A NOTE ON OVERDISPERSED EXPONENTIAL FAMILIES

BY

A. E. GELFAND and S. R. DALAL

TECHNICAL REPORT NO. 425
DECEMBER 5, 1989

PREPARED UNDER CONTRACT
N00014-89-J-1627 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

(4)

A NOTE ON OVERDISPERSED EXPONENTIAL FAMILIES

BY

A. E. GELFAND and S. R. DALAL

TECHNICAL REPORT NO. 425
DECEMBER 5, 1989

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A NOTE ON OVERDISPERSED EXPONENTIAL FAMILIES

A. E. Gelfand

S. R. Dalal

Summary

The issue of creating overdispersion in a given one parameter one dimensional exponential family, by extending it to a two parameter exponential family with the same support, is considered. An easily verifiable sufficient condition for this is derived. It is shown that a large class of families satisfy this condition and that this class includes Efron's (1986) and Lindsay's (1986) family as special cases. This class is also closely related to Jorgensen's (1988) Exponential Dispersion Models. UMP unbiased tests for testing overdispersion are exhibited and it is shown that in this context Cox's overdispersion test is a special case. A graphical display is developed to select a family of the general class which should be used with a given set of data to overdisperse a given target one parameter family. Two real data illustrations are given.

Keywords: Exponential family, overdispersion, exponential dispersion model, weighted least squares fitting, graphical displays for discrete models

1. Introduction

Overdispersion as an issue has been recognized by data analysts for many years. Samples are often found to be too heterogeneous to be explained by a one parameter family of models in the sense that the implicit mean-variance relationship in such a family is violated by the data; the sample variance is large compared with that predicted by inserting the sample mean into the mean-variance relationship. The natural remedy is to consider a larger collection of models, say a two parameter family. Historically the most frequently used means of doing this has been to mix the one parameter family with a two parameter family creating a two parameter marginal mixture family for the data. At the same value of the mean such mixing typically inflates the model variance. In fact Shaked (1980) shows that for a one parameter exponential family this is necessarily the case. Cox (1983) noted that for modest amounts of overdispersion a full specification of the mixing distribution was unnecessary; only its mean and variance (two parameters) are needed.

Does the assumption that the one parameter family is an exponential family, whence the mean-variance relationship (usually called the variance function) is immediately available through the normalizing function, enhance our ability to do modeling and inference for overdispersion? We argue that the answer is yes by developing a general class of two parameter exponential families which are overdispersed relative to a given one parameter exponential family. Customarily, the one-parameter exponential family has been mixed with a two parameter conjugate distribution (see e.g., Morris, 1982). The resulting overdispersed family of mixture models will often be awkward to work with since it will not be an exponential family. Hence straightforward optimal inference may be sacrificed. To model overdispersion Efron (1986) creates a so called double exponential family using Hoeffding's representation of a one

parameter exponential family through the Kullback-Leibler distance. In fact this family is a two parameter exponential family whose normalizing function has a very simple approximate form. Efron's arguments are all asymptotic relying upon the assumption that each observation is itself an average of a large number of observations. For any fixed sample size Efron's family is a special case of ours and therefore asymptotics are not needed to argue for overdispersion. Lindsay (1986) addresses the slightly different question of whether mixing a one parameter experimental family can produce a two parameter exponential family. He shows that using so-called reweighted infinitely divisible families as mixing distributions will achieve this. Again, our general two-parameter family includes Lindsay's. Within Jorgensen (1987) the one parameter exponential family is extended to a two parameter class of distributions which is called an exponential dispersion model (EDM). Extending our two parameter family to account for sample size yields a class of models having two interesting features. First this class is overdispersed relative to Jorgensen's two parameter EDM. Second this class is itself approximately an EDM.

In section 2 we offer our basic results. In section 3 we comment on sampling models by introducing sample size into our family. Finally in section 4 we develop a UMP unbiased test for overdispersion, propose a graphical display and illustrate our methods with two examples.

2. Basic Results

Consider the two-parameter exponential family $\mathcal{P}$, having densities, with respect to some σ-finite measure G, of the form

$$p_{\theta,\tau}(y) = \exp\{\theta y + \tau T(y) - \rho(\theta,\tau)\} \tag{2.1}$$

We write expectations under this density using the notation $E(S(y) \mid \theta,\tau)$. We presume that the

natural parameter space contains a two-dimensional rectangle which, by translation, can be taken to contain $\tau = 0$. We denote the one parameter model at $\tau = 0$ by

$$p_\theta(y) = \exp\{\theta y - \chi(\theta)\} \tag{2.2}$$

In (2.2), the "normalizing function" $\chi(\theta)$ is, in fact, the log moment generating function of G with $\chi'(\theta) \equiv \mu$ the mean of Y, and $\chi''(\theta)$ the variance. Since $\chi'(\cdot)$ is strictly increasing, it has a unique strictly increasing inverse function, say, $\theta = \eta(\mu)$.

In (2.1) let $\rho^{(r,s)} = \partial \rho^{r+s} / \partial \theta^r \partial \tau^s$, $r, s \geq 0$. We recall that $\rho^{(1,0)} = E(Y \mid \theta, \tau)$, $\rho^{(2,0)} = var(Y \mid \theta, \tau)$, $\rho^{(2,1)} = E(Y - E(Y))^2 (T(Y) - ET(Y)) \mid \theta, \tau)$ etc. In the sequel we suppose in (2.2) that corresponding to θ_0 the mean is μ_0. In (2.1) define $\theta_{\mu_0}(\tau)$ by the implicit function $E(Y \mid \theta, \tau) = \mu_0$ i.e., $\theta_{\mu_0}(\tau)$ indexes the curved subfamily of (2.1), $\mathcal{P}_{\mu_0}$, where the mean is μ_0. Implicit differentiation yields $\theta'_{\mu_0}(\tau) = -((\partial E(Y \mid \theta, \tau)/\partial \tau) / (\partial E(Y \mid \theta, \tau)/\partial \theta)) \mid_{(\theta_{\mu_0}(\tau), \tau)}$, i.e.,

$$\theta'_{\mu_0}(\tau) = -cov(Y, T \mid \theta_{\mu_0}(\tau), \tau) / var(Y \mid \theta_{\mu_0}(\tau), \tau) \tag{2.3}$$

Thus, by the association inequality if T is monotone $\theta_{\mu_0}(\tau)$ is.

As a definition the family of models (2.1) is said to be overdispersed relative to the family of models (2.2) if, keeping the means fixed in (2.1) the variance increases in τ. More precisely, we mean that for each subfamily, $\mathcal{P}_{\mu_0}$, $var(Y)$ increases in τ. Defining $g_{\mu_0}(\tau) = E(Y^2 \mid \theta_{\mu_0}(\tau), \tau)$, a sufficient condition for such overdispersion is $g_{\mu_0}'(\tau) \geq 0$ for $\tau \geq 0$. Underdispersion is defined by requiring $var(Y)$ decreasing in τ for each $\mathcal{P}_{\mu_0}$ with sufficient condition $g'_{\mu_0}(\tau) \leq 0$ for $\tau \geq 0$. Underdispersed models rarely occur in practice.

Lemma 2.1 provides a convenient analytic expression for $g_{\mu_0}'(\tau)$.

Lemma 2.1

$$g'_{\mu_0}(\tau) = cov\,(Y^2, T) - cov\,(Y, Y^2)\,cov\,(Y, T)\,/\,var\,(Y) \tag{2.4}$$

where expectations are taken at $(\theta_{\mu_0}(\tau),\tau)$ *in (2.4) and in the proof.*

Proof: Differentiating under the integral sign we have $g'_{\mu_0}(\tau) = \theta'_{\mu_0}(\tau)\,E\,(Y^3) + E\,(Y^2 T(Y)) - (d\,\rho(\theta_{\mu_0}(\tau), \tau)\,/\,d\,\tau)\,E\,(Y^2)$. Using (2.3) and the fact that $d\,\rho(\theta_{\mu_0}(\tau), \tau)\,/\,d\,\tau = \theta'_{\mu_0}(\tau)\,E\,(Y) + ET\,(Y)$ with simple manipulation yields (2.4).

We note that for arbitrary members of $\mathcal{P}$ the right hand side of (2.4) may be expressed in terms of derivatives of ρ:

$$cov\,(Y^2, T) - cov\,(Y, Y^2)\,cov\,(Y, T)\,/\,var\,(Y) = \rho^{(2,1)} - \rho^{(3,0)}\,(\rho^{(1,1)}\,/\,\rho^{(2,0)}) \tag{2.5}$$

We now state our basic result.

Theorem 2.1: *A sufficient condition that family (2.1) is overdispersed relative to family (2.2) is that* $T(y)$ *is convex.*

The proof of the result directly follows by Lemma 2.2 taken from Dalal, Kemperman and Mallows (1988).

Lemma 2.2: *Let* S_1 and S_2 *be both convex or both concave functions. Then for any random variable* Y

$$cov\,(S_1(Y), S_2(Y))\,var\,(Y) \geq cov\,(Y, S_1(Y))\,cov\,(Y, S_2(Y))$$

If Y *has support at more than 2 points, the inequality is strict provided either* S_1 *or* S_2 *is nonlinear.* Lemma 2.2 immediately provides a sufficient condition for (2.4) to be positive thus

proving the theorem.

An alternative proof of Theorem 2.1 arises from ideas contained in Shaked (1980). We may easily deduce the following modification of his Theorem 1.

Lemma 2.3: *Consider any pair of distinct densities f and g with respect to some dominating measure ν. If f/g is convex and $E_f(Y) = E_g(Y)$ then the number of sign changes for $f - g$ is two and the sequence is +,—,+.*

The conclusion of Lemma 2.3, again with $E_f(Y) = E_g(Y)$, implies that, provided expectations exist, $E_f(W(y) \geq E_g W(y)$ for all real convex W. (An elementary proof is given in Schweder (1982)). Thus by taking $W(y) = y^2$ we have $var_f(Y) \geq var_g(Y)$. Theorem 2.1 now follows by choosing any two members of $\mathcal{P}_{\mu_0}$, identifying as f the one with the larger τ. The convexity of $T(y)$ implies the convexity of f/g.

Indeed the convexity of T yields a somewhat stronger notion of overdispersion than our definition since, within $\mathcal{P}_{\mu_0}$, both of the proofs imply ordering by τ of expectations of an arbitrary convex function, i.e., for any arbitrary convex function S, $d/d\tau(ES(Y) \mid \theta_{\mu_0}(\tau), \tau) \geq 0$. This inequality allows for comparison of skewness, kurtosis, etc.

Writing Efron's (1986) family in the form (2.1) reveals $T(y) = y\,\eta(y) - \chi(\eta(y))$ ($\eta(y)$ is defined below (2.2)). Hence $T'(y) = \eta(y)$, a strictly increasing function, so his T is convex and his family is contained in (2.1). Lindsay (1986) shows that the family (2.1) arises by suitable mixing of (2.2) provided T is the log moment generating function of an infinitely divisible family of distributions, equivalently provided T' is absolutely monotone. If our interest is in overdispersion, we can allow the wider class of convex T's. This may serve to mitigate his concern (p.129) regarding finding T's such that $\exp(T(y))$ is integrable with respect to G. For

instance $T(y) = y \log y$ is convex but $T'(y)$ is not absolutely monotone on R^+. This T is used by Efron (1986) and in Section 4.4 below for the case when (2.2) is the Poisson family. Note that if $\exp(y^2)$ is integrable with respect to G we may argue that (2.1) with $t(y)=y^2$ approximates, for τ small, an arbitrary mixture of (2.2) provided the mixing distribution has finite second moment. This follows directly from Cox (1983,p.272).

The following corollary to Theorem 2.1 quantifies the relative overdispersion of (2.1) to (2.2) for τ small.

Corollary 2.1. *If in (2.1) T is convex and* τ is small, positive then

$$var(Y \mid \theta_{\mu_o}(\tau), \tau) / V(\mu_o) = 1 + a\,\tau + O(\tau^2) \tag{2.6}$$

where $a = g'_{\mu_0}(0) / V(\mu_0) > 0$ with $V(\mu) = (d\eta/d\mu)^{-1}$, the variance function associated with (2.2).

Proof. Write the numerator of the left hand side of (2.6) as $g_{\mu_0}(\tau) - \mu_0^2$ and expand in a Taylor series about $\tau = 0$.

We conclude this section by noting that in (2.1) the parameters τ and $\mu = E_{\theta,\tau}(Y)$ are orthogonal i.e. $E\,\partial^2 \log p_{\theta,\tau}(y) / \partial\mu\partial\tau = 0$, as can be verified by direct calculation . See Cox and Reid (1982) and Barndorff-Nielsen (1978, p.184) for further discussion.

3. Sampling Models and Asymptotics

To incorporate sample size into our models, suppose Z is the average over n independent replications of (2.2). Then by convolution the density of Z becomes

$$f_{\theta,n}(z) = \exp\{n(\theta z - \chi(\theta))\} \tag{3.1}$$

with respect to G_n, the corresponding convolution measure of G. That is, G_n has log moment generating function $n\chi(\theta)$. Treating n in (3.1)as a so-called dispersion parameter by allowing it to range over the subset of R^+ such that $n\chi(\theta)$ is a log moment generating function for some measure H_n, Jorgensen defines (3.1) to be an exponential dispersion model (EDM). Note that this dispersion parameter is not related to our notion of overdispersion. In fact we wish to formulate an overdispersed family of models relative to the EDM, (3.1).

Similar convolution of (2.1) does not produce an EDM. Here $\rho(\theta, \tau)$ is the log moment generating function of $\exp\{\tau T(y)\}d\,G(y)$ but $n\rho(\theta, \tau)$ will generally *not* be the log moment generating function of $\exp\{\tau T(y)\}d\,G_n(y)$. Rather the measure G_n will depend upon τ as well as n. Instead consider the extension of (3.1), paralleling (2.1), to the family of densities with respect to G_n of the form

$$f_{\theta,\tau,n}(z) = \exp\{n\theta z + m_n \tau T(z) - \rho_n(\theta,\tau)\} \tag{3.2}$$

where again T is convex, ρ_n is the normalizing function, and the sequence $m_n > 0$ is to be determined.

For fixed n hence m_n, Theorem 2.1 shows that on curved subfamilies of (3.2) where the mean, $n^{-1}\dfrac{\partial \rho_n}{\partial \theta}$, is held constant the variance will increase in τ. Thus if we consider independent replications of (3.2) with n constant (balanced samples), (3.2) serves as an overdispersed family of models for (3.1) regardless of how m_n is chosen. For unbalanced data extending (3.1) to (3.2) will lead to n varying over independent observations from (3.2). In this case we claim that the choice of m_n matters and that $m_n = n$ is the appropriate choice.

Sampling from (3.1) with varying n is interpreted as drawing averages based upon differing

sample sizes. Thus for (3.2) to suitably extend (3.1) the mean should be approximately constant over n. With regard to overdispersion consider the usual mixture model approach. If we mix (3.1) with some distribution H having mean μ_H and variance σ_H^2, the relative overdispersion of the resulting mixture distribution to (3.1) is $(n^{-1}E_H V(\mu) + \sigma_H^2) / n^{-1}V(\mu_H)$ which tends to ∞ as $n \to \infty$. That is, since the mixing distribution is assumed not to depend on n, taking additional observations within a population does not increase our knowledge regarding heterogeneity across populations. (An open extension of Lindsay's (1986) work is whether (3.1) can be mixed by a distribution free of n to produce a two-parameter exponential family. Our ensuing discussion suggests that the answer is no.)

We now argue roughly that regardless of the choice of m_n, the models (3.2) can not produce the "mixing type" of overdispersion relative to (3.1). They can achieve a limit for the relative overdispersion which is a constant > 1 and this occurs only when $m_n = n$. We note that such a limit arises in Efron's (1986) formulation. The discussion by Kent to Jorgenson (1987) alludes to this difference in "type" of overdispersion.

Suppose n is large with $m_n = n$. Expanding $\rho_n(\theta, \tau)$ about $\tau = 0$ we have for small τ

$$\rho_n(\theta, \tau) = \rho_n(\theta, 0) + \tau \frac{\partial \rho_n(\theta, \tau)}{\partial \tau} \Big|_{(\theta, 0)}$$

$$= n\,\chi(\theta) + \tau n E_n(T(Z) \mid Z \sim f_{\theta, n})$$

$$= n(\chi(\theta) + \tau T(\mu)) + 0(1) \tag{3.3}$$

where $\mu = \chi'(\theta)$. The last equality follows by expansion of $T(z)$ about μ. In fact we can take additional terms in the expansion of $\rho_n(\theta, \tau)$ and by similar argumentation eventually assert that $\rho_n(\theta, \tau)$ can be expressed in the form

$$\rho_n(\theta,\tau) = n\,\psi(\theta,\tau) + 0(1)\,. \tag{3.4}$$

Then for large n clearly $E_n(Z \mid \theta,\tau) = \dfrac{\partial\psi}{\partial\theta}$ so that the mean remains roughly constant across n. In fact for n large, under (3.4), (3.2) will be approximately an EDM and thus enjoy the same small dispersion asymptotics (see Jorgensen, 1987, p.135) as EDM's do. We note that Efron's (1986) double exponential family is of the form (3.2) with $m_n = n$ and (3.4) holding. Moreover we can extend our calculations of Section 2 to this case. The left hand side of (2.5) becomes $n^{-2}\delta(\theta,\tau) + o(n^{-2})$ where $\delta(\theta,\tau) = \psi^{(2,1)} - \psi^{(3,0)}\psi^{(1,1)} / \psi^{(2,0)}$. The relative overdispersion (2.6) becomes $1 + \dfrac{(n^{-2}\delta(\theta_0,0) + o(n^{-2}))\,n\tau}{n^{-1}\psi^{(2,0)}(\theta_0,0)} + O(\tau^2)$ which tends to a constant > 1. This argument generalizes Efron's Fact 2 (p. 711).

For a general sequence m_n, (3.3) becomes

$$n(\chi(\theta) + \frac{m_n}{n}\tau\,T(\mu)) + 0(1)\,. \tag{3.5}$$

If $m_n = o(n)$, (3.5) is $n\,\chi(\theta) + o(n)$ i.e. asymptotically (3.2) behaves like (3.1). If $n = o(m_n)$ from (3.5) we see that upon differentiation, the mean of (3.2) will not be stable over n. In fact it tends to ∞ as $n \to \infty$. Thus $m_n = n$ is the unique choice producing overdispersion of (3.4) relative to (3.1).

In Section 4 we confine ourselves to independent replications from (3.2) with n constant. More interesting problems would fit (3.2) allowing $n, \theta,$ and τ to vary across independent observations expressing, for the i^{th} observation, θ_i and τ_i through generalized linear models (see McCullagh and Nelder, 1983). Like Efron (1986), this could be incorporated in our setting.

4. Test and Displays for Overdispersion

4.1 UMP Unbiased Test

If $Y_1, \ldots, Y_k$ are an independent sample from (3.2) then standard theory gives a UMP unbiased test for overdispersion. That is, to test $H_0: \tau = 0$ vs. $H_A: \tau > 0$ we reject for

$$\Sigma T(Y_i) > c(\hat{\mu}) \tag{4.1}$$

where $\hat{\mu} = \Sigma Y_i / k$. Recall that the MLEs for μ, τ solve $\rho^{(1,0)} = \hat{\mu}$, $\rho_n^{(0,1)} = \Sigma T(Y_i)/k$. Since $\rho^{(0,1)}$ increases in τ we must have $\hat{\tau}$ increasing in $\Sigma T(Y_i)$ for fixed $\hat{\mu}$. We may write (4.1) as $\hat{\tau} > d(\hat{\mu})$. The concluding remark of section 2 shows that for k large $\hat{\mu}$ and $\hat{\tau}$ are approximately independent. Thus the unconditional test based upon the asymptotic normal distribution of $\hat{\tau}$ under H_0 will be approximately UMP unbiased. As Cox and Reid (1987, p.2) note, the asymptotic standard error of $\hat{\tau}$ is the same whether μ is known or not. In the special case $T(y) = y^2$, (4.1) can be written in the appealing form

$$\Sigma(Y_i - \bar{Y})^2 > c(\bar{Y}) \tag{4.2}$$

capturing our informal notion of overdispersion. In fact under the family (3.2), (4.2) is essentially Cox's (1983, p. 272) test for overdispersion.

4.2 Graphical Displays

We consider the case where (3.2) is a distribution on the nonnegative integers. Extension to continuous distributions could be similarly developed by partitioning the domain of Y into intervals. Our approach has its roots in the work of Gart (1969) and Ord (1970). These papers investigate standard overdispersion cases e.g. Beta Binomial to Binomial, Negative Binomial to Poisson, Binomial to Hypergeometric. Suppose then the class of models

$$p_{\theta,\tau}(y) = h(y)\exp\{\theta y + \tau T(y) - \rho(\theta,\tau)\},\quad y = 0,1,2,... \tag{4.3}$$

$p_{\theta,\tau}(y)$ a density with respect to counting measure. How might we develop a display to see the presence of overdispersion and to suggest a good T? We do not view this as an optimality problem. Allowing varying convex Ts in (4.3) would, for appropriate θ, τ, yield comparably fitting models. If $Y_1, \ldots, Y_k$ are a sample from (4.3) define $\hat{p}_y$ to be the observed proportion of Ys equal to y. Lindsay (1986) suggests examination of the log residual $r_y = \log(\hat{p}_y / p_{\hat{\theta},0}(y))$ where $\hat{\theta}$ is the MLE under (4.3) when $\tau = 0$. Suppose $\mu_0, \tau_0, \theta_{\mu_0}(\tau_0)$ are the true parameter values. Since $\hat{p}_y = p_{\theta_{\mu_0}(\tau_0),\tau_0}(y) + O_p(k^{-1/2})$ and $\hat{\theta} = \theta_{\mu_0}(0) + O_p(k^{-1/2})$ we can show that

$$r_y = \left\{\theta_{\mu_0}(\tau_0) - \theta_{\mu_0}(0)\right\} y + \tau_0 T(y) + \chi(\theta_{\mu_0}(0)) - \rho(\theta_{\mu_0}(\tau_0), \tau_0) + O_p(k^{-1/2}). \tag{4.4}$$

If we could remove the linear term in (4.4) we might more easily see whether $\tau > 0$ i.e. see the presence of overdispersion. Let

$$s_y = \log\{\hat{p}_{y+1}\, h(y) / (\hat{p}_y\, h(y+1))\}.$$

Then analogous to (4.4) we can show that

$$s_y = \theta_{\mu_0}(\tau_0) + \tau_0\{T(y+1) - T(y)\} + O_p(k^{-1/2}) \tag{4.5}$$

The linear term has been removed. Since, for T convex, $T(y+1) - T(y)$ increases in y, s_y should be increasing in y if overdispersion is present. An analogy here to unadjusted and adjusted (or partial) residual plots is noteworthy. A plot of r_y vs y corresponds to the former; a plot of s_y vs y to the latter. The latter display should be more effective in seeing overdispersion. Moreover since $T(y+1) - T(y)$ behaves like $T'(y)$ this plot allows us to readily see trends in T' and thus to suggest candidate T's. Note that because s_y is a function of two dependent random

variables (as is r_y) successful use of such displays will require k large and suitable truncation of y. Since our effort here is exploratory this should not cause concern.

4.3 Fitting the Model

Fitting of models (4.3) and the goodness of such fits is discussed extensively in Lindsay. In general ML estimation is unattractive because the *normalizing function is usually not* available in closed form. Weighted least squares is a straightforward alternative. We have investigated both $\ell_y = log\,(\hat{p}_y / h(y))$ and s_y (Lindsay uses r_y). Let $m+1$ be the smallest value of y such that $\hat{p} = 0$. For ℓ_y we minimize

$$\sum_{y=0}^{m} w_y \{\ell_y - (\theta_y + \tau T(y) + c)\}^2 \tag{4.6}$$

over θ, τ and c where $w_y = \hat{p}_y / (1 - \hat{p}_y)$. For s_y we minimize

$$\sum_{y=0}^{m-1} w_y \{s_y - (\theta + \tau (T(y+1) - T(y)))\}^2 \tag{4.7}$$

over θ and τ where $w_y = \hat{p}_{y+1}\hat{p}_y / (\hat{p}_{y+1} + \hat{p}_y)$. The weight w_y is (up to a constant) the reciprocal of the estimated variance of ℓ_y and s_y respectively. We ignored the covariances in the fitting on two grounds. First, Lindsay's theoretical work (Theorem 4.1) shows that the least squares estimates resulting from (4.6) are asymptotically efficient if the domain of (4.3) is bounded. Second, for the two data sets in section 4.4 the full covariance matrix amongst the ℓ_y or amongst the s_y can be estimated by the delta method. For both data sets for ℓ_y and for s_y the diagonal terms dominated the estimated inverse.

When using s_y, $\hat{\theta}$ and $\hat{\tau}$ immediately provide estimates of $p_{\theta,\tau}(y)$ in (4.3) up to the normalizing constant. This constant is then computed terminally to correctly standardize the

fitted cell probabilities. When using ℓ_y, $\hat{c}$ estimates the normalizing constant but in fact once $\hat{\theta}$ and $\hat{\tau}$ were obtained we ignored $\hat{c}$. Rather we again calculated c terminally to standardize.

Comparison between observed and fitted was done using Pearson's chi square statistic. When using ℓ_y we have $m+1$ cells with 3 parameters; when using s_y we have m cells with 2 parameters. Thus in either case $m-3$ degrees of freedom are associated with the goodness of fit statistic. Intuitively we might expect poorer fitting using the s_y. They are the log of a ratio of random variables and would thus be expected to be more variable than the ℓ_y. This is perhaps borne out for the second data set in section 4.4.

4.4 Two Examples

The data in Table 1 is taken from Sokal and Rohlf (1973, p.67) and has been examined by e.g. Shaked (1980). It consists of the frequency of males in 6115 sibships of size 12 in Saxony, 1876-85. Taking an initial binomial model the overall $\hat{p} = .519$, $n\hat{p}(1-\hat{p}) = 2.996$ while $S^2 = 3.490$ suggesting overdispersion. The UMPU test for overdispersion using (4.2) with a normal approximation is extremely significant. Figure 1 plots s_y vs y revealing the expected increasing pattern. In fact since Figure 1 reveals a roughly linear relationship $T(y) = y^2$ is suggested. Fitting using ℓ_y produced $\hat{\theta} = -.2585$, $\hat{\tau} = .0265$; the fitted probabilities are given under $p_y^{(1)}$ in Table 1. Fitting using s_y produced $\hat{\theta} = -.2463$, $\hat{\tau} = .0260$; the fitted probabilities are given under $p_y^{(2)}$ in Table 1. The fits are very close and both are excellent. For $p_y^{(1)}$ $\chi^2 = 15.41$, for $p_y^{(2)}$ $\chi^2 = 14.54$ with d.f. = 10.

The data in Table 2, originally collected by Thyrion (1961) is taken from Seal (1969) and has been analyzed by Lindsay (1986) and others. It consists of observed counts of accidents in a year for 9461 Belgian drivers. Taking an initial Poisson model $\hat{\lambda} = .0214$ with $S^2 = .0289$

suggesting overdispersion. Figure 2 plots s_y vs. y supporting this. In this situation integrable choices for $T(y)$ are limited. For example $T(y)=y^2$ or $T(y)=(y+1)\log(y+1)$ are not. $T_1(y)=e^{-y}$ was used by Lindsay. We use $T_2(y)=y\log y$ which arises from Efron's (1986) "double Poisson" example. Figure 2 suggests that T' is possibly concave which is satisfied by both T_1 and T_2. (Experimentation not presented shows that for T_2, (4.3) resembles, except in the far tails, a negative binomial distribution). Fitting using ℓ_y produced $\hat{\theta}=-1.833$, $\hat{\tau}=.7546$; the fitted probabilities are given under $p_y^{(1)}$ in Table 2. Fitting using s_y produced $\hat{\theta}=-1.792$, $\hat{\tau}=.6322$; and the fitted probabilities are given under $p_y^{(2)}$ in Table 2. The fits are similar. The goodness of fit test collapsing $y \geq 5$ has 3 d.f. For $p_y^{(1)}$ $\chi^2=25.92$, for $p_y^{(2)}$ $\chi^2=40.76$. While Lindsay's fits (p.131) appear to be better the comparison is unfair since he has really employed a 3 parameter model. In any event our discussion shows that Efron's family may not be adequate and that allowing more general convex T increases Lindsay's possibilities.

References

Bandorff-Nielsen, O. (1978). *Information and Exponential Families In Statistical Theory*. Wiley, Chichester.

Cox,D. R. (1983). "Some Remarks on Overdispersion," *Biometrika 70*, p. 269-274.

Cox, D. R. and Reid, N. (1987). "Parameter Orthogonality and Approximate Conditional Inference," *J. R. Statist. Soc.-B, 49*, p.1-39.

Dalal, S. R., Kemperman, J. H. B. and Mallows, C. L. (1988). "Some Moment inequalities Involving Generalized Convex Functions." Bellcore Technical Memorandum, Morristown, N.J.

Efron, B. (1986). "Double Experimental Families and Their Use in Generalized Linear Regression." *J. Am. Statist. Assoc. 81*, p. 709-721.

Gart, J. J. (1970). "Some Simple Graphically Oriented Statistical Methods for Discrete Data," *Random Counts in Models and Structure*, G. P. Patil, ed., Penn State Univ. Press.

Jorgensen, B. (1987). "Exponential Dispersion Models," *J. R. Statist. Soc.-B 49*, p.127-162.

Lindsay, B. (1986). "Exponential Family Mixture Models (with Least Squares Estimators)," *Ann. of Statist., 14* p. 124-137.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, Chapman Hall, London.

Ord, J.K. (1967). "Graphical Methods for a Class of Discrete Distributions," *J. R. Statist. Soc.-A 130*, p.232-238.

Schweder, T. (1982). "On The Dispersion of Mixtures," *Scand. J. of Statist.*, 9. p. 165-169.

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

Shaked, M. (1980). "On Mixtures from Exponential Families," *J. R. Statist. Soc.-B 42*, p. 192-198.

Sokal and Rohlf (1973). *Introduction to Biostatistics*. Freeman, San Francisco.

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

Table 1

Sibship data (Sokal and Rohlf, 1973) with fitted probabilities

y	Observed Counts	f_y	$p_y^{(1)}$	$p_y^{(2)}$
0	3	0.0005	0.0004	0.0004
1	24	0.0039	0.0037	0.0038
2	104	0.0170	0.0171	0.0177
3	286	0.0468	0.0508	0.0520
4	670	0.1096	0.1073	0.1088
5	1033	0.1689	0.1696	0.1706
6	1343	0.2196	0.2058	0.2057
7	1112	0.1818	0.1934	0.1922
8	829	0.1356	0.1395	0.1380
9	478	0.0782	0.0754	0.0743
10	181	0.0296	0.0290	0.0285
11	45	0.0074	0.0071	0.0070
12	7	0.0011	0.0008	0.0008

Table 2

Accident data (Seal, 1969) with fitted probabilities

y	Observed Counts	f_y	$p_y^{(1)}$	$p_y^{(2)}$
0	7840	0.8287	0.8286	0.8282
1	1317	0.1392	0.1325	0.1380
2	239	0.0253	0.0302	0.0276
3	42	0.0044	0.0068	0.0051
4	14	0.0015	0.0015	0.0009
5	4	0.0004	0.0003	0.0001
6	4	0.0004	0.0001	0.0
7	1	0.0001	0.0	0.0

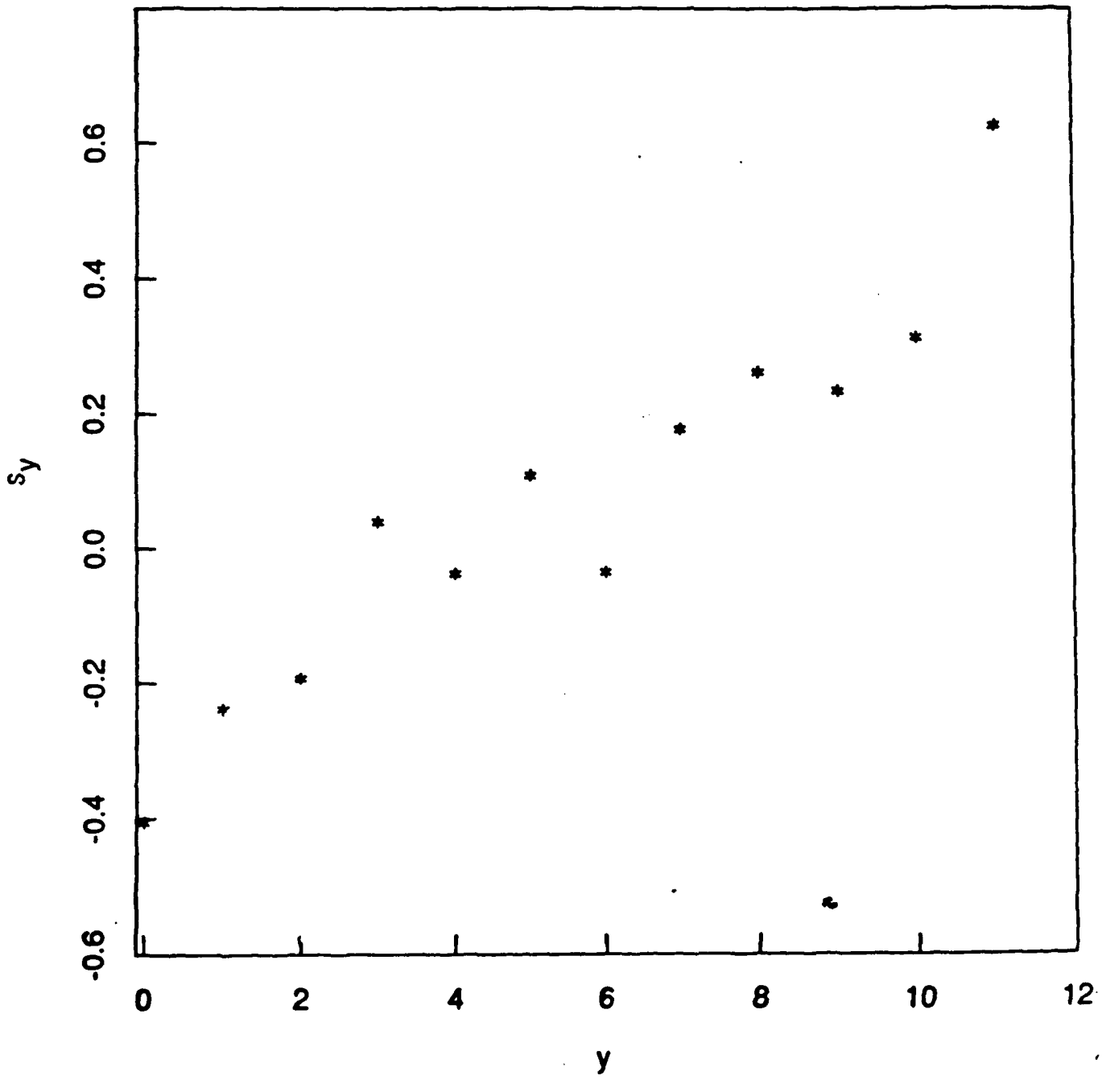

Figure 1: Plot of s_y vs. y for sibship data (Sokal & Rholf, 1973)

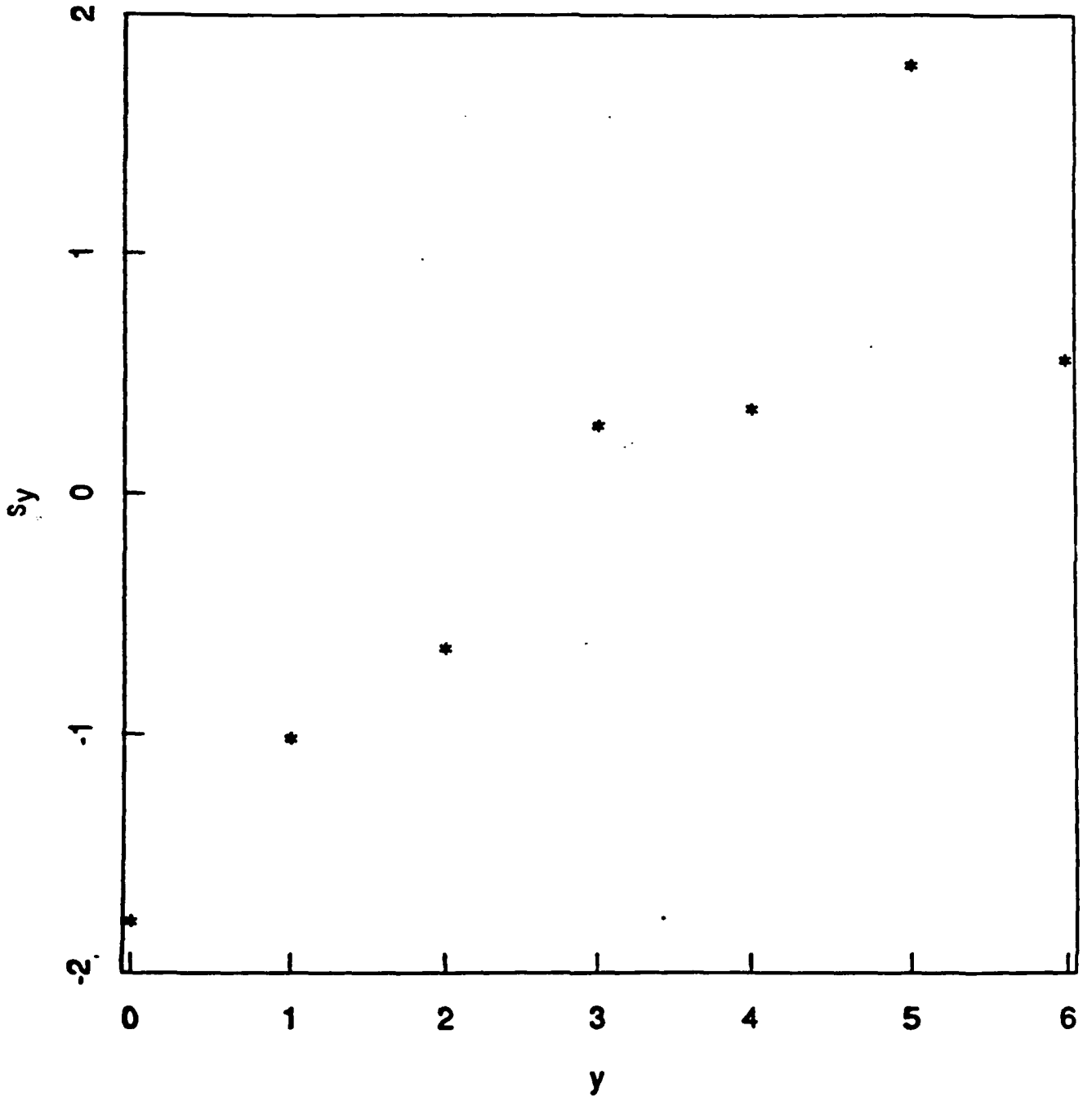

Figure 2: Plot of s_y vs. y for accident data (Seal, 1969)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 425	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Note On Overdispersed Exponential Families		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) A. E. Gelfand and S. R. Dalal		8. CONTRACT OR GRANT NUMBER(s) N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE December 5, 1989 13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Exponential family, overdispersion, exponential dispersion model, weighted least squares fitting, graphical displays for discrete models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 425

20. ABSTRACT

The issue of creating overdispersion in a given one parameter one dimensional exponential family, by extending it to a two parameter exponential family with the same support, is considered. An easily verifiable sufficient condition for this is derived. It is shown that a large class of families satisfy this condition and that this class includes Efron's (1986) and Lindsay's (1986) family as special cases. This class is also closely related to Jorgensen's (1988) Exponential Dispersion Models. UMP unbiased tests for testing overdispersion are exhibited and it is shown that in this context Cox's overdispersion test is a special case. A graphical display is developed to select a family of the general class which should be used with a given set of data to overdisperse a given target one parameter family. Two real data illustrations are given.