedom and so forth. However, none of these criteria is sharp enough to entirely free the statistician from model-choice quandries. A considerable amount of subjectivity must still go into the model-building process, just as in ordinary regression situations.

5. Total relative variance. A variety of diagnostic tools is available to assist model selection in standard regression situations. Similar tools are available for model selection in special exponential families. These ideas are developed in the next two sections, beginning here with the total relative variance, a simple measure of overall variability for an SEF density estimate. For example, looking ahead to Table 4 the reader can compare the total relative variances for the two galaxy-data SEF estimates in Figure 3: 6.8 for the quadratic model in the left panel versus 8.8 for the smoothing-only model in the right panel.

Let

$$\hat{\boldsymbol{\mu}} = \text{SEF}(\mathbf{s}; m, X) \tag{5.1}$$

indicate $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}$, the special exponential family estimate (3.2). First we will compute the $K \times K$ derivative matrix of $\hat{\boldsymbol{\mu}}$ with respect to $\mathbf{s}$, which leads immediately to a delta-method estimate of $\text{Cov}(\hat{\boldsymbol{\mu}})$. This derivative matrix involves the projection matrices

$$\hat{P} = X\left(X'\hat{D}X\right)^{-1}X' \quad \text{and} \quad \hat{Q} = \hat{D}^{-1} - \hat{P}, \tag{5.2}$$

where $\hat{D} = D(\hat{\boldsymbol{\mu}})$ as before, $\hat{P}$ is the symmetric projection matrix into the linear space spanned by the columns of X, in the inner product $\hat{D}$, the projection of vector $\mathbf{v}$ being $\hat{P}\hat{D}\mathbf{v}$, and $\hat{Q}$ is the projection orthogonal to X's column space. Because they represent orthogonal projections, we have

$$\hat{P}\hat{D}\hat{P} = \hat{P}, \qquad \hat{Q}\hat{D}\hat{Q} = \hat{Q} \quad \text{and} \quad \hat{P}\hat{D}\hat{Q} = 0. \tag{5.3}$$

LEMMA 2. *The derivative matrix of* $\hat{\boldsymbol{\mu}} = \mathrm{SEF}(\mathbf{s}; m, X)$ *with respect to* $\mathbf{s}$ *is*

$$\frac{d\hat{\boldsymbol{\mu}}}{d\mathbf{s}} = \hat{D}\hat{O} \tag{5.4a}$$

where, in terms of $\hat{H} = d\log\hat{\boldsymbol{\mu}}^o/d\mathbf{s}$, (3.3),

$$\hat{O} = \hat{P} + \hat{Q}\hat{D}\hat{H} = \hat{P} + \hat{H} - \hat{P}\hat{D}\hat{H}. \tag{5.4b}$$

If $\hat{\boldsymbol{\mu}}^o = M\mathbf{s}$, *then* $\hat{H} = D(1/\hat{\boldsymbol{\mu}}^o)M$ *and*

$$\hat{O} = \hat{P} + \hat{Q}D\big(e^{X'\hat{\beta}}\big)M. \tag{5.4c}$$

(The proof appears below.)

The canonical parameter vector for the Poisson family (3.1) is

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \big(\log(\mu_1), \log(\mu_2), \ldots, \log(\mu_k)\big). \tag{5.5}$$

Likewise, define $\hat{\boldsymbol{\eta}} = \log(\hat{\boldsymbol{\mu}})$ and $\hat{\boldsymbol{\eta}}^o = \log(\hat{\boldsymbol{\mu}}^o)$. Then we can write Lemma 2 as

$$\frac{d\hat{\boldsymbol{\eta}}}{d\mathbf{s}} = \hat{O} = \hat{P} + \hat{Q}\hat{D}\frac{d\hat{\eta}^o}{d\mathbf{s}}. \tag{5.6}$$

This decomposes $d\hat{\boldsymbol{\eta}}/d\mathbf{s}$ into a part $\hat{P}$ coming from the exponential family factor $e^{X\hat{\beta}}$ and an orthogonal part $\hat{Q}\hat{D}(d\hat{\boldsymbol{\eta}}^o/d\mathbf{s})$ coming from the adaptive choice of the carrier.

Lemma 2 leads directly to delta-method estimates of the covariance matrix of $\hat{\boldsymbol{\mu}}$, as in (3.6),

$$\overline{\mathrm{Cov}}(\hat{\boldsymbol{\mu}}) = \hat{D}\hat{O}\overline{D}\hat{O}'\hat{D} \quad \text{or} \quad \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}}) = \hat{D}\hat{O}\hat{D}\hat{O}'\hat{D} \tag{5.7}$$

with $\overline{D} = D(s)$. The diagonal elements give variance estimates $\overline{\mathrm{var}}(\hat{\mu}_k)$ or $\widehat{\mathrm{var}}(\hat{\mu}_k)$ for the individual components. For Poisson variables it is natural to measure variance relative to the etsimate $\hat{\mu}_k$. We define the *total relative variance* (TRV) estimate for $\hat{\boldsymbol{\mu}}$ to be

$$\overline{\mathrm{TRV}} = \sum_{k=1}^{K} \frac{\overline{\mathrm{var}}(\hat{\mu}_k)}{\hat{\mu}_k} \quad \text{or} \quad \widehat{\mathrm{TRV}} = \sum_{k=1}^{K} \frac{\widehat{\mathrm{var}}(\hat{\mu}_k)}{\hat{\mu}_k}. \tag{5.8}$$

COROLLARY 2. *For* $\hat{\boldsymbol{\mu}} = \mathrm{SEF}(\mathbf{s}; m, X)$ *the total relative variance estimates are*

$$\begin{aligned} \overline{\mathrm{TRV}} &= \mathrm{tr}\Big[\overline{D}\hat{P} + \big(\hat{H}\overline{D}\hat{H}'\big)\big(\hat{D}\hat{Q}\hat{D}\big)\Big] \quad \textit{or} \\ \widehat{\mathrm{TRV}} &= (p+1) + \mathrm{tr}\big(\hat{H}\hat{D}\hat{H}'\big)\big(\hat{D}\hat{Q}\hat{D}\big), \end{aligned} \tag{5.9}$$

where $p+1$ *is the number of columns of* X, $\overline{D} = D(\mathbf{s})$, $\hat{D} = D(\hat{\boldsymbol{\mu}})$ *and* $\hat{P}, \hat{Q}$ *are as in* (5.2).

The proof of Corollary 2 follows directly from (5.7) by writing $\overline{\text{TRV}}$ in the trace form $\text{tr}\,\hat{D}^{-1}\,\overline{\text{Cov}}(\hat{\mu})$ and using the orthogonality relationship (5.3), and similarly for $\widehat{\overline{\text{TRV}}}$. Computationally more efficient expressions for $\overline{\text{TRV}}$ and $\widehat{\overline{\text{TRV}}}$ appear in Remark I, Section 9.

Table 4 shows $\overline{\text{TRV}}$ for various SEF estimates for the galaxy data, as discretized in Figure 2. The three estimates correspond to three choices of X in (5.1): sef 2 is for X as in (4.5), sef 3 for X as in (4.6), and sef 0 for $X = \mathbf{1}$. This last choice is the smoothing-only estimate $\hat{\boldsymbol{\mu}}^o$ rescaled to $(n/\hat{\mu}^o_+)\hat{\boldsymbol{\mu}}^o$, so that it sums to $n = s_+$. The carrier $\hat{\boldsymbol{\mu}}^o$ for the SEF is

$$\hat{\boldsymbol{\mu}}^o = M(\lambda)\mathbf{s}, \tag{5.10}$$

where $M(\lambda)$ is the matrix (4.4).

All three TRV estimates decrease as the smoothing parameter λ increases because greater window width λ decreases the variability of $\hat{\boldsymbol{\mu}}^o$. If the carrier $\hat{\boldsymbol{\mu}}^o$ were prechosen instead of adaptive, then sef 2 would exceed sef 0 by about 5, this being the increased number of free parameters, and likewise sef 3 would exceed sef 2 by about 1. This is seen in (5.9) for the case $\hat{H} = 0$. In fact, this is nearly the case at the right side of Table 4, where λ is so large that $\hat{\boldsymbol{\mu}}^o$ has nearly constant entries. The difference between the three estimates decreases at smaller values of λ because the adaptability of the common carrier $\hat{\boldsymbol{\mu}}^o = M(\lambda)\cdot\mathbf{s}$ absorbs some of the difference in the exponential family fits $e^{X\hat{\beta}}$.

Small variability is a good property of course, but we also want $\hat{\boldsymbol{\mu}}$ to have small bias for estimating the true density vector $\boldsymbol{\mu}$. The next section puts TRV into the context of bias–variance tradeoffs for Poisson regression estimates.

PROOF OF LEMMA 2. In the notation following (5.5), $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^o + X\hat{\beta}$. Differentiating this with respect to $\mathbf{s}$ and using (3.4) gives

$$\frac{d\hat{\boldsymbol{\eta}}}{d\mathbf{s}} = \hat{H} + X(X'\hat{D}X)^{-1}X'\left[I - \hat{D}\hat{H}\right] = \hat{P} + \hat{H} - \hat{P}\hat{D}\hat{H}, \tag{5.11}$$

which is (5.6), the equivalent of Lemma 2. □

TABLE 4

Total relative variance $\overline{\text{TRV}}$ *for three* SEF *estimates, galaxy data, at increasing values of smoothing parameter* λ

	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$
sef 0 (smoothing only)	8.8[a]	5.2	1.2	0.4
sef 2 (4.5)	9.6	6.8[b]	5.2	5.2
sef 3 (4.6)	10.2	7.6	6.2	6.2

[a] Right panel of Figure 3.
[b] Left panel of Figure 3.

6. Degrees of freedom and estimated deviance. Selecting a good SEF estimate $\hat{\boldsymbol{\mu}}$ for a particular application involves making the usual tradeoffs between variance and bias. This section concerns estimating the total expected deviance of $\hat{\boldsymbol{\mu}}$ from the expectation vector $\boldsymbol{\mu}$. This is a measure of accuracy for $\hat{\boldsymbol{\mu}}$ that involves both variance and bias. An important role is played by the *degrees of freedom* of the estimator $\hat{\boldsymbol{\mu}}$, an idea related to the total relative variance of Section 5. The ideas here are an extension of those in Section 6.8 of Hastie and Tibshirani (1990).

Figure 5 relates to the *expected deviance* measure $\widehat{\text{EDEV}}$ developed below, a diagnostic measure for comparing the goodness-of-fit of different SEF models. It shows $\widehat{\text{EDEV}}$ for both the pain-score and galaxy examples. $\widehat{\text{EDEV}}$ is plotted versus the smoothing parameter λ for the normal kernel estimates (2.13) and (4.4) used to obtain $\hat{\boldsymbol{\mu}}^o = M(\lambda) \cdot \mathbf{s}$. The different curves correspond to different choices of X in the SEF formula (3.2). "Smoothing-only" refers to $X = \mathbf{1}$, which gives the estimate $\hat{\boldsymbol{\mu}}^o$ renormalized to sum to n. For the pain-score data, "linear" and "quadratic" are the cases referred to in (2.15). For the galaxy data, (4.5) and (4.6) are as in Table 4.

Notice that $\widehat{\text{EDEV}}(\lambda)$ has a sharp minimum as a function of λ for both smoothing-only cases. (At $\lambda = 0.63$ for the pain-score data and at $\lambda = 1.5$ for the galaxy data.) This is not true for the genuine SEF estimates. They allow the smoothing parameter λ to be chosen much larger without incurring too much EDEV penalty. In other words, *the SEF methodology allows us to oversmooth the density estimate*, with the advantages seen in Figure 4.

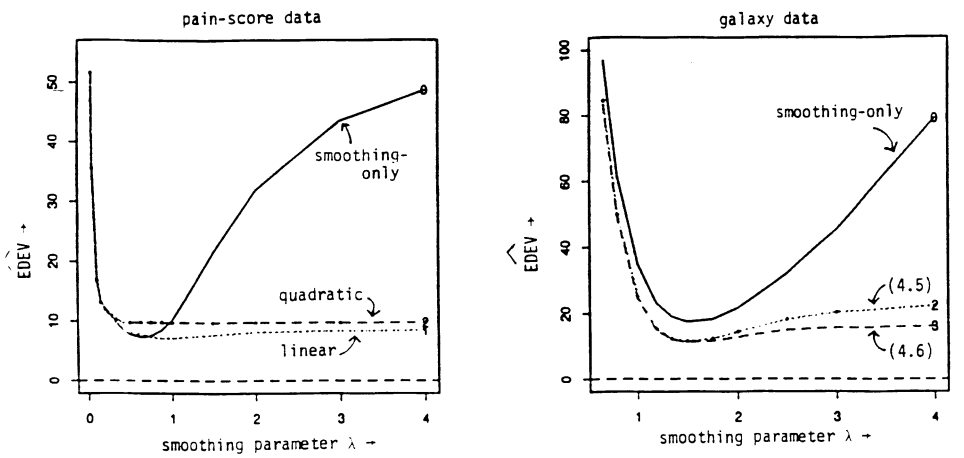

FIG. 5. *Expected deviance estimates* $\widehat{\text{EDev}}$, (6.22). *Left panel*: *SEF estimates for pain-score data* (2.13)–(2.15). *Right panel*: *SEF estimates for galaxy data as in Figure* 5. *In both cases the SEF estimates permit the use of larger smoothing parameters* λ, *compared to the smoothing-only estimates*.

In order to motivate $\widehat{\text{EDEV}}$, it helps to begin the discussion with the normal case. Suppose that the statistician observes a K-dimensional normal vector

$$\mathbf{z} \sim N_K(\mu, I) \tag{6.1}$$

and wishes to estimate μ using some linear estimator

$$\hat{\boldsymbol{\mu}} = S\mathbf{z}, \tag{6.2}$$

S being a $K \times K$ matrix. Model selection amounts to making a good choice of S.

Let err be the observed total residual squared error of $\hat{\boldsymbol{\mu}}$:

$$\text{err} = \|\mathbf{z} - \hat{\boldsymbol{\mu}}\|^2. \tag{6.3}$$

Also define

$$\boldsymbol{\nu} = E\{\hat{\boldsymbol{\mu}}\}, \qquad \text{DF} = \operatorname{tr} S \quad \text{and} \quad \text{TV} = \operatorname{tr} SS', \tag{6.4}$$

where DF stands for *degrees of freedom* and TV stands for *total variance*. Total variance TV equals $\sum_k \operatorname{var}(\hat{\mu}_k)$, and if S is a projection matrix, then $\operatorname{tr} S$ is the usual degrees of freedom. It is easy to prove the following two relationships:

$$E\{\text{err} - (K - 2\text{DF})\} = E\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} \tag{6.5a}$$

and

$$E\{\text{err} - (K - 2\text{DF} + \text{TV})\} = \|\boldsymbol{\nu} - \boldsymbol{\mu}\|^2. \tag{6.5b}$$

Both (6.5a) and (6.5b) play an important role in normal-theory model selection. The first of these is essentially the C_p or AIC criterion. Its extension to nonlinear estimators is Stein's unbiased risk estimate. Extensions of formula (6.5b) are used in hypothesis testing. The quantity $K - 2\text{DF} + \text{TV}$ equals the *residual degrees of freedom*:

$$\text{RDF} = \operatorname{tr}(I - S)(I - S)'. \tag{6.6}$$

If S is a projection matrix, then (6.5b) can be improved to

$$\text{err} \sim \chi^2_{\text{RDF}}(\|\boldsymbol{\nu} - \boldsymbol{\mu}\|^2), \tag{6.7}$$

where the notation indicates a noncentral chi-square variate with noncentrality parameter $\|\boldsymbol{\nu} - \boldsymbol{\mu}\|^2$. We test the adequacy of the estimate $\hat{\boldsymbol{\mu}} = S\mathbf{z}$ by comparing err to a central chi-square distribution χ^2_{RDF}.

We now return to the Poisson situation where we observe

$$\mathbf{s} \sim \text{Po}_K(\boldsymbol{\mu}) \tag{6.8}$$

and estimate $\boldsymbol{\mu}$ by some estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\mathbf{s})$, not necessarily of the SEF form (3.2), having expectation

$$\boldsymbol{\nu} \equiv E\{\hat{\boldsymbol{\mu}}(\mathbf{s})\}. \tag{6.9}$$

The *observed error* $\text{err} = \text{err}(\mathbf{s})$ in the Poisson context is

$$\text{err}(\mathbf{s}) = \text{Dev}(\mathbf{s}, \hat{\boldsymbol{\mu}}), \tag{6.10}$$

where Dev indicates the total Poisson deviance

$$\mathrm{Dev}(\boldsymbol{\mu}, \boldsymbol{\nu}) = 2\,\Sigma_k\{\log(\mu_k/\nu_k) - (\mu_k - \nu_k)\}.$$

The Poisson equivalents of K, DF and TV in (6.5) are

$$K(\boldsymbol{\mu}) = E\{\mathrm{Dev}(\mathbf{s}, \boldsymbol{\mu})\}, \qquad \mathrm{DF}(\boldsymbol{\mu}) = E\Big\{\sum_k (s_k - \mu_k)\log(\hat{\mu}_k)\Big\},$$
$$\mathrm{TRV}(\boldsymbol{\mu}) = 2\sum_k E\{\mu_k \log(\nu_k/\hat{\mu}_k)\}, \tag{6.11}$$

all expectations being with respect to $\mathbf{s} \sim \mathrm{Po}_K(\boldsymbol{\mu})$.

LEMMA 3. *In the Poisson situation* (6.8)–(6.11),

$$E\{\mathrm{err}(\mathbf{s}) - [K(\boldsymbol{\mu}) - 2\mathrm{DF}(\boldsymbol{\mu})]\} = E\{\mathrm{Dev}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\} \tag{6.12a}$$

and

$$E\{\mathrm{err}(\mathbf{s}) - [K(\boldsymbol{\mu}) - 2\mathrm{DF}(\boldsymbol{\mu}) + \mathrm{TRV}(\boldsymbol{\mu})]\} = \mathrm{Dev}(\boldsymbol{\mu}, \boldsymbol{\nu}). \tag{6.12b}$$

(The proof is given below.) Formulas (6.12a, b) are the Poisson versions of (6.5a, b).

In order to use Lemma 3, we can approximate $K(\boldsymbol{\mu})$, $\mathrm{DF}(\boldsymbol{\mu})$ and $\mathrm{TRV}(\boldsymbol{\mu})$ by their plug-in estimates $K(\hat{\boldsymbol{\mu}})$, $\mathrm{DF}(\hat{\boldsymbol{\mu}})$ and $\mathrm{TRV}(\hat{\boldsymbol{\mu}})$. Using bootstrap notation, let $\mathbf{s}^*$ given $\mathbf{s}$ have Poisson distribution

$$\mathbf{s}^*|\mathbf{s} \sim \mathrm{Po}_K(\hat{\boldsymbol{\mu}}(\mathbf{s})) \tag{6.13}$$

and let E_* indicate expectations with respect to (6.13), with $\mathbf{s}$ and $\hat{\boldsymbol{\mu}}(\mathbf{s})$ fixed. Then

$$\begin{aligned} K(\hat{\boldsymbol{\mu}}) &= \sum_k E_*\{\mathrm{Dev}(s_k^*, \hat{\mu}_k)\} \\ &= \sum_k E_*\{2[s_k^* \log(s_k^*/\hat{\mu}_k) - (s_k^* - \hat{\mu}_k)]\}. \end{aligned} \tag{6.14}$$

It is easy to evaluate (6.14) numerically since it is the sum of univariate Poisson deviances, the $\hat{\mu}_k$ being just fixed constants in the E_* expectations.

Using this same notation we can write the degrees of freedom estimate $\mathrm{DF}(\hat{\boldsymbol{\mu}})$ as

$$\mathrm{DF}(\hat{\boldsymbol{\mu}}) = E_*\Big\{\sum_k (s_k^* - \hat{\mu}_k)\log\hat{\mu}_k^*\Big\} = E_*(\mathbf{s}^* - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\eta}}^*, \tag{6.15}$$

where $\hat{\boldsymbol{\eta}}^* = \log(\hat{\boldsymbol{\mu}}^*) = \log(\hat{\boldsymbol{\mu}}(\mathbf{s}^*))$, the vector with kth component $\log\hat{\mu}_k^*$. Since by (5.6), $d\hat{\boldsymbol{\eta}}/d\mathbf{s} = \hat{O}$, (6.15) has the Taylor series approximation

$$\mathrm{DF}(\hat{\boldsymbol{\mu}}) \doteq E_*(\mathbf{s}^* - \hat{\boldsymbol{\mu}})'\hat{O}(s^* - \hat{\boldsymbol{\mu}}) = \mathrm{tr}\,\hat{D}\hat{O} \equiv \widehat{\mathrm{DF}}. \tag{6.16}$$

See Remark K. An alternative estimate is $\overline{\mathrm{DF}} = \mathrm{tr}\,\overline{D}\hat{O}$ as in (5.8).

The quadratic expansion $\log(\hat\mu/\nu) \doteq (\hat\mu/\nu - 1) - \frac{1}{2}(\hat\mu/\nu - 1)^2$ gives

$$E\left\{\mu_k \log\left(\frac{\nu_k}{\hat\mu_k}\right)\right\} \doteq \frac{\mu_k}{2}\frac{\operatorname{var}(\hat\mu_k)}{\nu_k^2} \doteq \frac{\operatorname{var}(\hat\mu_k)}{2\mu_k}, \tag{6.17}$$

permitting us to approximate (6.11) by

$$\operatorname{TRV}(\boldsymbol{\mu}) \doteq \sum_k \frac{\operatorname{var}(\hat\mu_k)}{\mu_k}. \tag{6.18}$$

The quantities $\overline{\mathrm{TRV}}$ and $\widehat{\mathrm{TRV}}$ in (5.8) are the obvious plug-in estimates for (6.18). The substitution of μ_k for ν_k in the denominator of (6.17) looks worrisome, but Remark K of Section 9 shows that the resulting error is asymptotically negligible.

The TRV estimates (5.8) can be written as

$$\overline{\mathrm{TRV}} = \operatorname{tr}\overline{D}\hat O'\hat D\hat O \quad \text{or} \quad \widehat{\mathrm{TRV}} = \operatorname{tr}\hat D\hat O'\hat D\hat O, \tag{6.19}$$

using (5.7), compared to the DF estimates

$$\overline{\mathrm{DF}} = \operatorname{tr}\overline{D}\hat O \quad \text{or} \quad \widehat{\mathrm{DF}} = \operatorname{tr}\hat D\hat O. \tag{6.20}$$

Notice the similarity to the normal-theory definitions in (6.4). Comparing (6.5b) with (6.12b), we can also define residual degrees of freedom RDF for the Poisson situation, estimated by

$$\overline{\mathrm{RDF}} = K(\hat{\boldsymbol{\mu}}) - 2\overline{\mathrm{DF}} + \overline{\mathrm{TRV}} \quad \text{or} \quad \widehat{\mathrm{RDF}} = K(\hat{\boldsymbol{\mu}}) - 2\widehat{\mathrm{DF}} + \widehat{\mathrm{TRV}}. \tag{6.21}$$

See Remark L. The quantity graphed in Figure 5 was the expected deviance estimate obtained from (6.12a), (6.14) and (6.20):

$$\widehat{\mathrm{EDEV}} = \operatorname{err}(\mathbf{s}) - K(\hat{\boldsymbol{\mu}}) + 2\widehat{\mathrm{DF}}. \tag{6.22}$$

The vertical scale in Figure 5 is misleading. For the galaxy data, reducing λ from 4 to 1.5 reduces $\widehat{\mathrm{EDEV}}$ for the quadratic SEF (4.5) from 22.2 to 11.5, a considerable amount. Is this significant? The observed deviance error (6.10), err, decreases by 45.3, while the residual degrees of freedom $\widehat{\mathrm{RDF}}$, (6.21), decreases by 29.3. This gives the naive chi-square significance value $\operatorname{prob}\{\chi^2_{29.3} > 45.3\} = 0.03$. A more trustworthy significance level could be obtained by Monte Carlo methods, bootstrapping with $\mathbf{s}^* \sim \operatorname{Prob}(\hat{\boldsymbol{\mu}})$, with $\hat{\boldsymbol{\mu}}$ the (1.5, 2) SEF vector.

The quantitative aspects of Figure 5 cannot be taken too literally. For instance, using $\overline{\mathrm{EDEV}}$ instead of $\widehat{\mathrm{EDEV}}$ moved the smoothing-only curve below the others for $\lambda < 2$ in the galaxy data. In general the careted (hat) formulas gave less erratic answers than the bar formulas, but there are really no strong reasons for preferring $\widehat{\mathrm{EDEV}}$. The fact is that it is usually difficult to estimate the performance of competing decision rules, and the SEF density estimates are no exception. Formulas like (5.9) and (6.12) help with model selection, but considerable subjectivity remains. The main point of Figure 5 is the qualitative one that the SEF estimates permit extensive oversmoothing.

PROOF OF LEMMA 3. Define $\mathscr{E}(\boldsymbol{\mu}) = E\{\operatorname{Dev}(\mathbf{s}^\dagger, \hat{\boldsymbol{\mu}})\}$, where $\mathbf{s}^\dagger \sim \operatorname{Po}_K(\boldsymbol{\mu})$ independent of $\mathbf{s}$, so $\mathscr{E}(\boldsymbol{\mu})$ is the expected prediction error of $\hat{\boldsymbol{\mu}}(\mathbf{s})$. Efron (1986)

shows that

$$\mathscr{E}(\boldsymbol{\mu}) = E\{\mathrm{err}(\mathbf{s}) + 2\mathrm{DF}(\mu)\}. \tag{6.23a}$$

We also have the identities

$$\mathscr{E}(\boldsymbol{\mu}) = K(\boldsymbol{\mu}) + E\{\mathrm{Dev}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\} \tag{6.23b}$$

and

$$E(\mathrm{Dev}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\} = \mathrm{Dev}(\boldsymbol{\mu}, \boldsymbol{\nu}) + \mathrm{TRV}(\boldsymbol{\mu}). \tag{6.23c}$$

All three of these relationships are easy to prove directly from the definition of the total Poisson deviance. Substituting (6.23b) into (6.23a) gives (6.12a). Substituting (6.23c) on the right side of (6.12a) gives (6.12b). □

7. Multisample problems. So far we have only considered one-sample problems. SEF estimates are particularly useful for investigating density differences in multisample situations. We use the exponential family model (1.2) for the different densities, with a shared carrier g_0 estimated nonparametrically, but with possibly different values of the exponential β parameters. An example will precede the theory.

Figure 6 concerns a two-sample application of SEF modelling. The data are the compliances of men in the Stanford arm of a randomized trial of the cholesterol-lowering drug Cholostyramine; see Efron and Feldman (1991). There were $n_1 = 172$ men in the control group and $n_2 = 165$ men in the treatment group. Compliance ran from 0 to 100%, so $\mathscr{Y} = [0, 100]$. A discretization $\mathscr{Y} = \bigcup_k \mathscr{Y}_k$ partitioned $\mathscr{Y}$ into $K = 46$ intervals of equal length. The left panel of Figure 6 shows the counts in the two groups. Compliance is significantly worse in the treatment group, as shown by standard two-sample tests.

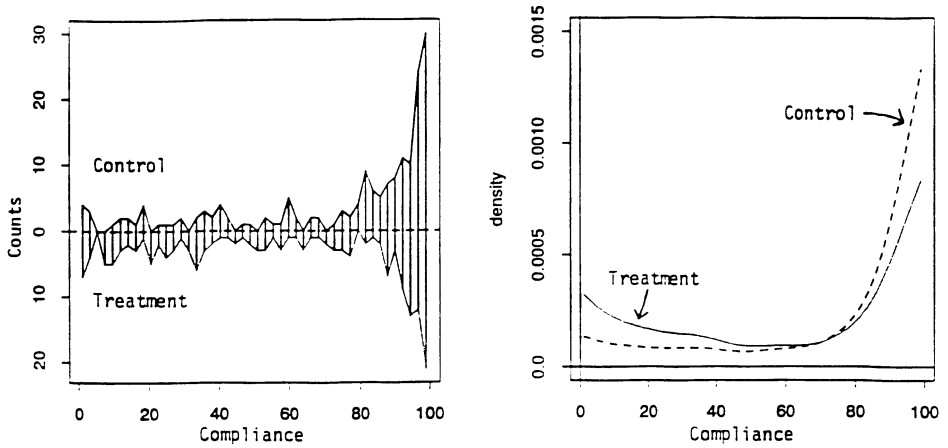

FIG. 6. *An application of SEF modelling to the Cholostyramine trial compliance data of Efron and Feldman* (1991). *Left panel*: *The count vectors for the control group and the treatment group*; $K = 46$ *equal divisions of* $\mathscr{Y} = [0, 100]$. *Right panel*: *SEF density estimates* (7.4) *for the two groups*; m *and* X *as in* (7.11). *The poorer compliance in the treatment group is graphically evident*.

The right panel of Figure 6 is the output of a two-sample SEF analysis. It neatly displays the compliance differences between the two groups in terms of their estimated densities. The ratio $\hat{g}_{\text{treat}}(y)/\hat{g}_{\text{cont}}(y)$ decreases almost linearly as y goes from 0 to 100%, but both densities are greatest at $y = 100\%$.

These densities were derived from a simple extension of the previous theory. In the multisample situation we observe independent random samples from L possibly different densities $g_1, g_2, \ldots, g_L$ on the same sample space $\mathcal{Y}$,

$$y_i(l) \overset{\text{i.i.d.}}{\sim} g_l(y) \quad \text{for } i = 1, 2, \ldots, n_l \text{ and } l = 1, 2, \ldots, L. \tag{7.1}$$

We discretize the problem as in Section 2, obtaining count vector $\mathbf{s}(l)$ for the lth sample,

$$s_k(l) = \#\{y_i(l) \in \mathcal{Y}_k\} \quad \text{for } k = 1, 2, \ldots, K \text{ and } l = 1, 2, \ldots, L. \tag{7.2}$$

The many-sample version of the Poisson regression model (3.1) is

$$\mathbf{s}(l) \overset{\text{ind.}}{\sim} \text{Po}_K(\boldsymbol{\mu}(l)) \quad \text{for } l = 1, 2, \ldots, L, \text{ with } \boldsymbol{\mu}(l) = \boldsymbol{\mu}^o e^{X\beta(l)}. \tag{7.3}$$

The SEF estimates corresponding to (3.2) are

$$\hat{\boldsymbol{\mu}}(l) = \hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}(l)}, \tag{7.4a}$$

where

$$\hat{\boldsymbol{\mu}}^o = m(\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(L)) \quad \text{and} \quad \hat{\beta}(l)\colon X'\left[\mathbf{s}(l) - \hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}(l)}\right] = 0. \tag{7.4b}$$

In this model, a common carrier $\hat{\boldsymbol{\mu}}^o$, estimated from all of the data $(\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(L))$ is modified by an exponential factor $e^{X\hat{\beta}(L)}$ which can vary with l.

The many-sample version of Lemma 1 in Section 3 is the following lemma:

LEMMA 4. *Let*

$$\hat{H}(l) = \frac{\partial \log(\hat{\boldsymbol{\mu}}^o)}{\partial \mathbf{s}(l)} \tag{7.5}$$

be the $K \times K$ matrix with kjth entry $\partial \log(\hat{\mu}_k^o)/ds_j(l)$. Then the $(p+1) \times K$ derivative matrix of $\hat{\beta}(l)$ with respect to $\mathbf{s}(j)$ is

$$\frac{\partial \hat{\beta}(l)}{\partial \mathbf{s}(j)} = \hat{G}(l)^{-1} Z'_{lj}, \tag{7.6a}$$

where

$$\hat{G}(l) = X'\hat{D}(l)X \quad \textit{and} \quad Z'_{lj} = X'_{lj}\left[\delta_{lj} I - \hat{D}(l)\hat{H}(j)\right], \tag{7.6b}$$

$\hat{D}(l) \equiv D(\hat{\boldsymbol{\mu}}(l))$, δ_{lj} equalling 1 or 0 as l does or does not equal j. If

$$\hat{\boldsymbol{\mu}}^o = M\mathbf{s}_+ \qquad \left(\mathbf{s}_+ \equiv \sum_l \mathbf{s}_l\right), \tag{7.7}$$

then

$$Z'_{lj} = X'_{lj}\left[\delta_{lj} I - D\left(e^{X\hat{\beta}(l)}\right) M\right]. \tag{7.8}$$

(The proof is nearly the same as for Lemma 1 and will not be given here.)

Lemma 4 leads to delta-method estimates of the covariance matrix for $\hat{\beta}(l)$, just as in (3.6),

$$\begin{aligned} \overline{\mathrm{Cov}}\left(\hat{\beta}(l)\right) &= \sum_{j=1}^{L} \hat{G}(l)^{-1}\left[Z'_{lj}\overline{D}(j)Z_{lj}\right]\hat{G}(l)^{-1} \quad \text{or} \\ \widehat{\mathrm{Cov}}\left(\hat{\beta}(l)\right) &= \sum_{j=1}^{L} \hat{G}(l)^{-1}\left[Z'_{lj}\hat{D}(j)Z_{jl}\right]\hat{G}(l)^{-1}, \end{aligned} \tag{7.9}$$

$\overline{D}(j) \equiv D(\mathbf{s}_j)$. We can also obtain covariance estimates for functions of the $\hat{\beta}(l)$. For example, if $\hat{\gamma} = \hat{\beta}(2) - \hat{\beta}(1)$, then

$$\begin{aligned} \frac{\partial\hat{\gamma}}{\partial\mathbf{s}(1)} &= \hat{G}(2)^{-1}Z'_{21} - \hat{G}(1)^{-1}Z'_{11}, \\ \frac{\partial\hat{\gamma}}{\partial\mathbf{s}(2)} &= \hat{G}(2)^{-1}Z'_{22} - \hat{G}(1)^{-1}Z'_{12} \end{aligned} \tag{7.10a}$$

and

$$\begin{aligned} \overline{\mathrm{Cov}}(\hat{\gamma}) &= \sum_{j=1}^{2}\left(\frac{\partial\hat{\gamma}}{\partial\mathbf{s}(j)}\right)\overline{D}(j)\left(\frac{\partial\hat{\gamma}}{\partial\mathbf{s}(j)}\right)', \\ \widehat{\mathrm{Cov}}(\hat{\gamma}) &= \sum_{j=1}^{2}\left(\frac{\partial\hat{\gamma}}{\partial\mathbf{s}(j)}\right)\hat{D}(j)\left(\frac{\partial\hat{\gamma}}{\partial\mathbf{s}(j)}\right)'. \end{aligned} \tag{7.10b}$$

SEF model (7.4) was used to estimate the two compliance densities in Figure 6. The kth row of X was quadratic in compliance,

$$x_k = \left(1, \tilde{y}_{(k)}, \tilde{y}^2_{(k)}\right), \tag{7.11a}$$

where $\tilde{y}_{(k)} = y_{(k)} - 50/100$, $y_{(k)}$ being the midpoint of $\mathcal{Y}_k$; $\hat{\boldsymbol{\mu}}^o = M\mathbf{s}_1$ as in (7.7), with

$$M = M(7), \tag{7.11b}$$

as in (2.14). The estimated difference $\hat{\gamma} = \hat{\beta}(\text{treatment}) - \hat{\beta}(\text{control})$ was

$$(0.26, -1.36, -0.44) \pm (0.22, 0.38, 1.51), \tag{7.12}$$

with the standard errors taken from $\widehat{\mathrm{Cov}}(\hat{\gamma})$, (7.10b). We see that the linear coefficient is significantly negative, t-value $= -3.6$, but the quadratic coefficient is not.

8. Other types of estimators. In the SEF methodology the application of an initial nonparametric smoother is followed by the fitting of an exponential family parametric model. Hjort and Glad's (1995) semiparametric density

estimator reverses this order, with the parametric model coming before the smoother. *Backfitting*, as applied to generalized linear models like (3.1), repeatedly iterates between the parametric and nonparametric fitting methods. This section discusses the SEF methodology in the context of these other possibilities.

As in Section 6, we begin with the normal case where the statistician wishes to estimate $\boldsymbol{\mu}$ having observed

$$\mathbf{z} \sim N_K(\boldsymbol{\mu}, I). \tag{8.1}$$

The linear model $\boldsymbol{\mu} = \boldsymbol{\xi} + X\beta$, with $\boldsymbol{\xi}$ a known *origin* vector and X a known $K \times (p+1)$ structure matrix, gives the estimate

$$\hat{\boldsymbol{\mu}} = \mathscr{P}(\mathbf{z}; \boldsymbol{\xi}) \equiv \boldsymbol{\xi} + P(\mathbf{z} - \boldsymbol{\xi}). \tag{8.2}$$

Here $P = X(X'X)^{-1}X'$ is the $K \times K$ projection matrix into $\mathscr{L}(X)$, the column space of X. In the normal case, $\mathscr{P}(\mathbf{z}; \boldsymbol{\xi})$ plays the role of the parametric exponential family model estimator. The role of the nonparametric smoother is played by

$$\mathscr{M}(\mathbf{z}; \boldsymbol{\xi}) = \boldsymbol{\xi} + M(\mathbf{z} - \boldsymbol{\xi}), \tag{8.3}$$

where M is some fixed $K \times K$ smoothing matrix such as $M(\lambda)$ in (2.13). Once again $\boldsymbol{\xi}$ represents a fixed and known origin. Often we take $\boldsymbol{\xi} = \mathbf{0}$, but it will be important here to consider more general choices of the origin.

The normal-theory analog of the SEF estimate (3.2) is

$$\begin{aligned} \hat{\mu}_{\text{sef}} &= \mathscr{P}(\mathbf{z}; \mathscr{M}(\mathbf{z}; \mathbf{0})) = (P + M - PM)\mathbf{z} \\ &= (P + QM)\mathbf{z}. \end{aligned} \tag{8.4}$$

Here $Q = I - P$ is the projection matrix into $\overset{\perp}{\mathscr{L}}(X)$, the orthocomplement to $\mathscr{L}(X)$. In other words, we begin with $\mathbf{0}$ as the origin, apply $\mathscr{M}$ to get an updated origin $\hat{\boldsymbol{\mu}}^o = \mathscr{M}(\mathbf{z}, \mathbf{0}) = M\mathbf{z}$ and finally take the estimate of $\boldsymbol{\mu}$ to be $\hat{\boldsymbol{\mu}} = \mathscr{P}(\mathbf{z}; \hat{\boldsymbol{\mu}}^o) = \hat{\boldsymbol{\mu}}^o + P(\mathbf{z} - \hat{\boldsymbol{\mu}}^o)$. Notice that

$$X'\mathbf{z} = X'\hat{\boldsymbol{\mu}} \tag{8.5}$$

according to (8.4), which says that $\mathbf{z}$ and $\hat{\boldsymbol{\mu}}$ have the same projection into $\mathscr{L}(X)$, namely, $P\mathbf{z}$. Equality (8.5) is the normal-theory analog of the moment-matching property (2.17).

Reversing the order of $\mathscr{P}$ and $\mathscr{M}$, as Hjort and Glad do in the density estimation problem, gives the estimator

$$\hat{\boldsymbol{\mu}}_{\text{HG}} = \mathscr{M}(\mathbf{z}; \mathscr{P}(\mathbf{z}; \mathbf{0})) = (P + MQ)\mathbf{z}. \tag{8.6}$$

This no longer enjoys the moment-matching property (8.5), but we can restore it with one further application of $\mathscr{P}$. This defines the *symmetrized estimator*

$$\hat{\boldsymbol{\mu}}_{\text{sym}} = \mathscr{P}(\mathbf{z}; \hat{\boldsymbol{\mu}}_{\text{HG}}) = (P + QMQ)\mathbf{z}. \tag{8.7}$$

If M is a symmetric matrix with eigenvalues between 0 and 1, then so is $P + QMQ$. This makes $\hat{\mu}_{\text{sym}}$ a formal Bayes estimator for $\boldsymbol{\mu}$ in situation (8.1), versus a normal prior distribution on $\boldsymbol{\mu}$ possibly having infinite variance in some directions.

The backfitting estimator $\hat{\boldsymbol{\mu}}_{\text{back}}$ is defined as follows in Chapter 5 of Hastie and Tibshirani (1990): suppose we can find vectors $\hat{\boldsymbol{\mu}}^o$ and $\hat{\boldsymbol{\mu}}^1$ such that

$$\hat{\boldsymbol{\mu}}^o = \mathscr{M}(\mathbf{z}; \hat{\boldsymbol{\mu}}^1) - \hat{\boldsymbol{\mu}}^1 \quad \text{and} \quad \hat{\boldsymbol{\mu}}^1 = \mathscr{P}(\mathbf{z}; \hat{\boldsymbol{\mu}}^o) - \hat{\boldsymbol{\mu}}^o. \tag{8.8a}$$

Then

$$\hat{\boldsymbol{\mu}}_{\text{back}} = \hat{\boldsymbol{\mu}}^o + \hat{\boldsymbol{\mu}}^1 \tag{8.8b}$$

[so $\hat{\boldsymbol{\mu}}_{\text{back}} = \mathscr{M}(\mathbf{z}; \hat{\boldsymbol{\mu}}^1) = P(\mathbf{z}; \hat{\boldsymbol{\mu}}^o)$.] Letting $\mathbf{v} = \mathbf{z} - \hat{\boldsymbol{\mu}}^1$, it is easy to show that

$$\mathbf{v} - M\mathbf{v} = (\mathbf{z} - \hat{\boldsymbol{\mu}}^1) - \hat{\boldsymbol{\mu}}^o \in \overset{\perp}{\mathscr{L}}(X), \tag{8.9}$$

which implies the moment-matching property (8.5). Hastie and Tibshirani show that $\hat{\boldsymbol{\mu}}_{\text{back}} = S_{\text{back}}\mathbf{z}$ for a matrix S_{back} that is symmetric and with eigenvalues in $[0, 1]$ if M has these same properties.

Further insight into the backfitting estimator can be gained by expressing it in an explicit form. The backfitting estimator satisfies

$$\hat{\beta} = (X'X)^{-1}X'(\mathbf{z} - \hat{\boldsymbol{\mu}}^0),$$
$$\hat{\boldsymbol{\mu}}^1 = M(\mathbf{z} - X\hat{\beta}).$$

Solving these equations yields the explicit expression

$$\hat{\beta} = [X'(I - M)X]^{-1}X'(I - M)\mathbf{z}.$$

This looks like a weighted least squares estimate with weight matrix $I - M$ or, equivalently, variance matrix $(I - M)^{-1}$. It has the same form as Hjort and Glad's estimator except that they use an unweighted least squares estimator. The weights $I - M$ can be justified from a mixed effects model in which $\boldsymbol{\mu}^1$ is a random effect.

We have discussed four linear estimators $\hat{\boldsymbol{\mu}} = S\mathbf{z}$, three of which have the moment-matching property (8.5). These three can be described as follows: the $\mathscr{L}(X)$ component of $\hat{\boldsymbol{\mu}}$ matches the $\mathscr{L}(X)$ component of $\mathbf{z}$; the $\overset{\perp}{\mathscr{L}}(X)$ component of $\hat{\boldsymbol{\mu}}$ equals the $\overset{\perp}{\mathscr{L}}(X)$ component of a point $\mathbf{v}$ in the flat space $\mathbf{z} \oplus \mathscr{L}(X)$; $\mathbf{v} = \mathbf{z}$ for $\hat{\mu}_{\text{sef}}$, $\mathbf{v} = Q\mathbf{z}$ for $\hat{\mu}_{\text{sym}}$ and $\mathbf{v}$ equals the point or points satisfying the orthogonality condition (8.9) for $\hat{\boldsymbol{\mu}}_{\text{back}}$. *Note*: All four estimators are the same if M and P commute, $MP = PM$.

We calculated the "equivalent kernels" for each of the four estimators, as in Figure 2.5 of Hastie and Tibshirani (1990). The calculation was done for the situation that produced the quadratic estimator in Figure 1: 40 equally spaced x values, P based on a matrix X with rows $(1, x, x^2)$ and M of the form (2.14). The smoothing parameter λ was chosen to give $\operatorname{tr} S = 5$ in each case. The SEF and backfitting kernels were remarkably similar, with both the HG and symmetrical kernels being slightly different.

We can define analogs to $\hat{\boldsymbol{\mu}}_{\text{sef}}$, $\hat{\boldsymbol{\mu}}_{\text{HG}}$, $\hat{\boldsymbol{\mu}}_{\text{sym}}$ and $\hat{\boldsymbol{\mu}}_{\text{back}}$ for the Poisson situation $\mathbf{s} \sim \text{Po}_K(\boldsymbol{\mu})$. Given a positive origin vector $\boldsymbol{\xi}$ and structure matrix X, we let the analog of (8.2) be

$$\mathscr{P}(\mathbf{s}; \boldsymbol{\xi}) \equiv D(\boldsymbol{\xi})e^{X\hat{\beta}(\xi)}, \quad \text{where } X'\left[\mathbf{s} - D(\boldsymbol{\xi})e^{X\hat{\beta}(\xi)}\right] = 0; \tag{8.10}$$

$\hat{\boldsymbol{\mu}} = \mathscr{P}(\mathbf{s}; \boldsymbol{\xi})$ is the MLE estimate of $\boldsymbol{\mu}$ in the "offset" generalized linear model $\boldsymbol{\mu} = D(\boldsymbol{\xi})e^{X\beta}$. The analog of (8.3) is

$$\mathscr{M}(\mathbf{s}; \boldsymbol{\xi}) = D(\boldsymbol{\xi})MD(1/\boldsymbol{\xi})\mathbf{s} = D(\boldsymbol{\xi})M(\mathbf{s}/\boldsymbol{\xi}); \tag{8.11}$$

$\hat{\boldsymbol{\mu}} = \mathscr{M}(\mathbf{s}; \boldsymbol{\xi})$ is a discretized version of what Hjort and Glad (1994) call a *nonparametric density estimate with a parametric start*, the "start" being the choice of $\boldsymbol{\xi}$.

The Poisson analogs of $\hat{\boldsymbol{\mu}}_{\text{sef}}$, $\hat{\boldsymbol{\mu}}_{\text{HG}}$ and $\hat{\boldsymbol{\mu}}_{\text{sym}}$ [(8.4)–(8.7)] are

$$\hat{\boldsymbol{\mu}}_{\text{sef}} = \mathscr{P}(\mathbf{s}; \mathscr{M}(\mathbf{s}; \mathbf{1})), \qquad \hat{\boldsymbol{\mu}}_{\text{HG}} = \mathscr{M}(\mathbf{s}; \mathscr{P}(\mathbf{s}; \mathbf{1})), \qquad \hat{\boldsymbol{\mu}}_{\text{sym}} = \mathscr{P}(\mathbf{s}; \hat{\boldsymbol{\mu}}_{\text{HG}}). \tag{8.12}$$

The backfitting estimate (8.8) is now defined by $\hat{\boldsymbol{\mu}}_{\text{back}} = \hat{\boldsymbol{\mu}}^o + \hat{\boldsymbol{\mu}}^1$, where $\hat{\boldsymbol{\mu}}^o$ and $\hat{\boldsymbol{\mu}}^1$ satisfy the fixed-point relationships

$$\hat{\boldsymbol{\mu}}^o = D(1/\hat{\boldsymbol{\mu}}^1)\mathscr{M}(\mathbf{s}; \hat{\boldsymbol{\mu}}^1) \quad \text{and} \quad \hat{\boldsymbol{\mu}}^1 = D(1/\hat{\boldsymbol{\mu}}^o)\mathscr{P}(\mathbf{s}; \hat{\boldsymbol{\mu}}^o). \tag{8.13}$$

Chapter 6 of Hastie and Tibshirani (1990) discusses an iterative algorithm for computing $\hat{\boldsymbol{\mu}}_{\text{back}}$.

The moment-matching property (2.17) is satisfied by $\hat{\boldsymbol{\mu}}_{\text{sef}}$, $\hat{\boldsymbol{\mu}}_{\text{sym}}$ and $\hat{\boldsymbol{\mu}}_{\text{back}}$. Each of these estimators can be written in the form $\hat{\boldsymbol{\mu}}^o e^{X\hat{\beta}}$, (3.26), with

$$\hat{\boldsymbol{\mu}}^o_{\text{sef}} = M\mathbf{s}, \qquad \hat{\boldsymbol{\mu}}^o_{\text{sym}} = \hat{\boldsymbol{\mu}}_{\text{HG}} \quad \text{and} \quad \hat{\boldsymbol{\mu}}^o_{\text{back}} = M(\mathbf{s}/\hat{\boldsymbol{\mu}}^1). \tag{8.14}$$

We have used $\hat{\boldsymbol{\mu}}_{\text{sef}}$ in all of our examples because it makes the computation of $\hat{H} = d\log(\hat{\boldsymbol{\mu}}^o)/d\mathbf{s}$, a crucial part of the formulas in Sections 2–7, so simple. [An even simpler choice, $\log(\hat{\boldsymbol{\mu}}^o) = H\mathbf{s}$ for some fixed matrix H, does not satisfy (5.14) and seems to have undesirable small-sample properties.]

In theory at least we can compute $\hat{H}$ for any choice of $\hat{\boldsymbol{\mu}}^o = m(\mathbf{s})$. Here, without proof, is $\hat{H}$ for $\hat{\boldsymbol{\mu}}^o_{\text{sym}}$:

LEMMA 5. *For* $\hat{\boldsymbol{\mu}}^o = \hat{\boldsymbol{\mu}}^o_{\text{sym}} = \mathscr{M}(\mathbf{s}; \mathscr{P}(\mathbf{s}, \mathbf{1}))$, *the derivative matrix* $\hat{H} = d\log(\hat{\boldsymbol{\mu}}^o)/d\mathbf{s}$ *is*

$$\hat{H} = D(\hat{\boldsymbol{\mu}}^{oo}/\hat{\boldsymbol{\mu}}^o)MD(1/\hat{\boldsymbol{\mu}}^{oo}) + \left[I - D(\hat{\boldsymbol{\mu}}^{oo}/\hat{\boldsymbol{\mu}}^o)MD(\mathbf{s}/\hat{\boldsymbol{\mu}}^{oo})\right]\hat{P}^{oo}, \tag{8.15a}$$

where

$$\hat{\boldsymbol{\mu}}^{oo} = \mathscr{P}(\mathbf{s}; \mathbf{1}) \quad \textit{and} \quad \hat{P}^{oo} = X\left[X'D(\hat{\boldsymbol{\mu}}^{oo})X\right]^{-1}X'. \tag{8.15b}$$

The symmetrized version of the $\hat{\boldsymbol{\mu}}_{\text{sef}}$ in Figure 3, $\hat{\mu}_{\text{sym}} = \mathscr{P}(\mathbf{s}; \mathscr{M}(\mathbf{s}; \hat{\boldsymbol{\mu}}_{\text{sef}}))$, gave similar but somewhat rougher contours than those in the left panel of Figure 4. There is no compelling theoretical reason for preferring $\hat{\boldsymbol{\mu}}_{\text{sym}}$ or $\hat{\boldsymbol{\mu}}_{\text{back}}$ to $\hat{\boldsymbol{\mu}}_{\text{sef}}$, though they seem closer in structure to Bayes and maximum likelihood estimators. In practice, the specific choices of M and X seem more crucial to successful estimation than does the choice between $\hat{\boldsymbol{\mu}}_{\text{sef}}$, $\hat{\boldsymbol{\mu}}_{\text{sym}}$ or $\hat{\boldsymbol{\mu}}_{\text{back}}$.

9. Remarks. The following remarks apply to the indicated sections.

REMARK A (Section 2). There is an interesting connection between moment-matching and function-preserving properties of smoothers. For a smoother $\hat{\mu} = M\mathbf{s}$, the condition $\Sigma\hat{\mu}_i = \Sigma s_i$ requires $\mathbf{1}'M = \mathbf{1}'$ (where $\mathbf{1}$ is a column of

ones), or equivalently $M'\mathbf{1} = \mathbf{1}$. Now most smoothers preserve constants so that $M\mathbf{1} = \mathbf{1}$. For symmetric M we see that matching the zeroth moment is the same as preserving the constant vector. However, most smoothers such as kernels are not symmetric, and hence will not match the zeroth moment exactly. Similarly, a smoother may preserve a vector $\mathbf{t}$ (so $M\mathbf{t} = \mathbf{t}$) without satisfying the moment-matching property $M'\mathbf{t} = \mathbf{t}$. In general, moment-matching is equivalent to function-preserving for the transpose of the smoother matrix.

REMARK B (Section 3). What is the true parameter β being estimated by $\hat{\beta}$? Suppose that $\mathbf{s} \sim \text{Po}_k(\boldsymbol{\mu})$, but that $\boldsymbol{\mu}$ is not necessarily of the form $\boldsymbol{\mu}(\beta)$ in (3.1). From $\boldsymbol{\mu}$ we determine $\boldsymbol{\mu}^o = m(\boldsymbol{\mu})$ and then β the solution vector to the equations $X'[\boldsymbol{\mu} - \boldsymbol{\mu}^o e^{X\beta}] = 0$. Under reasonable conditions $\hat{\beta}$ will be an asymptotically normal estimate for β, in the usual manner of an MLE. For instance if $\boldsymbol{\mu} = n\boldsymbol{\pi}$, with $\boldsymbol{\pi}$ fixed and $n \to \infty$, and if the $\boldsymbol{\mu}^o$ estimate is homogeneous, $m(c\mathbf{s}) = cm(\mathbf{s})$, then it is easy to show that

$$\sqrt{n}\,(\hat{\beta} - \beta) \to N_{p+1}(\mathbf{0}, A^{-1}BA^{-1}), \tag{9.1}$$

where, with $Z' = X'[I - D(e^{X\beta})\, d\boldsymbol{\mu}^o/d\boldsymbol{\mu}]$,

$$A = \lim_{n\to\infty} X'D\left(\frac{\boldsymbol{\mu}^o e^{X\beta}}{n}\right)X \quad \text{and} \quad B = \lim_{n\to\infty} Z'D\left(\frac{\boldsymbol{\mu}}{n}\right)Z. \tag{9.2}$$

The bias of $\hat{\beta}$ as an estimate of β is of order $1/n$.

REMARK C (Section 3). In case (2.13), $\hat{\boldsymbol{\mu}}^o = M\mathbf{s}$, the matrix Z in (3.4a) is a function of $\hat{\beta}$ but not of $\hat{\boldsymbol{\mu}}^o$, say $Z = Z(\hat{\beta})$. If $d\mathbf{s}$ is a K vector orthogonal to the columns of $Z(\hat{\beta})$, then

$$d\hat{\beta} = [X'\hat{D}X]^{-1}Z'(\hat{\beta})\, d\mathbf{s} = 0. \tag{9.3}$$

This implies that the level surfaces of constant $\hat{\beta}$ value are $K - (p+1)$-dimensional flat subspaces in the K-dimensional $\mathbf{s}$ space. Flat level surfaces tend to make delta-method covariance estimates such as (3.6) more accurate.

REMARK D (Section 4). Here are the total Poisson deviances:

$$\text{Dev}(\mathbf{s}, \hat{\boldsymbol{\mu}}) = \sum_k 2\left\{s_k \log\left(\frac{s_k}{\hat{\mu}_k}\right) - (s_k - \hat{\mu}_k)\right\} \tag{9.4}$$

for four choices of $\hat{\boldsymbol{\mu}}$: $\hat{\boldsymbol{\mu}}^o = M(2)\mathbf{s}$; $\hat{\boldsymbol{\mu}}^1$ the SEF based on this $\hat{\boldsymbol{\mu}}^o$ and $X = (\mathbf{1}, \mathbf{i}, \mathbf{j})$; $\hat{\boldsymbol{\mu}}^2$ the SEF (4.5); $\hat{\boldsymbol{\mu}}^3$ the SEF (9.3):

$\hat{\boldsymbol{\mu}}^0$	$\hat{\boldsymbol{\mu}}^1$	$\hat{\boldsymbol{\mu}}^2$	$\hat{\boldsymbol{\mu}}^3$
247.1	240.3	226.1	220.3.

(9.5)

The deviance decrease $D(\mathbf{s}, \hat{\boldsymbol{\mu}}^2) - D(\mathbf{s}\hat{\boldsymbol{\mu}}^3) = 5.77$ is much smaller than 3.75^2, the square of the corresponding t-value in Table 3. The naive t-value, calculated using the standard error estimate from (3.7), is the right one for approximating the deviance decrease. In this case the naive t-value is $0.000267/0.000111 = 2.74$, predicting $2.74^2 = 5.76$ for the deviance decrease.

REMARK E (Section 4). Results like (3.6) can be directly derived for continuous special exponential families (1.3) without going through the Poisson discretization argument. Suppose we estimate $g_0(y)$ in (1.2) by a continuous version of (2.13),

$$\hat{g}_0(y) = \frac{1}{n}\sum_{i=1}^{n} M(y|y_i), \tag{9.6}$$

where, for any y_i, $M(y|y_i)$ is a distribution over $\mathcal{Y}$ [not necessarily satisfying $\int_{\mathcal{Y}} M(y|y_i)\,dy = 1$]. We define the SEF density estimate to be $g_{\hat{\beta}}(y) = \hat{g}_0(y)\exp(\hat{\beta}_0 + t(y)\hat{\beta}_1)$, where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ satisfies the "maximum likelihood" equations

$$\int_{\mathcal{Y}} t(y)\, g_{\hat{\beta}}(y)\,dy = \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t(y_i), \tag{9.7}$$

as well as the constraint $\int_{\mathcal{Y}} g_{\hat{\beta}}(y)\,dy = 1$.

Define

$$z_i = \big(t(y_i) - \bar{t}\big) - \int_{\mathcal{Y}} (t(y) - \bar{t})\exp\big(\hat{\beta}_0 + (t(y) - \bar{t})\hat{\beta}_1\big) M(y|y_i)\,dy \tag{9.8}$$

and let $\widehat{\mathrm{Cov}}(t)$ indicate the covariance matrix of $t(y)$ for the distribution on $\mathcal{Y}$ corresponding to $g_{\hat{\beta}}(y)$. Then the continuous analog of (3.6a) is

$$\widehat{\mathrm{Cov}}\big(\hat{\beta}_1\big) = \frac{1}{n}\Big[\widehat{\mathrm{Cov}}(t)\Big]^{-1}\left[\sum_{i=1}^{n}\frac{z_i' z_i}{n}\right]\Big[\widehat{\mathrm{Cov}}(t)\Big]^{-1}. \tag{9.9}$$

(9.9) is the limit of (3.6a) as the discretization (2.1) becomes infinitely fine, after the superfluous parameter $\hat{\beta}_0$ is removed.

In order to use (9.9), we need to evaluate the integrals over $\mathcal{Y}$ involved in (9.7), (9.8) and $\widehat{\mathrm{Cov}}(t)$. The discretization argument effectively does such integrals by summation over $k = 1, 2, \ldots, K$. If $\mathcal{Y}$ is high dimensional, we might prefer to work in the continuous mode, doing the integrals by some more efficient algorithm such as componentwise Simpson rules.

REMARK F (Section 5). We will often be interested in the probability vector

$$\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\mu}}/\hat{\mu}_+ = \hat{\boldsymbol{\mu}}/n, \tag{9.10}$$

rather than in $\hat{\boldsymbol{\mu}}$ itself. Suppose that $\hat{\boldsymbol{\mu}}^o = m(\mathbf{s})$ is scale homogeneous,

$$m(c\mathbf{s}) = cm(\mathbf{s}) \qquad (c > 0). \tag{9.11}$$

The $\hat{\boldsymbol{\mu}} = \mathrm{SEF}(\mathbf{s}; m, X)$ is also scale homogeneous and $\hat{\pi}(c\mathbf{s}) = \hat{\pi}(\mathbf{s})$. Familiar homogeneity properties give

$$\mathbf{1}'\frac{d\hat{\boldsymbol{\mu}}}{d\mathbf{s}} = \mathbf{1}', \qquad \frac{d\hat{\boldsymbol{\mu}}}{d\mathbf{s}}\mathbf{s} = \hat{\boldsymbol{\mu}} \quad \text{and} \quad \frac{d\hat{\boldsymbol{\pi}}}{d\mathbf{s}} = \frac{1}{\hat{\mu}_+}(I - \hat{\boldsymbol{\pi}}\mathbf{1}')\frac{d\hat{\boldsymbol{\mu}}}{d\mathbf{s}}. \tag{9.12}$$

Using (9.12) it is not difficult to show that

$$\overline{\mathrm{Cov}}(\hat{\boldsymbol{\pi}}) \equiv \left(\frac{d\hat{\boldsymbol{\pi}}}{d\mathbf{s}}\right)\overline{D}\left(\frac{d\hat{\boldsymbol{\pi}}}{d\mathbf{s}}\right)' = \frac{1}{n^2}\left[\overline{\mathrm{Cov}}(\hat{\boldsymbol{\mu}}) - \frac{\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}'}{n}\right] \tag{9.13}$$

and that the diagonal elements of $\overline{\mathrm{Cov}}(\hat{\boldsymbol{\pi}})$ satisfy

$$\overline{\mathrm{trv}} \equiv n\sum_k \frac{\overline{\mathrm{var}}(\hat{\pi}_i)}{\hat{\pi}_k} \equiv \overline{\mathrm{TRV}} - 1. \tag{9.14}$$

Thus the comparisons in Table 4 remain valid for $\overline{\mathrm{trv}}$. The results for the equivalent of $\widehat{\mathrm{TRV}}$ are a little less neat.

REMARK G (Section 5). Let $\mathbf{p} = \mathbf{s}/n$, the vector of empirical probabilities, and define

$$\begin{aligned} &\bar{d} \equiv D(\mathbf{p}) = \overline{D}/n, \qquad \hat{d} \equiv D(\hat{\boldsymbol{\pi}}) = \hat{D}/n, \\ &\hat{p} \equiv X(X'\hat{d}X)^{-1}X' = n\hat{P}, \qquad \hat{q} \equiv \hat{d}^{-1} - \hat{p} = n\hat{Q}, \qquad \hat{h} \equiv n\hat{H}. \end{aligned} \tag{9.15}$$

[Under the homogeneity condition (5.14), $\hat{h} = d\log(\hat{\boldsymbol{\pi}}^o)/d\mathbf{p}$, for $\hat{\boldsymbol{\pi}}^o \equiv \hat{\boldsymbol{\mu}}^o/n$.] Then (5.9) can be written as

$$\overline{\mathrm{TRV}} = \mathrm{tr}\left[\bar{d}\hat{p} + \left(\hat{h}\bar{d}\hat{h}\right)'\left(\hat{d}\hat{q}\hat{d}\right]\right. \quad \text{or} \quad \widehat{\mathrm{TRV}} = (p+1) + \mathrm{tr}\left(\hat{h}\bar{d}\hat{h}\right)\left(\hat{d}\hat{q}\hat{d}\right) \tag{9.16}$$

not depending on n. This shows that $\overline{\mathrm{TRV}}$ and $\widehat{\mathrm{TRV}}$ are $O_p(1)$ as $n \to \infty$. See Remark K.

REMARK H (Section 5). Suppose that the discretization of $\mathscr{Y}$ becomes infinitely fine, with K going to infinity and the kth cell $\mathscr{Y}_k$ having volume Δ_k going to zero. Under sufficient regularity conditions, $\hat{\pi}(y_{(k)})/\Delta_k$ will approach $\hat{g}(y_{(k)})$, an estimate of the original continuous density with approximate variance

$$\overline{\mathrm{var}}\left\{\hat{g}(y_{(k)})\right\} = \frac{\overline{\mathrm{var}}\left\{\hat{\pi}(y_{(k)})\right\}}{\Delta_k^2}. \tag{9.17}$$

Then $\overline{\mathrm{trv}}$, (9.14), will approach

$$n\int_{\mathscr{Y}} \frac{\overline{\mathrm{var}}\{\hat{g}(y)\}}{\hat{g}(y)}\,dy = n\int_{\mathscr{Y}} \overline{\mathrm{CV}}(y)^2\hat{g}(y)\,dy, \tag{9.18}$$

with $\overline{\mathrm{CV}}(y) \equiv [\overline{\mathrm{Var}}(\hat{g}(y))]^{1/2}/\hat{g}(y)$, a coefficient of variation measure for $\hat{g}(y)$. From Remark G we see that $\overline{\mathrm{CV}}(y)$ is typically $O_p(1/\sqrt{n})$ as $n \to \infty$. The estimated average value for the coefficient of variation of SEF (4.5) was 0.15.

REMARK I (Section 5). A computationally more efficient expression than (5.9) is

$$\overline{\mathrm{TRV}} = \mathrm{tr}\,\overline{D}\left[\hat{P} + \hat{H}'\hat{D}\hat{H} - \hat{H}'\hat{D}\hat{P}\hat{D}\hat{H}\right]. \tag{9.19}$$

Each of the three terms in (9.19) can be evaluated using $O(K^2)$ multiplications rather than $O(K^3)$. This makes values of K as large as 1000 practical. Letting $\hat{G} = X'\hat{D}X$, the $O(K^2)$ computing formula is

$$\begin{aligned}\overline{\mathrm{TRV}} = {} & \operatorname{tr} \hat{G}^{-1}(X'\overline{D}X) + \|\overline{D}^{1/2}\hat{H}\hat{D}^{1/2}\|^2 \\ & + \operatorname{tr} \hat{G}^{-1}\big(X'\hat{D}\hat{H}\overline{D}^{1/2}\big)\big(\overline{D}^{1/2}\hat{H}'\hat{D}X\big)\end{aligned} \tag{9.20}$$

($\|a\|^2 \equiv \sum_i \sum_j a_{ij}^2$ for matrix a), and similarly for $\widehat{\mathrm{TRV}}$, substituting $\hat{D}$ for $\overline{D}$.

When $\hat{\boldsymbol{\mu}}^o = M\mathbf{s}$ the middle term in (9.20) equals $\sum_k \hat{\mu}_k A_k$, where

$$A_k = \sum_j \hat{M}_{kj}^2 \tilde{s}_j \Big/ \Big(\sum_j \tilde{M}_{kj} s_j\Big)^2 \qquad \big[\tilde{M}_{kj} \equiv (\hat{\mu}_k/\hat{\mu}_k^o) M_{kj}\big], \tag{9.21a}$$

and $\tilde{s}_j$ equals s_j for $\overline{\mathrm{TRV}}$ or $\hat{\mu}_j$ for $\widehat{\mathrm{TRV}}$. For the $\overline{\mathrm{TRV}}$ case $A_k \le 1$, but A_k can blow up for $\widehat{\mathrm{TRV}}$. The calculations of Section 6 replace A_k with

$$\min(A_k, 1) \tag{9.21b}$$

in the $\widehat{\mathrm{TRV}}$ case.

REMARK J (Section 6). The degrees of freedom formula (6.20) can be rewritten as

$$\overline{\mathrm{DF}} = \operatorname{tr} \overline{D}\big[\hat{P} + \hat{H} - \hat{P}\hat{D}\hat{H}\big] \quad \text{or} \quad \widehat{\mathrm{DF}} = (p+1) + \operatorname{tr} \hat{D}\big(\hat{H} - \hat{P}\hat{D}\hat{H}\big). \tag{9.22}$$

The $\hat{P}$ term corresponds to the $e^{X\hat{\beta}}$ part of the SEF definition (3.2), while the $\hat{H}$ term corresponds to the choice of the carrier $\hat{\boldsymbol{\mu}}^o$. The $\hat{P}\hat{D}\hat{H}$ subtraction term corrects for collinearity between the carrier and the exponential family terms. The degrees of freedom definition for $\hat{\boldsymbol{\pi}}$ rather than $\hat{\boldsymbol{\mu}}$, as in Remark F, subtracts 1 from formula (6.20) or (9.22). It takes $O(K^2)$ multiplications to compute $\overline{\mathrm{DF}}$ or $\widehat{\mathrm{DF}}$ from either formula.

REMARK K (Section 6). Write $\mathbf{s} = n\mathbf{p}$ as in Remark G and consider expression (6.15) for $\mathrm{DF}(\hat{\boldsymbol{\mu}})$ as $n \to \infty$ with $\mathbf{p}$ fixed. Assuming $m(\mathbf{s}) = cm(\mathbf{s})$, (9.11), the approximation $\widehat{\mathrm{DF}} = \hat{D}\hat{O}$ does not depend on n and so is $O_p(1)$ just as in (9.16). Using higher-order expansions it is easy to show that

$$\mathrm{DF}(\hat{\boldsymbol{\mu}}) - \widehat{\mathrm{DF}} = O_p(1/n), \tag{9.23}$$

so at least in this sense $\widehat{\mathrm{DF}}$ is a good approximation to $\mathrm{DF}(\hat{\boldsymbol{\mu}})$. The corresponding results hold for $\overline{\mathrm{DF}}$, $\overline{\mathrm{TRV}}$, and $\widehat{\mathrm{TRV}}$.

REMARK L (Section 6). The Poisson residual degrees of freedom formula

$$\mathrm{RDF}(\boldsymbol{\mu}) = K(\mu) - 2\mathrm{DF}(\mu) + \mathrm{TRV}(\boldsymbol{\mu}) \tag{9.24}$$

is always nonnegative. To prove this we use (6.12b) to write

$$\mathrm{RDF}(\boldsymbol{\mu}) = E\{\mathrm{Dev}(\mathbf{s}, \hat{\boldsymbol{\mu}})\} - \mathrm{Dev}(\boldsymbol{\mu}, \boldsymbol{\nu}), \tag{9.25}$$

and note that $E\{(\mathbf{s}, \hat{\boldsymbol{\mu}})\} = (\boldsymbol{\mu}, \boldsymbol{\nu})$. The proof is completed with the fact that the Poisson deviance $\mathrm{Dev}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is a jointly convex function of $(\boldsymbol{\mu}, \boldsymbol{\nu})$.

The RDF estimates (6.21) are not necessarily nonnegative. They become so if we replace $K(\boldsymbol{\mu})$ in (6.21) with the Taylor series estimates

$$\overline{K} = \operatorname{tr} \overline{D}\hat{D}^{-1} \quad \text{or} \quad \hat{K} = \operatorname{tr} \hat{D}\hat{D}^{-1} = (p+1), \tag{9.26}$$

but these were poor approximations in our examples.

REMARK M (Section 8). We tried a "loess" smoother [Cleveland and Gross (1992)] for the galaxy data, in the form of a generalized additive Poisson model [Hastie and Tibshirani (1990)]. We used a degree 2 loess model, that is, one that fits second degree polynomials locally. This gave an expected deviance picture similar to Figure 5. This is not surprising given the similarity between the moment-matching and function-preserving properties mentioned in Remark A.

REFERENCES

AITKEN, M. (1993). Model choice in single samples from the exponential and double exponential families using the posterior Bayes factor. Technical report, Tel-Aviv Univ., Israel.

CLEVELAND, W. C. and GROSS, E. (1992). Locally weighted regression. In *Statistical Models in S* (J. Chambers and T. Hastie, eds.). Wadsworth, Belmont, CA.

COX, D. and REID, N. (1987). Parametric orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

EFRON, B. (1986). How biased is the apparent error rate of a logistic regression? *J. Amer. Statist. Assoc.* **81** 461–470.

EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425.

EFRON, B. (1992). Six questions raised by the bootstrap. In *Bootstrap Proceedings Volume* (R. LaPage and L. Billard, eds.). Wiley, New York.

EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials (with comments and rejoinder). *J. Amer. Statist. Assoc.* **86** 9–25.

EFRON, B. and PETROSIAN, V. (1992). A simple test of independence for truncated data with applications to redshift surveys. *Astrophysical Journal* **399** 345–352.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, New York.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

HJORT, N. and GLAD, I. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** 882–904.

JONES, M. C. (1993). Kernel density estimation when the bandwidth is large. *Australian J. Statist.* **35** 319–326.

KOOPERBERG, C. and STONE, C. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

LEHMANN, E. (1983). *Theory of Point Estimation*. Wiley, New York.

LINDSEY, J. (1974a). Comparison of probability distributions. *J. Roy. Statist. Soc. Ser. B* **36** 38–47.

LINDSEY, J. (1974b). Construction and comparison of statistical models. *J. Roy. Statist. Soc. Ser. B* **36** 418–425.

LINDSEY, J. and MERSCH, G. (1992). Fitting and comparing probability distributions with log linear models. *Comput. Statist. Data Anal.* **13** 373–384.

LOH, E. and SPILLAR, E. (1986). Photometric redshift of galaxies. *Astrophysical Journal* **303** 154–161.

MARDIA, K. V., KENT, J. T. AND BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.

OLKIN, I. and SPIEGELMAN, C. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82** 858–865.

PIERCE, D. AND PETERS, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 701–737.

STONE, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–170.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA