The Annals of Statistics
1986, Vol. 14, No. 1, 124–137

EXPONENTIAL FAMILY MIXTURE MODELS (WITH LEAST-SQUARES ESTIMATORS)[1]

BY BRUCE G. LINDSAY

The Pennsylvania State University

For an arbitrary one parameter exponential family density it is shown how to construct a mixing distribution (prior) on the parameter in such a way that the resulting mixture distribution is a two (or more) parameter exponential family. Reweighted infinitely divisible distributions are shown to be the parametric mixing distributions for which this occurs. As an illustration conditions are given under which a parametric mixture of negative exponentials is in the exponential family. Properties of the posterior are given, including linearity of the posterior mean in the natural parameter. For the discrete case a class of simply-computed yet fully-efficient least-squares estimators is given. A Poisson example is used to demonstrate the strengths and weaknesses of the approach.

1. Introduction. The problem is this: suppose one is given an exponential family of densities for random variable X which have the form

$$f(x; \tau) = e^{\tau x - k(\tau)}, \qquad \tau \in \Omega, \tag{1.1}$$

with respect to an arbitrary sigma-finite measure μ on the real line. How can one then create a parametric family of mixing distributions (or equivalently, priors) on τ so that the resulting mixture density (the marginal for X) is still of exponential form? That is, we seek a parametric family $Q(\cdot; \alpha, \beta)$ of distributions on the natural parameter space Ω such that for some functions $t(x)$ and $k(\alpha, \beta)$ one has

$$f(x; \alpha, \beta) =_{\text{def}} \int f(x; \tau)\, dQ(\tau; \alpha, \beta) = e^{\alpha x + \beta t(x) - k(\alpha, \beta)}. \tag{1.2}$$

The solution to this problem is given in Theorem 2.1; the eligible class of functions $t(\cdot)$ are simply cumulant generating functions for infinitely divisible distributions. Section 2 of this paper provides a description of this class.

The posterior distributions corresponding to the priors of this paper will be shown to be the exponential family tilts of the corresponding infinitely-divisible distribution, where the *exponential family tilt* of the distribution $F(x)$ is the parametric family of distributions defined by

$$dF_\theta(x) = \frac{e^{\theta x}}{\int e^{\theta x}\, dF(x)}\, dF(x).$$

Although the main thrust of this paper is to develop some understanding of the structure given to the mixture problem by (1.2), attention will also be given to modelling and estimation in Sections 3 and 4. In particular, weighted least-

Received June 1984; revised September 1985.
[1]This research was supported by NSF Grant MCS80-03081.
AMS 1980 *subject classifications.* Primary 62F10; secondary 62F20.
Key words and phrases. Mixtures, random effects, exponential family, weighted least squares.

squares estimators of the parameters (α, β) in (1.2) will be given, together with a proof of their full efficiency in finite discrete models. Falling outside the domain of this paper are densities $f(x; \theta, \tau)$ involving other "unmixed" parameters θ. Lindsay (1985) demonstrates that the method of construction of this paper can be useful in such models.

Before proceeding to the solution, let us consider its ramifications, for which some background and context is essential. Mixture densities of the form $\int f(x; \tau)\, dQ(\tau)$ arise in a number of important settings. For example, suppose that for $i = 1$ to n each random variable X_i is an independent observation from a stratum i which has an associated parameter τ_i, but since the strata are sampled from a population of strata, the τ_i are themselves random variables from a distribution Q. In the usual normal linear models theory, this generates the one-way random-effects model, with Q generally being restricted to being normal itself. Another way the mixture model arises is as a natural model for overdispersion ("heavy-tails") relative to the basic density $f(x; \tau)$. This is perhaps most dramatically evidenced by the convex shape, as a function of x, of the logarithm of the ratio of the mixed density $\int f\, dQ$ to the basic density f, a property which will be utilized later in the paper. See Shaked (1980) and Schweder (1982) for further results regarding the relative dispersion of mixtures.

One motivation for considering new methods for generating mixing/prior distributions is the often awkward nature of the marginal distributions for X when the standard conjugate family is used. Although conjugate families have many attractive features in the Bayesian mode of inference [cf. Diaconis and Ylvisaker (1979)] even there it seems extremely limiting to have just one practical family of priors. Dalal and Hall (1983) consider discrete mixtures of conjugates as a method of increasing flexibility. Although the methods introduced here might also be useful in this regard, their main virtues are necessarily frequentist. From the Bayesian perspective perhaps the most interesting development is that this theory yields a class of priors for which the posterior mean of the natural parameter τ is linear in the prior's parameters [see (1.4)]. Another useful feature is that the posterior distributions are generally quite simple.

Consider the possible advantages of an exponential family mixture model in a frequentist setting. Suppose that one has a distribution Q so that (1.2) holds. Observe that $\beta = 0$ corresponds to the original unmixed model, with $\alpha = \tau$. It will be seen in the construction of Q that $\beta > 0$ will correspond to the presence of mixing, while $\beta < 0$ represents "underdispersion" relative to the basic model. Thus there will exist a uniformly-most-powerful unbiased test for the presence of the random effect $(\beta > 0)$ against the null hypothesis of no effect $(\beta = 0)$ based on the conditional distribution of $t(X)$ given $X = x$, or, with a random sample, $\Sigma t(X_i)$ given $\Sigma X_i = x$.

A second advantage to this family of mixing densities is that when there is a choice of the function $t(x)$ via selection of Q, there is flexibility in the choice of the mixed model. (It will be shown that the range of choices for $t(\cdot)$ are limited by an integrability constraint.) Furthermore, since the mixing effect shows up in the value of an observable, $t(x)$, simple graphical techniques can be useful both in the choice of $t(x)$ and considering the overall validity of the mixture model.

To confirm this last point, note that

$$\log[f(x;\alpha,\beta)/f(x,\tau)] = a + (\alpha - \tau)x + \beta t(x)$$

for $a = -k(\alpha,\beta) + k(\tau)$. Suppose the x-sample space is discrete. Denote by $n(x)$ the number of observations of x in a random sample of size n and by $\hat{\tau}$ the maximum likelihood estimator of τ in the basic density (1.1). Let $\hat{p}(x) = n(x)/n$. Then

$$(1.3)\quad r(x) =_{\text{def}} \log[\hat{p}(x)/f(x;\hat{\tau})] \to \text{ a.s. } a + bx + ct(x) \quad \text{as } n \to \infty,$$

for parametrically determined values a, b, c. For a sufficiently large sample size a graph of $(x, r(x))$ is thus diagnostic for the form of $t(\cdot)$. (A Poisson example will be given in Section 3.) In (1.3) we have used the maximum likelihood estimator of τ to standardize $r(x)$ as the graph then indicates departures in fit from the no-mixing model, with $r(x)$ near zero indicating a good fit at value x. In this regard note that $2\Sigma r(x)n(x)/n$ is the likelihood ratio goodness-of-fit statistic for testing the adequacy of the basic model containing no-mixing against an arbitrary multinomial alternative. More generally, with adequate data in a continuous model one possibility is to group the sample space into intervals, as in the chi-square goodness-of-fit test, and use a suitably redefined version of $(x, r(x))$ for diagnostics for the form of $t(x)$ and the validity of the mixture model.

Along this line, note that (1.3) suggests estimation of parameters by least squares. In Section 4 it is shown that fully efficient estimators can be so derived in discrete models.

A third useful feature of this class is the linearity of the posterior mean of the natural parameter. Lindsay (1985) has given a class of mixture problems where the linearity of the posterior mean yields a direct solution to an optimality problem. The linearity of the posterior mean and the form of the posterior variance follow from these easily derived relationships valid for exponential family mixtures:

$$(1.4)\qquad \begin{aligned} D_x \log \int f(x;\tau)\, dQ(\tau) &= E[\tau|x], \\ D_x^2 \log \int f(x;\tau)\, dQ(\tau) &= \operatorname{Var}[\tau|x]. \end{aligned}$$

These show that for density (1.2)

$$(1.5)\qquad \begin{aligned} E[\tau|x] &= \alpha + \beta t'(x), \\ \operatorname{Var}[\tau|x] &= \beta t''(x). \end{aligned}$$

In contrast with (1.5) under the conjugate prior it is the posterior mean of the mean value parameter $k'(\tau) = E_\tau[x]$ which is linear in the data x [see Diaconis and Ylvisaker (1979)].

Despite their virtues the methods of mixture modelling discussed herein are not statistical panaceas. For a start, numerical integration or summation will typically be necessary for maximum likelihood estimation. A second difficulty is that in a more complex model involving several parameters one cannot turn the sampling variation of one parameter τ into a "random effect" of the form

considered here unless the model fits rather narrow structural constraints. Thirdly, the eligible class of mixing distributions will generally change with the sample size n of each stratum. Finally, the priors are unconventional. Indeed, the most important aspect of the proposed models may not be their practicality but the insight gained by turning the presence of mixing into an observable phenomenon through the statistics $t(x)$.

2. The reweighted infinitely divisible densities. In this section the basic relationships between infinitely-divisible distributions and exponential family mixture models are established. First let P_β be a family of infinitely divisible distributions, with β a positive-valued parameter defined by the moment generating function (m.g.f.) relationship

$$\int e^{\tau x}\, dP_\beta(\tau) = e^{\beta t(x)}. \tag{2.1}$$

Note that $t(0) = 0$. We assume $t(x)$ is finite on the sample space of X. Important examples of this relationship are given by

Distribution	$t(x)$
Normal $(0, \beta)$	$x^2/2$
Poisson (β)	$\exp(x) - 1$
Gamma $(\beta, 1)$	$-\log(1 - x)$

For $k(\tau)$ defined from (1.1) by $k(\tau) = \log\int\exp(\tau x)\, d\mu(x)$, we define the k-reweighted distributions P_β^* by the relationship

$$dP_\beta^*(\tau) = c(\beta)e^{k(\tau)}\, dP_\beta(\tau), \tag{2.2}$$

where $c(\cdot)$ is the necessary standardizing function; we assume, of course, that the defining integral is finite, else the distribution is not defined. The following theorem establishes the main result.

THEOREM 2.1. *If dP_β^* is defined as in* (2.2), *then*

$$\int e^{\tau x - k(\tau)}\, dP_\beta^*(\tau) = c(\beta)e^{\beta t(x)}. \tag{2.3}$$

Conversely, suppose (2.3) *holds for some function $t(\cdot)$ on a set of x-values which has a point of accumulation and which includes* 0, *on an interval of β-values with left endpoint* 0. *Then* (2.2) *holds for a family of infinitely divisible distributions with* log m.g.f. $\beta(t(x) - t(0))$.

PROOF. Integrating the left side of (2.2) gives (2.3) directly. For the converse, we define from P_β^* a family of probability measures

$$d\tilde{P}_\beta(\tau) = d(\beta)e^{-k(\tau)}\, dP_\beta^*(\tau),$$

where $d(\beta)$ is the standardizing constant. Note that we are just standardizing by the inverse of the value of the density (2.3) at $x = 0$. Hence (2.3) may be

rewritten as

$$\int e^{\tau x}\, d\tilde{P}_\beta(\tau) = e^{\beta(t(x)-t(0))}. \tag{2.4}$$

If this relationship holds for a set of x-values with a point of accumulation, then the right-hand side of (2.4) completely determines the distribution $\tilde{P}_\beta$. It follows that the distribution of $\tilde{P}_\beta$ is the n-fold convolution of the distribution $\tilde{P}_{\beta/n}$. Hence $\tilde{P}_\beta$ is infinitely divisible and $\beta(t(x) - t(0))$ is the log m.g.f. for it. □

In the next proposition the families of infinitely divisible distributions are extended by several parameters.

PROPOSITION 2.2. (a) *Suppose that $\beta t(x)$ is the* log m.g.f. *of an infinitely divisible family of probability distributions P_β. Then*

$$t^*(x) = \alpha x + \beta[t(\sigma x + \theta) - t(\theta)]$$

is also the log m.g.f. *for a family P_β^* of infinitely divisible distributions, for arbitrary choice of real parameters α, σ, and θ, subject to $t(\theta) < \infty$. If the support of P_β is contained in a half infinite interval, then P_β^* has support shifted to the right by α.*

(b) *If $\beta_1 t_1(x), \ldots, \beta_k t_k(x)$ are each the* log m.g.f. *for a family of infinitely divisible distributions, then so is $\beta_1 t_1(x) + \cdots + \beta_k t_k(x)$.*

PROOF. First, we construct by exponential family tilt the density

$$dP_{\beta,\theta}(\tau) = c(\beta,\theta)e^{\theta\tau}\, dP_\beta(\tau) = e^{\theta\tau - \beta t(\theta)}\, dP_\beta(\tau). \tag{2.5}$$

This distribution has the m.g.f. in x:

$$e^{\beta(t(x+\theta)-t(\theta))}. \tag{2.6}$$

If Y has this m.g.f., then $\sigma Y + \alpha$ has the log m.g.f. $t^*(x)$ specified in the lemma, part (a).

Part (b) simply indicates the closure of infinitely divislbe distributions under convolutions. □

In Table 1 Proposition 2.2(a) is used to generate several classes of functions $t(\cdot)$. Using part (b) one can create multiparameter exponential family mixture models. In regard to computing the posterior distributions on τ, notice that the joint density of (τ, x) is proportional to

$$e^{\tau x - k(\tau)} e^{k(\tau)}\, dP_\beta(\tau)$$

so that for each β the family of distributions in parameter x is the exponential family tilt of the distribution P_β. If $\phi_\beta(s)$ is the moment generating function for dP_β, then the m.g.f. for the posterior given x is $\phi_\beta(s + x)/\phi_\beta(x)$. Since the prior is the expectation of the posterior over the marginal distribution of X, the priors corresponding to the families in Table 1 are mixtures of normals, Poissons, and gammas, respectively.

TABLE 1
Three important infinitely divisible distributions and the corresponding mixture structure

Infinitely divisible distribution	$t(x)$	Parameter space Ω must contain	Posterior	Posterior Mean	Posterior Variance
$N(\alpha, \sigma^2)$	$\alpha x + \sigma^2 x^2/2$	$(-\infty, \infty)$	$N(\alpha + \sigma^2 x, \sigma^2)$	$\alpha + \sigma^2 x$	σ^2
$\alpha + \sigma Y$ where $Y \sim \text{Po}(\beta)$	$\alpha x + \beta(e^{\sigma x} - 1)$	$\sigma R^+ + \alpha$	$\alpha + \sigma Y$ $Y \sim \text{Po}(\beta e^{\sigma x})$	$\alpha + \sigma\beta e^{\sigma x}$	$\sigma^2 \beta e^{\sigma x}$
$\alpha + \sigma Y$ where $Y \sim G(\beta, \lambda)$ and $\sigma = \pm 1$	$\alpha x + \beta \log \frac{\lambda}{(\lambda - \sigma x)}$ $(\sigma x < \lambda)$	$\sigma R^+ + \alpha$	$\alpha + \sigma Y$ $Y \sim G(\beta, \lambda - \sigma x)$	$\alpha + \frac{\sigma\beta}{\lambda - \sigma x}$	$\frac{\beta}{(\lambda - \sigma x)^2}$

Since all infinitely divisible distributions are the weak-convergence limits of convolutions of generalized Poisson distributions one might consider

$$t(x) = \alpha x + \sum \beta_j (e^{\sigma_j x} - 1) \tag{2.7}$$

to be a general format for the class of models. This is correct when the X-sample space is bounded. However, the constraint that $\int \exp(t(x))\, d\mu(x) < \infty$ otherwise provides some (possibly severe) limitations. Indeed, on an infinite sample space the possibility of generating an exponential family that models overdispersion may be considerably reduced by the integrability constraint. The following proposition, a direct application of Fubini's Theorem, gives us an alternative test for the eligibility of families $\{P_\beta\}$.

PROPOSITION 2.3.

$$\int e^{\beta t(x)}\, d\mu(x) < \infty \quad \textit{if and only if} \quad \int e^{k(\tau)}\, dP_\beta(\tau) < \infty.$$

EXAMPLES. From Proposition 2.3 for the normal $(\theta, 1)$ distribution one has the requirement on P_β that

$$\int \exp(\tau^2/2)\, dP_\beta(\tau) < \infty.$$

From this it is clear that the tails of $dP_\beta(\tau)$ must decline faster than $\exp(-\tau^2/2)$, and so no generalized Poisson will have a convergent integral. We note, however, that in some problems a truncation of the sample space may not be unreasonable, in which case there is no difficulty with convergence for the corresponding truncated exponential family.

For the negative exponential density, $\tau \exp(-\tau x)$, Proposition 2.3 yields the requirement that $\int \tau^{-1}\, dP_\beta(\tau) < \infty$. Since this will hold whenever there is no support at 0, correspondingly there is a rich class of exponential class mixture models, with an interesting contrast to the arbitrary mixture density.

PROPOSITION 2.4. *Let* $f(x;\tau) = \tau e^{-\tau x}$ *for* $x \geq 0$, $\tau > 0$. *Then if* $f(x)$ *is a density on* $x \geq 0$, *we have*

(a) $f(x)$ *is a mixture of* $f(x,\tau)$ *densities if and only if* $f(x)$ *is completely monotone.*

(b) *If* $f(x) = c(\beta)\exp(-\beta\psi(x))$ *for* $\beta > 0$ *is a mixture of* $f(x;\tau)$ *densities with* $f(0) < \infty$ *then* $\psi'(x) = D_x[-\log f(x)]$ (*the posterior mean of* τ) *is completely monotone.*

(c) *Conversely, if* $\psi(x)$ *has a completely monotone derivative* ψ' *on* $(0,\infty)$, *then there exists an exponential mixture density*

$$f(x;\alpha,\beta) = c(\alpha,\beta)\exp(-\alpha x - \beta\psi(x)) = \int f(x;\tau)\, dP^*_{\alpha,\beta}(\tau)$$

for $\alpha > 0$, $\beta \geq 0$.

PROOF. (a) Feller (1971), page 464, # 11.

(b) Follows from Feller (1971), page 450, Theorem 1. □

3. Exploratory analysis. As mentioned in Section 1, the form of the exponential family model suggests that an exploration of the mixture structure of a data set might potentially be conducted by analysis of the logarithmic residuals

$$r(x) = \log[\hat{p}(x)/f(x;\hat{\tau})].$$

In this section a data set is used to verify this idea, with emphasis on two attributes: (a) a graphical analysis illustrates important aspects of structure and (2) the problem of finding an appropriate mixture model is similar to that of finding an appropriate higher order term in a regression model. The example will also serve to illustrate several potential limitations on this simple approach which arise from the unbounded nature of the sample space.

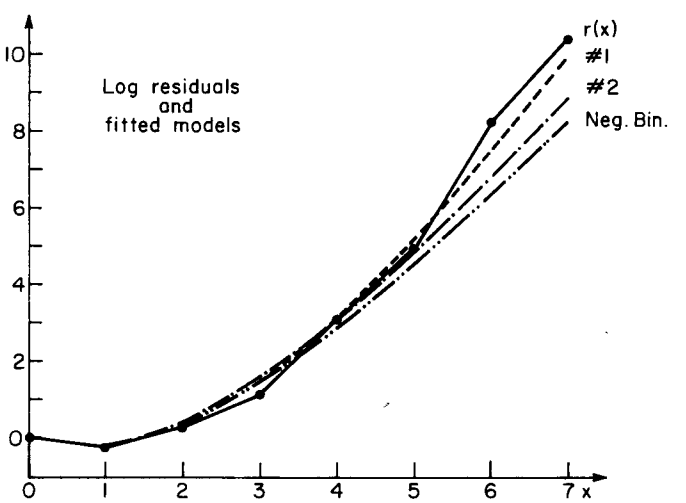

FIG. 1. *Graph of logarithmic residuals r(x) and three fitted models discussed in text.*

TABLE 2
Observed and estimated counts of the number of accidents for 9461 drivers

X	Observed count	$n\hat{f}_1$	$n\hat{f}_2$	Negative binomial
0	7840	7846.81	7852.20	7846.93
1	1317	1298.18	1265.93	1288.48
2	239	238.60	259.42	256.52
3	42	54.44	62.41	54.05
4	14	15.12	15.88	11.70
5	4	4.90	3.97	2.58
6	4	1.77	0.95	0.57
7	1	0.69	0.21	0.13
8 +	0	0.47	0.05	0.04
	9461			

A graphical presentation of the number of accidents in a year of driving by 9461 drivers in Belgium is given in Figure 1. The data appears in Table 2. The data was given by Thyrion (1961) and given further analysis by Seal (1971), Simar (1976), and Lambert and Tierney (1984). Assuming the number of accidents in a homogeneous population is Poisson, there is good reason to model the observed distribution as a mixture of Poissons. In Figure 1 the residuals $r(x)$ are plotted against x, revealing the convex shape characteristic of a mixed distribution [see (1.4)]. Also plotted on Figure 1 is the maximum likelihood fit of $r(x)$ by the negative binomial density, which is the mixture model arising from the conjugate mixing distribution. One can see that it appears to inadequately describe the tail behavior of the empirical density $\hat{p}(x)$.

The convergence criterion of Proposition 2.3 for the Poisson f with mean λ and natural parameter $\tau = \log \lambda$ is

$$\int e^{e^\tau} dP_\beta(\tau) < \infty. \tag{3.1}$$

This converges for every $dP_\beta(\tau)$ of the generalized Poisson form $\alpha - \sigma X$ for which σ is positive and X is Poisson (β). This implies that there exist exponential family mixture densities of the form

$$\log f(x; \alpha, \beta) = \alpha x + \beta e^{\sigma x} - k(\alpha, \beta) - \log x! \tag{3.2}$$

for σ positive.

The shape of the $r(x)$ graph suggests a reasonable fit could be obtained from a function $t(x)$ with decreasing second derivative and so class (3.2) seems promising. Figure 1 also shows the fit of exponential family Poisson mixtures for two choices of σ in (3.2)

$$\#1: \quad \log f_1(x) = \alpha x + \beta \exp(-0.25x) - k(\alpha, \beta) - \log x!,$$

$$\#2: \quad \log f_2(x) = \alpha x + \beta \exp(-0.50x) - k(\alpha, \beta) - \log x!.$$

(The choice of σ is a delicate issue which will be discussed in a remark at the end of this section.) From the parameter estimates (α, β) for the two exponential models one can directly compute estimates of the posterior mean $E[\log \mu|x]$ and variance $V(\log \mu|x)$ via (1.4).

The parameter estimates used to fit these exponential models were obtained by minimizing over (α, β, k) the weighted sum of squares

$$\sum \hat{p}(x)(r(x) - \alpha x - \beta t(x) + k)^2. \tag{3.3}$$

One nice feature of these estimators is that—unlike maximum likelihood estimation—one does not need to compute the summation constant $k(\alpha, \beta)$; one can estimate it and thereby eliminate the need for iterative methods. What is surprising is that in a discrete problem with finite support these estimates are fully efficient. This is shown in the next section. Of course, the term $k(\hat{\alpha}, \hat{\beta})$ must be computed terminally in order to correctly standardize the fitted density.

For comparison purposes Pearson's chi-square goodness-of-fit statistic for these models was computed by grouping all data from 5 on. The values for f_1, f_2, and the negative binomial were 3.38, 13.50, and 14.69, respectively. Even allowing a degree of freedom for the selection of σ, the density f_1 is a superior fit.

Of course, this Poisson example does not have a finite sample space, and the use of the least-squares method here illustrates some limitations in this approach. Since the value of magnitude factor $k(\alpha, \beta)$ is estimated, it is primarily the shape of the density being fit. There is no penalty for a poor fit in regions where $\hat{p}(x) = 0$. In the example, $\hat{p}(x) = 0$ for $x \geq 8$, and in particular, if one used model (3.2) with $\sigma = -0.10$, then the solution is inadequate. It gives a bimodal density with the larger mode lying beyond the range of the data.

Information about the appropriateness of the least-squares fit can be obtained by evaluating $\Delta = k(\hat{\alpha}, \hat{\beta}) - \hat{k}$. This is the shift on the logarithmic scale between the least-squares fit and the correctly standardized fit for each x. Since it will be shown in Proposition 4.4 that Δ is also the Kullback–Leibler information distance between the empirical density and the restandardized fit, it follows that if Δ is small, the observed data are sure to be close to the model. In this regard Proposition 4.4 implies that $2n\Delta$ would be the likelihood ratio goodness-of-fit statistic against the general multinomial alternative except that the least-squares estimators of (α, β) are substituted for the maximum likelihood estimators. Since $2n\Delta$ is therefore greater than the likelihood ratio statistic, it provides a conservative test of the fit of the model. In the above models #1 and #2 the differences Δ were quite small, being 0.0003 and 0.0009, respectively. The shift in log fit is negligible on the scale of Figure 1.

REMARK. When σ is treated as a free parameter, as was done implicitly above, the resulting family of distributions is no longer exponential family and the method of least squares becomes a nonlinear problem in the parameter σ. The method of least squares could still be applied to choose $\hat{\sigma}$, but in the given example the unbounded sample space makes this fail, as the solutions falls in the range where least squares gives a poor solution. Thus if the statistical user were

to need more than the above results, a more precise maximum likelihood solution is called for. We do note, however, what has been gained by the exploratory analysis: an idea of the data structure along with a good set of initial parameter estimates.

4. Least-squares estimators. In this section it is shown that certain least-squares estimators are fully efficient estimators for finite discrete exponential families. Two theorems are given; in the first, the standardizing constant for the density must be computed, in the other it is estimated. Conceptually, the first is more attractive, as it forces the correctly standardized density to lie near the empirical density; but the second least-squares estimator can be computed directly by ordinary weighted regression, and when the model fits the data well the standardization seems to be well estimated, as was seen above.

This section is not meant to be a complete study of the properties of least-squares estimators for discrete exponential families; indeed the theorems herein raise further questions which will be addressed elsewhere. The preliminary results are offered here merely to indicate that the exponential family models being considered are not as computationally unattractive as one might have supposed.

THEOREM 4.1. *Suppose that* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, $\mathbf{t}(x) = (t_1(x), \ldots, t_p(x))$, *and that*

$$f_{\boldsymbol{\theta}}(x) = \exp(\boldsymbol{\theta} \cdot \mathbf{t}(x) - k(\boldsymbol{\theta}))h(x)$$

is a discrete exponential family density with respect to counting measure on a finite support set $\{x_1, \ldots, x_s\}$ *on which the functions* $1, t_1(x), \ldots, t_p(x)$ *are linearly independent. Then the vector* $\hat{\boldsymbol{\theta}}$ *which minimizes*

$$R(\boldsymbol{\theta}) = \sum \hat{p}(x)\{\log[\hat{p}(x)/h(x)] - \boldsymbol{\theta} \cdot \mathbf{t}(x) + k(\boldsymbol{\theta})\}^2$$

is an asymptotically efficient estimator of $\boldsymbol{\theta}$.

PROOF. This is an element of a type of estimating equation considered by Rao (1961). Define the probability estimates for each x by $\hat{p}(x) = n(x)/n$. Then the estimating equation is of the form

$$\sum \hat{p}(x)(\log \hat{p}(x) - \log f_{\boldsymbol{\theta}}(x))(f_{\boldsymbol{\theta}}'/f_{\boldsymbol{\theta}}) = 0.$$

It is easily shown to satisfy Lemma 3 of Rao, and hence is first-order efficient. (Rao's lemma is for univariate θ, but is easily extended.) □

REMARK. We note that for $n = 1$, the solution is the maximum likelihood estimator. Moreover, one can apply the results of Rao concerning second-order efficiency to show that this estimator of $\boldsymbol{\theta}$ has the same second-order efficiency as the minimum chi-square estimator and the minimum Kullback–Leibler distance estimator.

THEOREM 4.2. *Under the same conditions as Theorem* 4.1 *the component* $\boldsymbol{\theta}$ *of the vector* $(\boldsymbol{\theta}, k)$ *which minimizes*

$$R^*(\boldsymbol{\theta}, k) = \sum_x \hat{p}(x)\{\log[\hat{p}(x)/h(x)] - \boldsymbol{\theta} \cdot \mathbf{t}(x) + k\}^2$$

is an asymptotically efficient estimator of $\boldsymbol{\theta}$.

PROOF. Although this could again be proved using Rao [see Remark (1) below], a direct proof offers some insight. Consider univariate θ. The solution in θ to the weighted least-squares problem may be obtained by first regressing out the constant term. For univariate θ this gives

$$\hat{\theta} = \frac{\sum \hat{p}(x)\{\log(\hat{p}(x)/h(x))\}(t(x) - \bar{t})}{\sum p(x)(t(x) - t)^2},$$

where $\bar{t} = \sum \hat{p}(x)t(x)$. Moreover,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\sqrt{n}\sum \hat{p}(x)\{\log \hat{p}(x) - \log f_{\theta_0}(x)\}(t(x) - \bar{t})}{\sum \hat{p}(x)(t(x) - \bar{t})^2}.$$

As $n \to \infty$, the denominator converges in probability to Var T. If we view the numerator of $\hat{\theta} - \theta_0$ as a function of $\mathbf{p}$, say $g(\mathbf{p})$ and use the delta method to find its limiting distribution about $\mathbf{p} = \mathbf{f}_{\theta_0}$, we obtain the result. □

By the delta method, $k(\hat{\theta})$ is a fully efficient estimator of $k(\theta)$. The following corollary shows that the least-squares estimator $\hat{k}$ is also.

COROLLARY 4.3. *The least-squares estimator* $\hat{k}$ *of the standardizing constant* $k(\theta_0)$ *is asymptotically efficient.*

PROOF. We write

$$\hat{k} = -\sum \hat{p}(x)\log[\hat{p}(x)/h(x)] + \hat{\theta}\bar{t}$$

and so

$$\hat{k} - k_0 = (\hat{\theta} - \theta_0)\bar{t} - \sum \hat{p}(x)\big(\log \hat{p}(x) - \log f_{\theta_0}(x)\big).$$

The second term has, when multiplied by $-2n$, an asymptotic $\chi^2(s-1)$ distribution, hence

$$\sqrt{n}(\hat{k} - k_0) = \sqrt{n}(\hat{\theta} - \theta_0)\bar{t} + o_p(1),$$

which implies the limiting distribution $N(0, E^2(T)/\text{Var}(T))$, thus achieving the required lower bound for the variance of an estimator of k. □

We define the Kullback–Leibler information distance $K(p, f)$ between two discrete densities by

$$K(p, f) = \sum_x p(x)\log[p(x)/f(x)].$$

The following proposition establishes some structural properties of the weighted least-squares solution with regard to this distance.

PROPOSITION 4.4. (i) *The least-squares estimators* $(\hat{\alpha}, \hat{\beta}, \hat{k})$ *satisfy*

$$\hat{k} \leq \log\left\{ \sum_{p(x)>0} h(x) e^{\hat{\alpha}x + \hat{\beta}t(x)} \right\} \leq k(\hat{\alpha}, \hat{\beta}).$$

(ii) *The Kullback–Leibler information distance between* $\hat{p}(x) = n(x)/n$ *and the density* $\hat{f} = f(x; \hat{\alpha}, \hat{\beta})$ *is*

$$K(\hat{p}, \hat{f}) = k(\hat{\alpha}, \hat{\beta}) - \hat{k}.$$

PROOF. Let $k^*(\hat{\alpha}, \hat{\beta})$ be the middle term in the above inequalities. We have from the least-squares equations (with intercept) that the weighted residuals sum to zero:

$$\sum \hat{p}(x)[\log \hat{p}(x) - \log \tilde{f}(x)] = 0, \tag{4.1}$$

where $\tilde{f}(x)$ is the fitted value using $\hat{k}$. However, the information inequality for probability mass functions implies

$$\sum \hat{p}(x)[\log \hat{p}(x) - \log f^*(x)] \geq 0, \tag{4.2}$$

where $f^*(x)$ is any probability mass function with support inclusive of that of $\hat{p}(x)$. In particular, $f^*(x) = \exp(\hat{k} - k^*(\hat{\alpha}, \hat{\beta}))\tilde{f}(x)$ is a density on the range $\{x: \hat{p}(x) > 0\}$. Using f^* in (4.2), together with (4.1), gives result (i) of the lemma. Use $\hat{f} = \exp(\hat{k} - k(\hat{\alpha}, \hat{\beta}))\tilde{f}(x)$ on the full range to get (ii). □

REMARKS. (1) One can generalize Theorem 4.2 to considering the problem of estimating θ in $f_\theta(x) = g_\theta(x)\exp(-k(\theta))$, where $g_\theta(x)$ is some positive function of θ and x and $\exp(k(\theta)) = \Sigma_x g_\theta(x)$. Then the problem

$$\min_{(k,\theta)} \sum \hat{p}(x)\big(\log \hat{p}(x) + k - \log g_\theta(x)\big)^2$$

becomes after minimization over k

$$\min_{\theta} \sum \hat{p}(x)\Big(\log \hat{p}(x) - \log g_\theta(x) - \sum \hat{p}(y)\log \hat{p}(y) + \sum \hat{p}(y)\log g_\theta(y)\Big)^2.$$

Again, using an adapted version of Rao's Lemma 3, one can demonstrate full first order efficiency for this functional of $\hat{\mathbf{p}}$.

(2) Although weighted least-squares estimators have been used in various guises in the statistical literature for discrete data analysis [e.g., Grizzle, Starmer, and Koch (1969)] to the author's knowledge this is the first explicit recognition that the standardizing constant can be treated as an unknown intercept. Its use seems to be implicitly advocated in Gabriel and Zamir (1979).

(3) A small computer experiment substantiated that the least-squares estimators can have reasonable efficacy in small samples. Samples of size 10 and 20 were taken from the binomial (2, 0.75) distribution. Expected counts of (0, 1, 2) were therefore (0.625, 3.75, 5.625) and (1.25, 7.5, 11.25), respectively. Since the

maximum likelihood estimator of θ: = $\log(p/1-p)$ is infinite on points of the sample space, and the least-squares estimator is undefined on some points we cannot precisely define mean square error here. However, in 362 Monte Carlo repetitions of the sample size 10 experiment (the trials terminated at the first "bad" sample) the sample mean square error of the m.l.e. was 0.292 and the mean square error of the least-squares estimator was 0.338. Note that the asymptotic variance for this problem is $0.26\bar{6}$. In the second Monte Carlo experiment, after 1150 iterations the same mean square errors were 0.139 and 0.130 for the m.l.e. and least-squares estimators, respectively (cf. asymptotic variance $0.13\bar{3}$).

5. Concluding remarks. Some useful comparisons can be made between the mixture dispersion models of this paper and the dispersion models used in the generalized linear models of McCullagh and Nelder (1983). In the univariate-x case the quasi-likelihood methods that they develop are exactly maximum likelihood methods when the underlying density has the form (McCullagh, 1983)

$$f(x;\theta,\sigma^2) = h(x)\exp\big(\sigma^{-2}(\theta x - b(\theta)) - c(\sigma^2, x)\big). \tag{5.1}$$

Here σ^2 represents a dispersion parameter, and for comparison purposes think of $\sigma^2 = 1$ as being the basic exponential family model. The practical advantages of this structure are extremely important: the mean of X and indeed the likelihood equations in the θ-parameter do not involve σ^2, so regression modelling in the mean value of X is straightforward. Moreover, since (as will be shown) σ^2 is effectively a sample size parameter, the model easily accommodates observations which are themselves means of varying sample sizes. The exponential families discussed earlier in this paper do not in general have these nice modelling properties, being directly applicable only to a random sample from one population.

On the other hand, the model given by (5.1) is only a single possible representation of dispersion. In fact it is one generated by convolutions rather than mixtures, as we now demonstrate.

The moment generating function for (5.1) in variable s is

$$\phi(s) = e^{\sigma^{-2}(b(\theta+s)-b(\theta))}. \tag{5.2}$$

For (5.2) to be a moment generating function for all $\sigma^2 > 0$ implies that it is infinitely divisible; it is in fact [as can be seen from (2.6)] the exponential family tilt in θ of an infinitely divisible distribution with convolution parameter $\beta = \sigma^{-2}$.

Thus one cannot in general create a model of the form (5.1) to represent dispersion about a baseline exponential model. However, (5.2) will represent a true moment generating function for all positive integer values of σ^{-2}; these are just the convolutions of the baseline density. Thus one could interpret σ^{-2} in (5.1) as being the unknown sample size of the sample from which the observation, the sample mean $\bar{x}$, was taken.

Acknowledgment. I wish to thank those involved in the editorial process for their helpful comments.

REFERENCES

DALAL, S. R. and HALL, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. Roy. Statist. Soc. Ser. B* **45** 278–286.

DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–282.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

GABRIEL, K. R. and ZAMIR, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21** 489–498.

GRIZZLE, J. E., STARMER, C. F. and KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25** 489–505.

LAMBERT, D. and TIERNEY, L. (1984). Asymptotic properties of maximum likelihood estimators in the mixed Poisson model. *Ann. Statist.* **12** 1388–1400.

LINDSAY, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13** 914–931.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 531–546.

SCHWEDER, T. (1982). On the dispersion of mixtures. *Scand. J. Statist.* **9** 165–169.

SEAL, H. L. (1969). *Stochastic Theory of a Risk Business*. Wiley, New York.

SHAKED, M. (1980). On mixtures from exponential families. *J. Roy. Statist. Soc. Ser. B* **42** 192–198.

SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209.

THYRION, P. (1961). Contribution a l'etude des bonus pour non sinistre en assurance automobile. *Astin Bull.* **1** 142–162.

DEPARTMENT OF STATISTICS
219 POND LAB
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802