Robust and Efficient Estimation by Minimising a Density Power Divergence

BY AYANENDRANATH BASU

Applied Statistics Unit, Indian Statistical Institute,
203 B. T. Road, Calcutta 700 035, India

IAN R. HARRIS

Department of Mathematics, Northern Arizona University,
Flagstaff, Arizona 86011, U.S.A.

NILS L. HJORT

Department of Mathematics and Statistics, University of Oslo,
P.B. 1053 Blindern, N-0316 Oslo, Norway

AND M. C. JONES

Department of Statistics, The Open University,
Milton Keynes, MK7 6AA, United Kingdom

SUMMARY

A minimum divergence estimation method is developed for robust parameter estimation and model fitting. The proposed approach uses new density-based divergences which, unlike existing density-based minimum divergence methods (e.g. minimum Hellinger distance estimation), avoid the use of nonparametric density estimation and associated complications such as bandwidth selection. The proposed class of 'density power divergences' is indexed by a single parameter α which can be varied to study the trade-off between robustness and efficiency. The method can be viewed as a robust extension of maximum likelihood estimation, since the class of divergences contains the Kullback–Leibler divergence when $\alpha = 0$. Choices of α near zero afford robustness while retaining efficiency close to that of maximum likelihood.

Some key words: Asymptotic efficiency; Divergence; Influence function; Maximum likelihood; Robustness.

1. INTRODUCTION

In parametric estimation, density-based minimum divergence methods, i.e. methods which estimate the parameter through minimising a data-based estimate of some appropriate divergence between the assumed model density and the "true" density underlying

the data, have a long history. These procedures include the classical maximum likelihood method as well as the minimum chi-square methods based on the families of chi-square distances studied by several authors (e.g. Neyman, 1949, Rao, 1963, Cressie & Read, 1984, Lindsay, 1994). Beran (1977), using Hellinger distance, was the first to use the technique of density-based minimum divergence estimation in continuous models to develop parameter estimates with good robustness properties relative to maximum likelihood. Among others, Tamura & Boos (1986) and Simpson (1987) have followed up on this line of research. Under some regularity conditions these methods have full asymptotic efficiency at the model. However, in continuous models the methods suffer from the drawback that it is necessary to use some nonparametric smoothing technique such as kernel density estimation to produce a continuous estimate of the true density — they therefore involve all the associated complications such as bandwidth selection. See also Cao, Cuevas & Fraiman (1995). Basu & Lindsay (1994) considered another modification of this approach where the model is smoothed with the same kernel as the data to reduce the dependence of the procedure on the smoothing method.

The present paper introduces a new family of density-based divergence measures, to be called density power divergences. (Note that these measures are *not* closely related to the 'power divergences' of Cressie & Read, 1984.) The family is indexed by a single parameter α which controls the trade-off between robustness and asymptotic efficiency of the parameter estimates which are the minimisers of this family of divergences. When $\alpha = 0$, the density power divergence is the Kullback–Leibler divergence (Kullback & Leibler, 1951) and the method is maximum likelihood estimation; when $\alpha = 1$, the divergence is the mean squared error, and a robust but inefficient minimum mean squared error estimator ensues. For any α, the estimation procedure has the considerable advantage of not requiring any nonparametric smoothing. Various examples are explored to investigate the interplay between robustness and efficiency. It is found that some of the estimators have strong robustness properties with little loss in asymptotic efficiency relative to maximum likelihood under model conditions.

The rest of the paper is organised as follows. In §2 we develop the estimation procedure considered in this paper, discuss some of its properties, and establish the asymptotic normality of the estimators. A robust model choice criterion is also suggested. In §3 we investigate the performance of the estimators in several common parametric families, study the breakdown of the methods in the normal model, and illustrate the performance of the method in some examples. In §4 we develop robust regression procedures utilising these ideas. Concluding remarks are presented in §5. Our work is related to, but different from, that of Windham (1995); see §2.2.

2. The density power divergence and related inference

2·1. The minimum L_2 distance estimator

Consider a parametric family of models $\{F_t\}$, indexed by the unknown parameter $t \in \Omega \subset R^s$, possessing densities $\{f_t\}$ with respect to Lebesgue measure, and let $\mathcal{G}$ be the class of all distributions having densities with respect to Lebesgue measure. Define the minimum L_2 distance, or minimum mean squared error, functional $T_1(\cdot)$ by the requirement that for every G in $\mathcal{G}$,

$$\int \{g(z) - f_{T_1(G)}(z)\}^2 \, dz = \min_{t \in \Omega} \int \{g(z) - f_t(z)\}^2 \, dz \tag{2.1}$$

where g is the density of G. (For the sake of keeping a clear focus in our presentations we have defined the class of densities $\mathcal{G}$ as above, but the results hold for discrete models as well.) Normally $T_1(G)$ would indeed exist and be unique, and we shall assume this to be the case. Suppose also that the parametric family is identifiable in the sense that $t_1 \neq t_2$ implies that $\{z: f_{t_1}(z) \neq f_{t_2}(z)\}$ is a set of positive Lebesgue measure. The minimum L_2 distance functional is then Fisher consistent in the sense that $T_1(F_\theta) = \theta$, uniquely.

Note that the L_2 distance $\int \{g(z) - f_t(z)\}^2 \, dz$ between g and f_t can be represented as $\int f_t^2(z) \, dz - 2 \int f_t(z) \, dG(z) + C$; the quantity C is independent of the parameter t, so does not affect the minimisation procedure. Given a random sample $X_1, \ldots, X_n$ from the true distribution with density g, one can actually minimise

$$\int f_t^2(z) \, dz - 2 \int f_t(z) \, dG_n(z) = \int f_t^2(z) \, dz - 2n^{-1} \sum_{i=1}^{n} f_t(X_i) \tag{2.2}$$

with respect to t, where G_n is the empirical distribution function, to obtain the minimum L_2 distance estimator of the best fitting parameter. Notice that this does not require a smooth nonparametric estimate of g, in contrast to the work of Cao et al. (1995).

Under differentiability of the model and appropriate regularity conditions, the minimum L_2 distance estimators can be obtained by solving the estimating equation

$$n^{-1} \sum_{i=1}^{n} u_t(X_i) f_t(X_i) - \int u_t(z) f_t^2(z) \, dz = 0, \tag{2.3}$$

where $u_t(z) = \partial \log f_t(z) / \partial t$ is the maximum likelihood score function. Note that the above estimating equation is unbiased when $g = f_t$.

For the sake of illustration, let $\{F_t\}$ be a location model, with location parameter t, in which case $\int f_t^2(z) \, dz$ is independent of t, and the minimum L_2 distance estimator is now the maximiser of $\sum_i f_t(X_i)$, with corresponding estimating equation $\sum_i u_t(X_i) f_t(X_i) = 0$. This contrasts with the maximum likelihood estimator which maximises $\sum_i \log f_t(X_i)$, with

the corresponding estimating equation being $\sum_i u_t(X_i) = 0$. For a random variable X in the exponential family with t being the mean value parameter, $u_t(z)$ equals $(z-t)/\sigma^2$, where σ^2 is the variance of X; thus the sample mean is the maximum likelihood estimator for the mean parameter in these families, suggesting the robustness problems of maximum likelihood. On the other hand, for several parametric models such as the normal, $u_t(z)f_t(z)$ is a *bounded* function of z for fixed t, as is the influence function of the minimum L_2 distance functional. This downweighting of the score function is probabilistic, rather than geometric, in the sense that greater downweighting is provided for observations that have lower probabilities of occurrence under the model, as opposed to for observations that are simply far from other observations.

A few examples of the robustness of some variants of the minimum L_2 distance estimator in the normal model have been presented by Brown & Hwang (1993), while studying minimising the L_2 distance between a normal density and a histogram estimating g. Consideration of the small contribution of outliers to L_2 distance based on histograms or kernel density estimates makes this robustness intuitively apparent. See also Terrell (1993), Hjort (1994) and Jones & Hjort (1994).

Unfortunately, however, the robustness of the minimum L_2 distance estimator is achieved at a fairly stiff price in asymptotic efficiency, as we will see later. In order to generate robust estimates with better efficiencies we introduce a family of divergences, and the estimators obtained by minimising these divergences, bridging the gap between maximum likelihood and minimum L_2. Many of these estimators combine strong robustness properties with high asymptotic efficiency.

2·2. The minimum density power divergence estimator

Define the divergence $d_\alpha(g, f)$ between density functions g and f to be

$$d_\alpha(g,f) = \int \left\{ f^{1+\alpha}(z) - \left(1+\frac{1}{\alpha}\right) g(z) f^\alpha(z) + \frac{1}{\alpha} g^{1+\alpha}(z) \right\} dz \quad \text{for } \alpha > 0. \tag{2.4}$$

When $\alpha = 0$, the integrand in the expression (2.4) is undefined, and we define the divergence $d_0(g, f)$ as

$$d_0(g,f) = \lim_{\alpha \to 0} d_\alpha(g,f) = \int g(z) \log(g(z)/f(z))\, dz.$$

Notice that $d_0(g, f)$ is a version of the Kullback–Leibler divergence.

In the estimation procedure that we discuss in this paper, we are most interested in smaller values of $\alpha \geq 0$, say between zero and one, although values greater than one can be considered too. The procedure typically becomes less and less efficient as α increases as we will see later.

THEOREM 2·1. *The quantity $d_\alpha(g, f)$ is a divergence in that it is nonnegative for all $g, f \in \mathcal{G}$ and is equal to zero if and only if $f \equiv g$ almost everywhere.*

Proof. Suppose $\alpha > 0$. Consider the integrand at a fixed value y. Factor out the term $g^{1+\alpha}(y)/\alpha$ and define $x = f(y)/g(y)$. To show that $d_\alpha(g, f)$ is nonnegative it is sufficient to show that the factored integrand, $I_\alpha(x) = \alpha x^{1+\alpha} - (1+\alpha)x^\alpha + 1$, is nonnegative for all $x \geq 0$. Clearly $I_\alpha(1) = 0$, and in fact this is the unique minimum for $x \geq 0$; the derivative of I_α is strictly negative for all $x < 1$, zero for $x = 1$ and strictly positive for all $x > 1$. Thus $d_\alpha(g, f)$ is nonnegative and is equal to zero if and only if $g = f$ identically.

When $\alpha = 0$, it is well known that the above Kullback–Leibler divergence is nonnegative and is equal to zero if and only if $g \equiv f$. □

The family of divergences d_α, as a function of α, will be called the class of density power divergences. Under the set-up of the previous section, the following is a simple consequence of Theorem 2.1: for any given α the minimum density power divergence functional at G, defined by the requirement $d_\alpha(g, f_{T(G)}) = \min_{t\in\Omega} d_\alpha(g, f_t)$, is Fisher consistent; we will denote this functional by $T_\alpha(G)$. In addition, the minimum density power divergence estimator $\widehat{\theta}$, generated by minimising

$$H_n(t) = \int f_t^{1+\alpha}(z)\, dz - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f_t^\alpha(X_i) \equiv n^{-1} \sum_{i=1}^{n} V_{n,t}(X_i) \tag{2.5}$$

with respect to t, is weakly consistent for $\theta = T_\alpha(G)$ as well (see Theorem 2.2). We assume here that $T_\alpha(G)$ exists and is unique, as will normally be the case. Verifying this is perhaps most easily done on a case by case basis, and would depend on the parameter space and the complexity of the $\{f_t\}$ family as well as on the true density g.

Consider the functional $T_0(\cdot)$. Given the data, $T_0(G_n)$ maximises $\int \log f_t(z)\, dG_n(z)$, and is therefore the maximum likelihood estimate of the parameter if it exists. On the other hand, the value $\alpha = 1$ gives precisely the L_2 distance between the densities discussed in §2.1. Thus, for $0 < \alpha < 1$, the class of density power divergences provides a smooth bridge between the L_2 distance and the Kullback–Leibler divergence.

Some motivation for the form of the divergence (2.4) can be obtained by again looking at the location model, where $\int f_t^{1+\alpha}(z)\, dz$ is independent of t. In this case, the proposed estimators maximise $\sum_i f_t^\alpha(X_i)$, with the corresponding estimating equations having the form

$$\sum_{i=1}^{n} u_t(X_i) f_t^\alpha(X_i) = 0. \tag{2.6}$$

Equation (2.6) can be viewed as a weighted version of the efficient maximum likelihood score equation. When $\alpha > 0$, (2.6) provides a relative-to-the-model downweighting for outlying

observations; observations that are are wildly discrepant with respect to the model will get nearly zero weights. In the fully efficient case $\alpha = 0$, all observations, including very severe outliers, get weights equal to one. By choosing a value of α close to zero, one makes all the weights closer to 1 compared to the minimum L_2 method, improving the asymptotic efficiency of the procedure. The proposed estimators, therefore, represent compromises between efficiency and robustness, with the degree of compromise controlled by the tuning parameter α.

For general families it can be checked easily that the estimating equations have the form

$$U_n(t) = \int u_t(z) f_t^{1+\alpha}(z)\, dz - n^{-1} \sum_{i=1}^{n} u_t(X_i) f_t^{\alpha}(X_i) = 0. \tag{2.7}$$

Again this estimating equation is unbiased when $g = f_t$. Notice that this has the appealing advantage that it does not require a smooth estimate of g which is necessary in other robust density-based minimum divergence approaches (e.g. Beran, 1977, Cao et al., 1995); thus the bandwidth selection problem and rate of convergence results for the kernel density estimator are no longer relevant.

We now present the asymptotic distribution of the minimum density power divergence estimators, when the data are generated from the true distribution G not necessarily in the model. (In the following, θ represents the best fitting value of the parameter, whereas t denotes a generic element of Ω.) Let $X_1, \ldots, X_n$ be independent and identically distributed with distribution G with corresponding density g, $T_\alpha(G) = \theta = (\theta_1, \ldots, \theta_s)$, and let $\widehat{\theta} = \widehat{\theta}_n$ be the minimiser of (2.5). Let $K = K(\theta)$ be the covariance matrix of $T = f_t^{\alpha}(X) u_t(X)$ under G i.e.

$$K = \int u_\theta(z) u_\theta^T(z) f_\theta^{2\alpha}(z) g(z)\, dz - \xi_\theta \xi_\theta^T \quad \text{and} \quad \xi_\theta = \int u_\theta(z) f_\theta^{\alpha}(z)\, g(z) dz. \tag{2.8}$$

For any given α, make the following assumptions:

A1: The distributions F_t and G have common support, so that the set A on which the densities are greater than zero is independent of t.

A2: There is an open subset ω of the parameter space Ω containing the best fitting parameter θ such that for almost all $z \in A$, and all $t \in \omega$, the density $f_t(z)$ is three times differentiable with respect to t and the third partial derivatives are continuous with respect to t.

A3: The integral $\int f_t^{1+\alpha}(z)\, dz$ can be differentiated three times with respect to t, and the derivative can be taken under the integral sign.

A4: The matrix $J = J(t)$, defined by

$$J(t) = \int u_\theta(z) u_\theta^T(z) f_\theta^{1+\alpha}(z)\, dz + \int (i_\theta(z) - \alpha u_\theta(z) u_\theta^T(z))(g(z) - f_\theta(z)) f_\theta^{\alpha}(z)\, dz, \tag{2.9}$$

where $i_t(x) = -\partial\{u_t(x)\}/\partial t$, the so called information function of the model, is positive definite for all $t \in \omega$.

A5: There exist functions $M_{jkl}(x)$ such that $|\partial^3 V_{n,t}(x)/\partial t_j \partial t_k \partial t_l| \leq M_{jkl}(x)$ for all $t \in \omega$, where $E_G[M_{jkl}(X)] < \infty$ for all j, k and l, where E_G denotes expectation with respect to G.

THEOREM 2·2. *Under the above conditions, with probability tending to 1 as $n \to \infty$, there exists $\widehat{\theta}_n$ such that*

(i) $\widehat{\theta}_n$ is consistent for θ, and

(ii) $n^{1/2}(\widehat{\theta}_n - \theta)$ is asymptotically multivariate normal with (vector) mean zero and covariance matrix $J^{-1}KJ^{-1}$.

The proof of Theorem 2.2 follows closely the proof of Theorem 6.4.1 of Lehmann (1983) (which is for the maximum likelihood estimator) with appropriate modifications to cope with our density power divergence and the allowance of distributions outside the model. The proof is omitted to save space; full details may be obtained from the first author.

For simplicity of notation, the subscript α has been dropped from the quantities $\widehat{\theta}$, H_n, $V_{n,t}$, U_n, M_{jkl}, as well as the matrices J and K. In addition, $\widehat{\theta}_n$ will revert to $\widehat{\theta}$ in what follows. The simplified formulae occurring when G is in the model will be considered in §3.

Note that the divergence given by (2.4) is close to a weighted L_2 distance (Hjort, 1994) in the sense that, for fixed α, and f close to g, $d_\alpha(g, f)$ becomes close to

$$\frac{1}{2}(1+\alpha)\int g^{\alpha-1}(z)\{f(z) - g(z)\}^2\, dz. \tag{2.10}$$

Observe how minimum L_2 corresponds (exactly) to a unit weighting, maximum likelihood corresponds to a $1/g$ weighting, and minimum density power divergence for $0 < \alpha < 1$ corresponds to an intermediate $1/g^\gamma$ for $0 < \gamma < 1$ weighting. Unlike (2.10), however, the beauty of (2.4) is that, ignoring the last term because it does not depend on f, g appears only as a multiplier of terms in f. Thus while f will be replaced by f_t, g can appropriately be replaced by its empirical version, and there is no need to introduce any smoothing into the formulation. (The same holds, of course, for maximum likelihood estimation.)

The idea of downweighting with respect to the model rather than the data is also the motivating principle of Windham (1995). Windham describes a fixed point algorithm that also uses density power weighting. Windham's procedure is equivalent to choosing t such that

$$\frac{\sum_i u_t(X_i) f_t^\alpha(X_i)}{\sum_i f_t^\alpha(X_i)} = \frac{\int u_t(z) f_t^{1+\alpha}(z)\, dz}{\int f_t^{1+\alpha}(z)\, dz}. \tag{2.11}$$

If f_t is a location family then (2.11) reduces to (2.6), and thus for this special case Windham's procedure is identical to ours. In general (2.11) does not reduce to (2.7). Insights into the relationship between the two methods and practical comparisons between them are the subject of a further paper currently in preparation.

2·3. Influence function and standard error

Let $G_\epsilon(z) = (1-\epsilon)G(z) + \epsilon\chi_y(z)$, $0 < \epsilon < 1$, where $\chi_y(z)$ is the distribution function of the random variable which puts all its mass on y. By direct differentiation of equation (2.7) (with G_ϵ in place of the implicit G_n) with respect to ϵ, one gets the influence function of the density power divergence functional to be

$$I_\alpha(G, y) = \frac{\partial}{\partial\epsilon} T_\alpha(G_\epsilon)|_{\epsilon=0} = J^{-1}(u_\theta(y) f_\theta^\alpha(y) - \xi_\theta), \quad \text{where } \theta = T_\alpha(G)$$

and ξ_θ and J are as in (2.8) and (2.9). Assuming that J and ξ_θ are finite, this is a bounded function of y whenever $u_\theta(y) f_\theta^\alpha(y)$ is bounded. This is true, for example, for any $\alpha > 0$ in the normal location-scale problem, unlike other density based minimum divergence procedures such as those based on the Hellinger distance. The influence functions for the estimation of the normal mean when $\sigma = 1$ are plotted in Figure 1 for several values of α; note their redescending nature for all $\alpha > 0$.

* * * Figure 1 about here * * *

The asymptotic variance of ($\sqrt{n}$ times) the minimum density power divergence estimator can be consistently estimated in a sandwich fashion by using the above influence function. Let $K_i = u_\theta(X_i) f_\theta^\alpha(X_i) - \xi_\theta$, and $\widehat{K_i}$ be the corresponding quantity evaluated at $\widehat{\theta}$, with G_n in place of G. Let $\widehat{K} = (n-1)^{-1} \sum_i (\widehat{K_i} \widehat{K_i}^T)$. Then the asymptotic variance of $\sqrt{n}$ times the parameter estimates can be consistently estimated by $\widehat{J}^{-1} \widehat{K} \widehat{J}^{-1}$, where $\widehat{J}$ is obtained from J by replacing θ with $\widehat{\theta}$, with G_n in place of G. Consistent estimates of the asymptotic variance of the method can also be obtained by the jackknife and bootstrap techniques.

2·4. Equivariance

The maximum likelihood method has two important equivariance properties; estimates are equivariant with respect to both reparametrisations and transformation of the data. Our minimum density power divergence method shares the first general property: if the model is reparametrised to $\psi = \psi(\theta)$ with a one-one transformation, then the density power divergence estimate of ψ is simply $\widehat{\psi} = \psi(\widehat{\theta})$, in terms of the density power divergence estimate of θ, using the same α. This follows from definition (2.5).

The second maximum likelihood property does not generally hold for the new estimation method, however. If data are transformed from X_i to $Y_i = h(X_i)$, then the minimum density power divergence estimator, say θ^*, is defined as the minimiser of

$$\int \{f_t(\xi(y))|\xi'(y)|\}^{1+\alpha}\, dy - n^{-1}(1+\alpha^{-1})\sum_{i=1}^{n}\{f_t(X_i)|\xi'(h(X_i))|\}^{\alpha},$$

where $X_i = \xi(Y_i)$ is the inverse transformation. We see by comparison with (2.5) that θ^* is equal to $\widehat{\theta}$ only if $\xi'(y)$ is a non-zero constant. Thus the estimation method is equivariant under a $Y_i = aX_i + b$ type data transformation, but not under other transformations (unless $\alpha = 0$).

2·5. Hypothesis testing

As a consequence of Theorem 2.2, one can readily construct Wald and score type tests for the null hypothesis $H_0: \theta = \theta_0$. We work under model conditions although it is possible to test hypotheses about θ outside the model too, where $\theta = T_\alpha(G)$. Under the null, the asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$ is normal with mean zero and covariance matrix $C(\theta) = J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0)$. As a result the Wald type statistic $n(\widehat{\theta} - \theta_0)^T C^{-1}(\widehat{\theta})(\widehat{\theta} - \theta_0)$ has an asymptotic $\chi^2(s)$ distribution. Similarly the score type statistic given by $nU_n^T(\theta_0)K^{-1}(\theta_0)U_n(\theta_0)$ has, under the null, the same asymptotic $\chi^2(s)$ distribution and is asymptotically equivalent to the Wald type statistic.

For the composite null hypothesis, let $\theta = (\theta_1^T, \theta_2^T)^T$ where θ_1 lies in an s_1-dimensional subspace of Ω. Consider the null hypothesis $H_0: \theta_1 = \theta_{0,1}$ where θ_2 is unspecified. Let $\widehat{\theta} = (\widehat{\theta}_1^T, \widehat{\theta}_2^T)^T$ and $\widehat{\theta}_N = (\theta_1^{*T}, \theta_2^{*T})^T$ be the minimum density power divergence estimates of the parameter without any restriction and under the null hypothesis respectively. Let $C_{11}^{-1}(\theta)$ and $K_{11}^{-1}(\theta)$ be the $s_1 \times s_1$ blocks corresponding to θ_1 in $C^{-1}(\theta)$ and $K^{-1}(\theta)$. Also let $U_{1,n}$ be the component of the density power score function corresponding to θ_1. Then the Wald type and the score type statistics, given by $n(\widehat{\theta}_1 - \theta_{0,1})^T C_{11}^{-1}(\widehat{\theta})(\widehat{\theta}_1 - \theta_{0,1})$ and $nU_{1,n}(\widehat{\theta}_N)K_{11}^{-1}(\widehat{\theta}_N)U_{1,n}(\widehat{\theta}_N)$ are both asymptotically $\chi^2(s_1)$.

2·6. A robust model choice criterion

Model choice criteria of the Akaike information variety penalise a model's achieved maximum log-likelihood with a term which depends suitably on the complexity of the candidate model. The arguments used to motivate and construct these criteria are typically asymptotic in nature, relying on the large-sample behaviour of maximum likelihood estimators. A similar route can be followed for the present type of robust estimators, working with d_α of (2.4) instead of d_0. We will in fact argue in favour of the following strategy. For each

candidate model M, compute the d_α divergence estimate $\widehat{\theta}_M$, which by the theory above has an associated estimated variance matrix of the form $n^{-1}\widehat{J}_M^{-1}\widehat{K}_M\widehat{J}_M^{-1}$. Then evaluate the robust information criterion

$$\mathrm{RIC}_M = H_n(\widehat{\theta}_M) + (1+\alpha)n^{-1}\,\mathrm{Tr}(\widehat{J}_M^{-1}\widehat{K}_M). \tag{2.12}$$

In the end choose the model with the smallest value of RIC. The limiting version of this as α tends to zero can be shown to be the same as maximising the achieved log-likelihood minus $\mathrm{Tr}(\widehat{J}_M^{-1}\widehat{K}_M)$ (involving suitable $\alpha = 0$ definitions of $\widehat{J}_M$ and $\widehat{K}_M$, see §2.3). But this is quite close to the traditional Akaike method which takes the view that the models are (approximately) correct, in particular entailing $J_M = K_M$ (with $\alpha = 0$); that is, the trace above becomes the number of parameters in the model.

To show how RIC evolves, agree first to put

$$\rho = \int f_{\widehat{\theta}}^{1+\alpha}(y)\,dy - \left(1+\frac{1}{\alpha}\right)\mathrm{E}\{f_{\widehat{\theta}}^{\alpha}(X_{n+1})\,|\,\mathrm{data}\},$$

which is a fixed constant away from being the distance $d_\alpha(g, f_{\widehat{\theta}}(\cdot))$ from truth to estimated model, and the conditional expectation operation is with respect to a new observation X_{n+1}, independent of previous data. We think of ρ as the predictive quality of the estimated model, and aim for models with as small ρ values as possible. Note, using (2.5), that $H_n(\widehat{\theta})$, being the minimum of an empirical process, will tend to undershoot the real ρ. A more balanced method is via cross validation,

$$\widehat{\rho}_x = \int f_{\widehat{\theta}}^{1+\alpha}(z)\,dz - \left(1+\frac{1}{\alpha}\right)n^{-1}\sum_{i=1}^{n} f_{\widehat{\theta}_{(i)}}^{\alpha}(X_i),$$

writing $\widehat{\theta}_{(i)}$ for the estimate obtained by leaving X_i out of the data set. The $\hat{\rho}_x$ estimator is almost unbiased for ρ, and can indeed be used as a model selection criterion.

It is fruitful to work out an approximation which is less intensive computationally. This can be done via influence functions. We find

$$\widehat{\theta} = T((1-n^{-1})\widehat{G}_{(i)} + n^{-1}\chi_{X_i}) \doteq \widehat{\theta}_{(i)} + n^{-1}I_\alpha(\widehat{G}, X_i),$$

involving the leave-X_i-out version of the empirical distribution in addition to χ_{X_i} as in §2.3. This leads to

$$f_{\widehat{\theta}_{(i)}}^{\alpha}(X_i) \doteq \widehat{f}_i^{\alpha}(1 - n^{-1}\alpha\widehat{u}_i^T\widehat{I}_i),$$

where $\widehat{f}_i$, $\widehat{u}_i$ and $\widehat{I}_i = \widehat{J}^{-1}(\widehat{u}_i\widehat{f}_i^{\alpha} - \widehat{\xi})$, with $\widehat{\xi} = n^{-1}\sum_{i=1}^n \widehat{u}_i\widehat{f}_i^{\alpha}$, are the natural empirical versions of $f_\theta(X_i)$, $u_\theta(X_i)$ and $I_\alpha(G, X_i)$. But this yields

$$\widehat{\rho}_x \doteq H_n(\widehat{\theta}) + (1+\alpha)\,n^{-2}\sum_{i=1}^{n}\widehat{f}_i^{\alpha}\widehat{u}_i^T\widehat{I}_i = H_n(\widehat{\theta}) + (1+\alpha)\,n^{-1}\mathrm{Tr}(\widehat{J}^{-1}\widehat{K})$$

after some calculations.

There is an alternative route to establishing this RIC formula, more akin to deductions one may find in the literature for the AIC criterion. The derivation above, however, does not presume that the models worked with are actually correct, and exhibits the cross-validation formula as a selection criterion of separate interest.

3. Special parametric families: efficiency, breakdown and examples

Suppose that the true distribution g belongs to the parametric family $\{f_t\}$, θ being the true value of the parameter. Then the formulae for J, K and ξ_θ simplify to

$$J = \int u_\theta(z) u_\theta^T(z) f_\theta^{1+\alpha}(z)\, dz, \tag{3.1}$$

$$K = \int u_\theta(z) u_\theta^T(z) f_\theta^{1+2\alpha}(z)\, dz - \xi_\theta \xi_\theta^T \quad \text{and} \quad \xi_\theta = \int u_\theta(z) f_\theta^{1+\alpha}(z)\, dz. \tag{3.2}$$

Note that in the limit $\alpha \to 0$, J and K both become equal to the classic Fisher information. These formulae can be used to investigate the asymptotic efficiency of the estimators, and in particular to judge how much is lost relative to the maximum likelihood estimator under model conditions. In the following subsection, some examples for particular parametric families are considered. We will define the asymptotic relative efficiency of an estimator to be the ratio of the asymptotic variance of the maximum likelihood estimator to that of the estimator in question.

3·1. Efficiencies for particular families

(a) Mean of univariate normal. For a location family $\xi_\theta = 0$. Letting f_θ be the $N(\mu, \sigma^2)$ density with known σ^2 and u_θ the score function with respect to the mean parameter μ, elementary integration gives

$$K = (2\pi)^{-\alpha} \sigma^{-(2+2\alpha)} (1+2\alpha)^{-3/2} \quad \text{and} \quad J = (2\pi)^{-\alpha/2} \sigma^{-(2+\alpha)} (1+\alpha)^{-3/2}.$$

The asymptotic variance of $n^{1/2}$ times the estimator of μ is then given by

$$\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/2} \sigma^2. \tag{3.3}$$

Since the asymptotic variance of $n^{1/2}$ times the maximum likelihood estimator is σ^2, the asymptotic relative efficiency of the density power divergence estimator is easy to compute. For $\alpha = 0.25$ it is 0.941, for example, already quite close to one. Results for different values of α are given in the first row of Table 1.

* * * Table 1 about here * * *

(b) Standard deviation of univariate normal. Again, let f_θ be the $N(\mu, \sigma^2)$ density but treat both parameters as unknown. Calculations for the two 2×2 matrices J and K show that both have zeros off the diagonals, that is, the estimators $\widehat{\mu}$ and $\widehat{\sigma}$ are asymptotically independent. The limiting distribution for $\sqrt{n}(\widehat{\mu} - \mu)$ is therefore as found in case (a) even when σ is unknown.

Here, we concentrate on estimation of σ. Lengthy calculations show that the asymptotic variance of $n^{1/2}$ times the estimator is

$$\frac{(1+\alpha)^2}{(2+\alpha^2)^2}\left\{\frac{2Q(\alpha)}{(1+2\alpha)^{5/2}} - \alpha^2\right\}\sigma^2$$

where $Q(\alpha) = 1 + 3\alpha + 5\alpha^2 + 7\alpha^3 + 6\alpha^4 + 2\alpha^5$. Efficiency calculations (compare with $\sigma^2/2$) are presented in the second row of Table 1. Small α density power divergence estimation continues to retain high efficiency. The values in Table 1 clearly show that the minimum L_2 distance estimators of μ and σ are quite inefficient; see also Hjort (1994).

(c) Exponential distribution. For the density $f_\theta(x) = \theta^{-1}\exp(-x/\theta)$, $x > 0$, the quantities K and J in the asymptotic variance of $n^{1/2}$ times the minimum density power divergence estimator of θ are given by

$$K = \left\{\frac{1+4\alpha^2}{(1+2\alpha)^3} - \frac{\alpha^2}{(1+\alpha)^4}\right\}\theta^{-(2+2\alpha)} \quad \text{and} \quad J = \frac{1+\alpha^2}{(1+\alpha)^3}\theta^{-(2+\alpha)}.$$

The asymptotic variance is then given by

$$\frac{(1+\alpha)^2 P(\alpha)\theta^2}{(1+\alpha^2)^2(1+2\alpha)^3}$$

where $P(\alpha) = 1 + 4\alpha + 9\alpha^2 + 14\alpha^3 + 13\alpha^4 + 8\alpha^5 + 4\alpha^6$. Again the asymptotic variance of $n^{1/2}$ times the maximum likelihood estimator is θ^2, so the asymptotic relative efficiencies are easily obtained. They are given for certain α in the third row of Table 1. Again, efficiencies remain high for small α.

(d) Mean of multivariate normal. The family is $N_p(\mu, \Sigma)$. The limiting covariance matrix of $n^{1/2}$ times the minimum density power divergence estimator of μ (whether or not Σ is known) can be shown to be

$$\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{p/2+1}\Sigma.$$

Thus one loses efficiency for increasing p if α is kept fixed.

(e) Poisson distribution. Calculation of the asymptotic variance of the estimator can be carried out numerically, although not via a closed-form formula. It involves an infinite but rapidly convergent sum. In Table 1 we also provide the asymptotic relative efficiencies of the estimators for two different values of the mean parameter λ and several choices of α. Note that the Poisson results are very similar to those for normal μ for $\lambda \geq 10$.

3·2. Breakdown in the normal distribution

The breakdown point of an estimator, crudely described as the proportion of bad observations that an estimator can tolerate before it becomes completely uninformative, is one of the descriptors of the robustness of the method. Here we determine the gross-error breakdown point (Hampel, Ronchetti, Rousseuw & Stahel, 1996, p.97) of the minimum density power divergence estimator of the parameters of the normal distribution under a particular contamination.

Let $\alpha > 0$ and let g be the $N(\mu, \sigma^2)$ density, written $\phi_{\mu,\sigma}(\cdot) = \sigma^{-1}\phi(\sigma^{-1}(\cdot - \mu))$, f_t the $N(m, s^2)$ density and $q(z) = (1-\epsilon)g(z) + \epsilon\delta_x(z)$, where δ is the Dirac delta function and $x \to \infty$. The data are a random sample from q and the target parameters are $\theta = (\mu, \sigma)$.

Consider the maximiser of

$$\begin{aligned}\psi(m,s) &\equiv (1+\alpha)\int q(z) f_t^{\alpha}(z)\,dz - \alpha\int f_t^{\alpha+1}(z)\,dz \\ &= (1+\alpha)\{(1-\epsilon)\int \phi_{\mu,\sigma}(z)\phi_{m,s}^{\alpha}(z)\,dz + \epsilon\int \delta_x(z)\phi_{m,s}^{\alpha}(z)\,dz\} - \alpha\int \phi_{m,s}^{1+\alpha}(z)\,dz\end{aligned}$$

with respect to m and s. If location breakdown occurs, the value of m which maximises the above goes to ∞, if scale breakdown occurs, the maximising value of s goes to 0 or ∞ (Hampel et al., 1986, p.98).

To evaluate $\psi(m,s)$, the following result, provable by elementary calculations, is useful:

$$\int \phi_{c,d}(z)\phi_{m,s}^{\alpha}(z)\,dz = \frac{\exp[-\alpha(c-m)^2/\{2(s^2+\alpha d^2)\}]}{(2\pi)^{\alpha/2}s^{\alpha}\left(1+\frac{\alpha d^2}{s^2}\right)^{1/2}}.$$

Write $A = \sigma/s$. It follows that $\psi_1(m,A) \equiv (2\pi)^{\alpha/2}\sigma^{\alpha}\psi(m,s)$ is given by

$$\begin{aligned}\psi_1(m,A) &= A^{\alpha}\left(\frac{(1+\alpha)(1-\epsilon)\exp[-\alpha(\mu-m)^2/\{2(s^2+\alpha\sigma^2)\}]}{(1+\alpha A^2)^{1/2}}\right. \\ &+ \left.\epsilon(1+\alpha)\exp\{-\alpha A^2(x-m)^2/(2\sigma^2)\} - \frac{\alpha}{(1+\alpha)^{1/2}}\right).\end{aligned}$$

We now wish to maximise this quantity over A (rather than s) and m. First, $\psi_1(m,0) = 0$. For $A > 0$, $\psi_1(m,A)$ consists essentially of two ridges which have heights

$$A^{\alpha}\left\{\frac{(1+\alpha)(1-\epsilon)}{(1+\alpha A^2)^{1/2}} - \frac{\alpha}{(1+\alpha)^{1/2}}\right\} \text{ at } m = \mu$$

and

$$A^{\alpha}\left\{\epsilon(1+\alpha)-\frac{\alpha}{(1+\alpha)^{1/2}}\right\} \text{ at } m=x.$$

If the $m=x$ ridge height is negative, i.e. $\epsilon < K \equiv \alpha/(1+\alpha)^{3/2}$, $A=0$ would be optimal if the $m=\mu$ ridge height is negative for all $A>0$ too. The latter happens if $\epsilon > 1-K$. However, $1-K < \epsilon < K$ is impossible because $K < 1/2$. So, $A=0$ cannot maximise $\psi_1(m,A)$.

However, if the $m=x$ ridge height is positive, the value along this ridge tends to ∞ as $A \to \infty$. The values along the $m=\mu$ ridge, however, stay finite: even if they are positive somewhere, they will have a finite maximum at a finite A and tend to a negative quantity as $A \to \infty$. That is, the maximum, if the $m=x$ ridge is positive, is at $m=x$ and $s=0$. This makes sense because for enough bad points, the normal fit tends to match δ_x with mean $x \to \infty$ and variance zero. This is simultaneous location and scale breakdown in the sense that location "explodes" and scale "implodes" (Hampel et al., 1986, p.98).

Breakdown therefore occurs if

$$\epsilon > \alpha/(1+\alpha)^{3/2}.$$

The breakdown point increases monotonically from zero when $\alpha \simeq 0$ (in line with the zero breakdown of the maximum likelihood estimator which can easily be shown separately) to $1/(2\sqrt{2}) = 0.354$ when $\alpha = 1$. (In fact, the breakdown continues to increase until its maximal value of $2/(3\sqrt{3}) = 0.385$ at $\alpha = 2$, but by then the efficiency of the estimator is unacceptably low.)

3·3. Examples

In our first example we consider Newcomb's light speed data (Stigler, 1977). The data set can be found in many elementary texts, including Moore & McCabe (1993). The data were also analysed by Brown & Hwang (1993), who were trying to fit the "best approximating normal distribution" to the corresponding histogram. The limiting case of their approach generates the normal distribution whose mean and standard deviation are the minimum L_2 distance estimates of μ and σ under a normal model. This estimator, it was observed, quite successfully downweighted the extreme outliers in the Newcomb data.

* * * Table 2 and Figure 2 about here * * *

For this dataset, Table 2 gives the values of the minimum density power divergence estimates of μ and σ for various values of α under the normal model. These estimators

exhibit strong outlier resistance properties even for quite small values of α. When α is as small as 0.1 (for which the minimum density power divergence estimator of σ has an efficiency loss of only 2.4% under the model) the estimate of σ is 5.39, fairly close to the estimate obtained for $\alpha = 1$. A visual representation of this is provided in Figure 2, where the normal densities $N(\widehat{\mu}, \widehat{\sigma}^2)$, for $\alpha = 1$, 0.5, 0.25, 0.1 and 0 are superimposed on a histogram of the Newcomb data. Except when the maximum likelihood estimator is used, all the normal densities fit the main body of the histogram quite well, even the one with $\alpha = 0.1$.

In the second example our estimation method is applied to chemical mutagenicity data previously analysed by Simpson (1987) in the context of minimum Hellinger distance estimation. In the sex linked recessive lethal test in drosophila (fruit flies), male flies are exposed to different doses of a chemical to be screened. They are then mated with unexposed females and for each male the number of daughter flies carrying a recessive lethal mutation on the X chromosome is noted. One such experiment with 34 males resulted in 23, 7, 3 and 1 males having 0, 1, 2 and 91 such daughters respectively. Note that the last value of 91 is a very large outlier. Simpson considered a Poisson fit for these data, and found that the minimum Hellinger distance estimate of the mean parameter λ successfully downweights the large outlier, unlike the maximum likelihood method.

* * * Table 3 about here * * *

Here we compute the minimum density power divergence estimates for these data under the Poisson(λ) model. The results are presented in Table 3. As expected the more robust members of the family downweight the large outlier successfully. However, what is more interesting is that this downweighting can be observed even for very small values of α. The procedure apparently loses robustness for some α between 0.01 and 0.001. For comparison, the maximum likelihood estimate of λ after deleting this outlier is 0.394, and the minimum Hellinger distance estimate of λ for these data (with and without the outlier) is 0.364.

The last example involves hypothesis testing in the normal model on a set of telephone line fault data presented in Welch (1987), which was also previously analysed by Simpson (1989). The data in Table 4 represent the difference of the inverse fault rates between the test and the control in 14 matched pairs. Here we do a parametric test under the $N(\mu, \sigma^2)$ model of the hypothesis H_0: $\mu = 0$ versus H_1: $\mu > 0$, where σ is unspecified, using the above data. We perform one-sided Wald type and score type tests of the null hypothesis.

* * * Tables 4 and 5 about here * * *

For a random sample $X_1, X_2, \ldots, X_n$ from the $N(\mu, \sigma^2)$ distribution, letting $\widehat{\theta} = (\widehat{\mu}_\alpha, \widehat{\sigma}_\alpha)$ and $\widehat{\theta}_N = (0, \widehat{\tau}_\alpha)$ be the unrestricted and the null estimates of $\theta = (\mu, \sigma)$, the Wald and score type statistics W_α and R_α have the form

$$W_\alpha = \frac{n^{1/2}\widehat{\mu}_\alpha}{(1+\frac{\alpha^2}{1+2\alpha})^{3/4}\widehat{\sigma}_\alpha} \quad \text{and} \quad R_\alpha = \frac{(1+2\alpha)^{3/4}}{n^{1/2}\widehat{\tau}_\alpha}\left[\sum_{i=1}^n X_i \exp\Big(-\frac{\alpha X_i^2}{2\widehat{\tau}_\alpha^2}\Big)\right].$$

Under the null hypothesis these statistics have asymptotic $N(0,1)$ distributions.

The results of our analysis of the telephone fault data using the Wald type test are presented in Table 5, where the statistics and their one-sided p-values are presented for several values of α, the p-values being calculated under a normal distribution. Because of the presence of the large outlier the likelihood based methods fail to detect the improvement of the test method over the control; the more robust methods provide a better picture of the comparison of the two sets of data. Similar results (not shown) arose when using the score type test.

In the above examples, we successfully used a simple bisection method for the one parameter case and Newton-Raphson in the two parameter cases, with fast results. Computational questions for larger and more difficult problems are left for future research.

4. Density power divergence estimation in regression models

It is important to extend the estimation methods to regression type situations, where response data y are to be explained through covariate information x. Here we propose such an extension. We also indicate briefly how statistical inference using the resulting robust regression estimators can be carried out.

4·1. Estimation method

Assume that a parametric regression model $f_\beta(y \,|\, x)$ is proposed for data (x_1, y_1), $\ldots, (x_n, y_n)$, where the model family is smooth in its, say, p-dimensional parameter β. The standard assumption in such situations is that the Y_is are conditionally independent given $x_1, \ldots, x_n$. The estimation methods we propose below are intended to work in all such cases, and inference can be carried out conditionally on the observed covariate values. We think of the x_is as coming from a suitable covariate distribution Q in the covariate measurement space $\mathcal{X}$. Thus averages $n^{-1}\sum_{i=1}^n h(x_i)$ will under very mild ergodic conditions tend in probability to limits $\int h(x)\, dQ(x) = \mathrm{E}_Q h(x)$, provided these are finite.

Let there be a true density $g(y \mid x)$ for Y given $X = x$. Consider the x-conditional version of the divergence (2.4),

$$d_\alpha(g(\cdot \mid x), f_\beta(\cdot \mid x)) = \int \left\{ f_\beta^{1+\alpha}(y \mid x) - \left(1 + \frac{1}{\alpha}\right) g(y \mid x) f_\beta^\alpha(y \mid x) + \frac{1}{\alpha} g^{1+\alpha}(y \mid x) \right\} dy, \tag{4.1}$$

from true density $g(\cdot \mid x)$ to parametrically modelled $f_\beta(\cdot \mid x)$. Our proposal is to use $\widehat{\beta}$, the parameter value that minimises

$$H_n(\beta) = n^{-1} \sum_{i=1}^n \int f_\beta^{1+\alpha}(y \mid x_i)\, dy - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^n f_\beta^\alpha(Y_i \mid x_i). \tag{4.2}$$

Observe that this tends almost surely to $\mathrm{E}_Q \int \{ f_\beta^{1+\alpha}(y \mid x) - (1 + \alpha^{-1}) f_\beta^\alpha(y \mid x) g(y \mid x) \}$ dy. But this means that $H_n(\beta)$ plus the term $\alpha^{-1} \mathrm{E}_Q \int g^{1+\alpha}(y \mid x)\, dy$, which is parameter-independent, tends to the natural overall divergence measure

$$D_\alpha[\text{truth}, \text{model}] = \int d_\alpha[g(\cdot \mid x), f_\beta(\cdot \mid x)]\, Q(dx). \tag{4.3}$$

4·2. Large-sample behaviour

Results from §2 can be generalised to the present setting, under mild regularity conditions. The first result is that $\widehat{\beta}$ tends in probability to the least false parameter β_0 that minimises (4.3). Note that what is the 'best parametric approximation' $f_{\beta_0}(y \mid x)$ actually depends not only on the real $g(y \mid x)$ but also on the distribution of covariates. The Q distribution is irrelevant only if the model is correct.

Next consider the limit distribution of $\widehat{\beta}$. This involves the model score function $u_\beta(y \mid x)$ which is now $\partial \log f_\beta(y \mid x)/\partial\beta$ and the model information function $i_\beta(y \mid x) = -\partial^2 \log f_\beta(y \mid x)/\partial\beta\partial\beta^T$. The vector of first derivatives of $H_n(\beta)$, modulo a multiplicative constant that we remove, is

$$U_n(\beta) = n^{-1} \sum_{i=1}^n u_\beta(y_i \mid x_i) f_\beta^\alpha(y_i \mid x_i) - n^{-1} \sum_{i=1}^n \int u_\beta(y \mid x_i) f_\beta^{1+\alpha}(y \mid x_i)\, dy.$$

And the second order derivatives are

$$\begin{aligned} I_n(\beta) \;=\; & n^{-1} \sum_{i=1}^n \{ \alpha u_\beta(y_i \mid x_i) u_\beta(y_i \mid x_i)^T - i_\beta(y_i \mid x_i) \} f_\beta^\alpha(y_i \mid x_i) \\ & - n^{-1} \sum_{i=1}^n \int \{ (1+\alpha) u_\beta(y \mid x_i) u_\beta(y \mid x_i)^T - i_\beta(y \mid x_i) \} f_\beta^{1+\alpha}(y \mid x_i)\, dy. \end{aligned}$$

Of course, $U_n(\widehat{\beta}) = 0$. Also, note that $U_n(\beta_n)$ has mean zero, where β_n minimises (4.3) when Q is the empirical distribution of the covariates. Consider the variance matrix of

$\sqrt{n}\, U_n(\beta_n)$, conditionally on the x_i's. This is $K_n(\beta_n)$, where

$$K_n(\beta) = n^{-1}\sum_{i=1}^n \int u_\beta(y\,|\,x_i)u_\beta(y\,|\,x_i)^T g(y\,|\,x_i) f_\beta^{2\alpha}(y\,|\,x_i)\,dy - n^{-1}\sum_{i=1}^n \xi_i\xi_i^T$$

and $\xi_i = \int u_\beta(y\,|\,x_i)g(y\,|\,x_i)f_\beta^\alpha(y\,|\,x_i)\,dy$. To save space, we shall not be explicit about regularity conditions in this section, but we shall assume sufficient conditions to be in force to ensure (i) that $-I_n(\widetilde{\beta}_n)$ tends in probability to a positive definite J, for each sequence $\widetilde{\beta}_n$ such that $\widetilde{\beta}_n - \beta_n$ goes to zero in probability; (ii) that the matrix $K_n(\beta_n)$ tends in probability to a positive definite K; and (iii) that $\sqrt{n}\, U_n(\beta_n)$ tends in distribution to $\mathcal{N}(0,K)$. It then follows that

$$\sqrt{n}(\widehat{\beta} - \beta_n) \to_d \mathcal{N}(0, J^{-1}KJ^{-1}). \tag{4.4}$$

These regularity requirements are not strict. The first and second essentially involve the law of large numbers, with some extra continuity and/or uniformity required for the first, while the third holds under Lindeberg type circumstances.

The matrices in (4.4) can be expressed in terms of expectations with respect to the covariate. In fact

$$\begin{aligned} J \;=\; & \mathrm{E}_Q \int u_{\beta_0}(y\,|\,x)u_{\beta_0}(y\,|\,x)^T f_{\beta_0}^{1+\alpha}(y\,|\,x)\,dy \\ & + \alpha\,\mathrm{E}_Q \int u_{\beta_0}(y\,|\,x)u_{\beta_0}(y\,|\,x)^T\{f_{\beta_0}^{1+\alpha}(y\,|\,x) - g(y\,|\,x)f_{\beta_0}^\alpha(y\,|\,x)\}\,dy \\ & - \mathrm{E}_Q \int i_{\beta_0}(y\,|\,x)\{f_{\beta_0}^{1+\alpha}(y\,|\,x) - g(y\,|\,x)f_{\beta_0}^\alpha(y\,|\,x)\}\,dy \end{aligned}$$

and

$$K = \mathrm{E}_Q \int u_{\beta_0}(y\,|\,x)u_{\beta_0}(y\,|\,x)^T g(y\,|\,x) f_{\beta_0}^{2\alpha}(y\,|\,x)\,dy - \mathrm{E}_Q\xi\xi^T.$$

where $\xi = \int u_{\beta_0}(y\,|\,x)g(y\,|\,x)f_{\beta_0}^\alpha(y\,|\,x)\,dy$.

For small α the (4.3) divergence is close to a Q-weighted version of x-conditional Kullback–Leibler divergence, which corresponds to ordinary maximum likelihood analysis. With $\alpha = 1$ the method would correspond to a form of L_2 regression analysis. Also note that when all covariates are equal the estimation methods and performance results of Section 2 are essentially retrieved.

Estimators are necessary for J and K in order to carry out inference, and such are readily constructed. Considering K first, estimate integrals with respect to $g(y\,|\,x)$ using averaging over y_j's. For example,

$$\widehat{K} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n u_{\widehat{\beta}}(y_j\,|\,x_i)u_{\widehat{\beta}}(y_j\,|\,x_i)^T f_{\widehat{\beta}}^{2\alpha}(y_j\,|\,x_i) - n^{-1}\sum_{i=1}^n \widehat{\xi}_i\widehat{\xi}_i^T,$$

where $\widehat{\xi}_i = n^{-1}\sum_{j=1}^n u_{\widehat{\beta}}(y_j \mid x_i) f_{\widehat{\beta}}^{\alpha}(y_j \mid x_i)$. For J there are a couple of natural possibilities. One proposal stems from looking at J as the limit in probability of $-\mathrm{E} I_n(\beta_0)$. Note that these estimators are nonparametric and model-robust; their construction does not require the parametric model to hold. The important point is that $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$ will converge in probability to the real limiting variance matrix for $\sqrt{n}\widehat{\beta}$, even outside parametric model conditions. Again, there are also other ways of estimating the variance via jackknifing and bootstrapping.

The simplifications under model conditions are that

$$J = \mathrm{E}_Q \int u_{\beta_0}(y \mid x) u_{\beta_0}(y \mid x)^T f_{\beta_0}^{1+\alpha}(y \mid x)\, dy$$

while

$$K = \mathrm{E}_Q \int u_{\beta_0}(y \mid x) u_{\beta_0}(y \mid x)^T f_{\beta_0}^{1+2\alpha}(y \mid x)\, dy - \mathrm{E}_Q \xi \xi^T,$$

where $\xi = \int u_{\beta_0}(y \mid x) f_{\beta_0}^{1+\alpha}(y \mid x)\, dy$.

4·3. Example: robust linear regression

Take the standard linear regression model $y_i = x_i^T\beta + \sigma e_i$, where β and the x_i's are p-dimensional and the e_i's are independent standard normals. The minimum d_α method is to solve the $p+1$ equations

$$n^{-1}\sum_{i=1}^n \phi^\alpha(\widehat{e}_i)\widehat{e}_i x_i = 0, \qquad n^{-1}\sum_{i=1}^n \phi^\alpha(\widehat{e}_i)(\widehat{e}_i^2 - 1) = \int (v^2-1)\phi^{1+\alpha}(v)\, dv$$

where $\widehat{e}_i = (y_i - x_i^{\mathrm{t}}\widehat{\beta})/\widehat{\sigma}$ and $\phi = \phi_{0,1}$. A suitably engineered iterative computational scheme will find the solutions $(\widehat{\beta}, \widehat{\sigma})$.

Using the large-sample results above we may derive the approximate distribution of the estimators. For illustrational purposes we are content here to give the results under the model conditions of linearity and normality. Calculations, not given, show that the variance matrix in the approximating normal distribution for $\widehat{\beta}$ is

$$\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/2} \Big(\sum_{i=1}^n x_i x_i^T\Big)^{-1} \sigma^2.$$

This is the natural analogue of the result for the normal location parameter in (3.3). The efficiency relative to the maximum likelihood estimator is the same as discussed there (see, in particular, Table 1). Likewise, for σ, we find the same efficiency figures as in the normal scale model as in (c) of §3.1; again, see Table 1.

The density power divergence methodology also extends directly to, for example, robust Poisson regression. The model choice methodology of §2.6 can also with some effort be generalised to the present framework with covariates, resulting in a version of (2.12).

5. Concluding remarks

This paper has introduced a general family of divergences, indexed by a parameter α, which generates a corresponding family of estimators. This family includes maximum likelihood estimation as the limiting case of $\alpha = 0$. It is shown that increasing α leads to estimators which are far more robust than the maximum likelihood estimators, and have little loss in efficiency. Several examples suggest that an α of between 0.1 and 0.25 will work well. The method can be applied to any parametric family, and also to models with covariates, as the extension to regression situations shows. One of the main advantages of this family of divergences over other proposed families such as the Hellinger distance is that no smoothing of the empirical density function is needed in the case of continuous densities.

There can be no universal way of selecting an appropriate α parameter when applying our estimation methods. It specifies the underlying distance measure and typically dictates to what extent the resulting methods become more statistically robust than the maximum likelihood methods, and should be thought of as an algorithmic parameter. One way of selecting it is to fix the efficiency loss, at the ideal parametric model employed, at some low level, like five or ten percent. A related idea is to fix the maximum level of the influence curve at some acceptable level. Other ways could in some practical applications involve prior notions of the extent of contamination of the model.

Acknowledgements

The authors would like to thank Professor Probal Chaudhuri for helpful comments.

References

Basu, A. & Lindsay, B.G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **48**, 683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445–463.

Brown, L.D. & Hwang, J.T.G. (1993). How to approximate a histogram by a normal density. *Amer. Statist.* **47**, 251–255.

Cao, R., Cuevas, A. & Fraiman, R. (1995). Minimum distance density-based estimation. *Comput. Statist. Data Anal.* **20**, 611–631.

Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc.* Ser. B **46**, 440–464.

Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J. & Stahel, W.A. (1986). *Robust Statistics; the Approach Based on Influence Functions.* New York: Wiley.

Hjort, N.L. (1994). Minimum L_2 and robust Kullback–Leibler estimation. In *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, eds. P. Lachout and J.Á. Víšek, pp. 102–105. Prague: Academy of Sciences of the Czech Republic.

Jones, M.C. & Hjort, N.L. (1994). Comment on Brown & Hwang (1993). *Amer. Statist.* **48**, 353–354.

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lehmann, E.L. (1983). *Theory of Point Estimation.* New York: Wiley.

Lindsay, B.G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–1114.

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the Practice of Statistics*, 2nd ed. New York: W.H. Freeman.

Neyman, J. (1949). Contribution to the theory of χ^2 tests. In *Proceedings of the First Berkeley Symposium in Mathematics and Statistics*, pp. 239–273. Berkeley, CA: University of California Press.

Rao, C.R. (1963). Criteria of estimation in large samples. *Sankhya* Ser. A **25**, 189–206.

Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82**, 802–807.

Simpson, D.G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *J. Amer. Statist. Assoc.* **84**, 107–113.

Stigler, S.M. (1977). Do robust estimators work with real data? (with discussion). *Ann. Statist.* **5**, 1055–1098.

Tamura, R.N. & Boos, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.* **81**, 223–239.

Terrell, G.R. (1993). Spline density estimates. *Amer. Statist. Assoc. Proc. Statist. Comput. Sec.*, 255–260.

Welch, W.J. (1987). Rerandomizing the median in matched-pairs designs. *Biometrika* **74**, 609–614.

Windham, M.P. (1995). Robustifying model fitting. *J. Roy. Statist. Soc.* Ser. B **57**, 599–609.

FIGURE LEGENDS

Fig. 1. Influence functions for estimation of a normal mean (known variance) for various choices of α.

Fig. 2. A histogram of the Newcomb data with superimposed normal densities fitted using density power divergence parameter estimation with various values of α.

Table 1: Asymptotic relative efficiencies of the density power divergence estimators.

Model α:	0.00	0.02	0.05	0.10	0.25	0.50	1.00
Normal μ	1.000	0.999	0.997	0.988	0.941	0.838	0.650
Normal σ	1.000	0.999	0.993	0.976	0.888	0.731	0.541
Exponential (θ)	1.000	0.998	0.991	0.968	0.858	0.684	0.509
Poisson$(\lambda = 3)$	1.000	0.999	0.997	0.988	0.944	0.850	0.679
Poisson$(\lambda = 10)$	1.000	0.999	0.997	0.988	0.941	0.840	0.656

Table 2: Estimated parameters for the Newcomb data under the normal model.

α	0.00	0.02	0.05	0.10	0.25	0.50	1.00
$\widehat{\mu}$	26.21	26.74	27.44	27.60	27.64	27.52	27.29
$\widehat{\sigma}$	10.66	8.92	5.99	5.39	5.04	4.90	4.67

Table 3: Estimated parameters for the drosophila data under the Poisson model.

α	0.00	0.001	0.01	0.02	0.05	0.10	0.25	0.50	1.00
$\widehat{\lambda}$ (all observations)	3.059	2.056	0.447	0.394	0.393	0.392	0.386	0.374	0.365
$\widehat{\lambda}$ (outlier deleted)	0.394	0.394	0.394	0.393	0.392	0.390	0.382	0.366	0.349

Table 4: The telephone line fault data. The observations represent the inverse fault rate differences in 14 pairs of areas ((test − control) $\times 10^5$).

−988	−135	−78	3	59	83	93	110	189	197	204	229	269	310

Table 5: Test statistics and p-values for the Wald type test for the telephone fault data.

α	0.01	0.10	0.25	0.50	1.00
$\widehat{\mu}$	42.8	96.0	124.7	131.1	142.2
$\widehat{\sigma}$	305.6	209.2	133.4	136.9	139.5
statistic	0.52	1.65	3.24	3.20	3.30
p-value(normal)$\times 10^5$	30085	4960	60	68	48

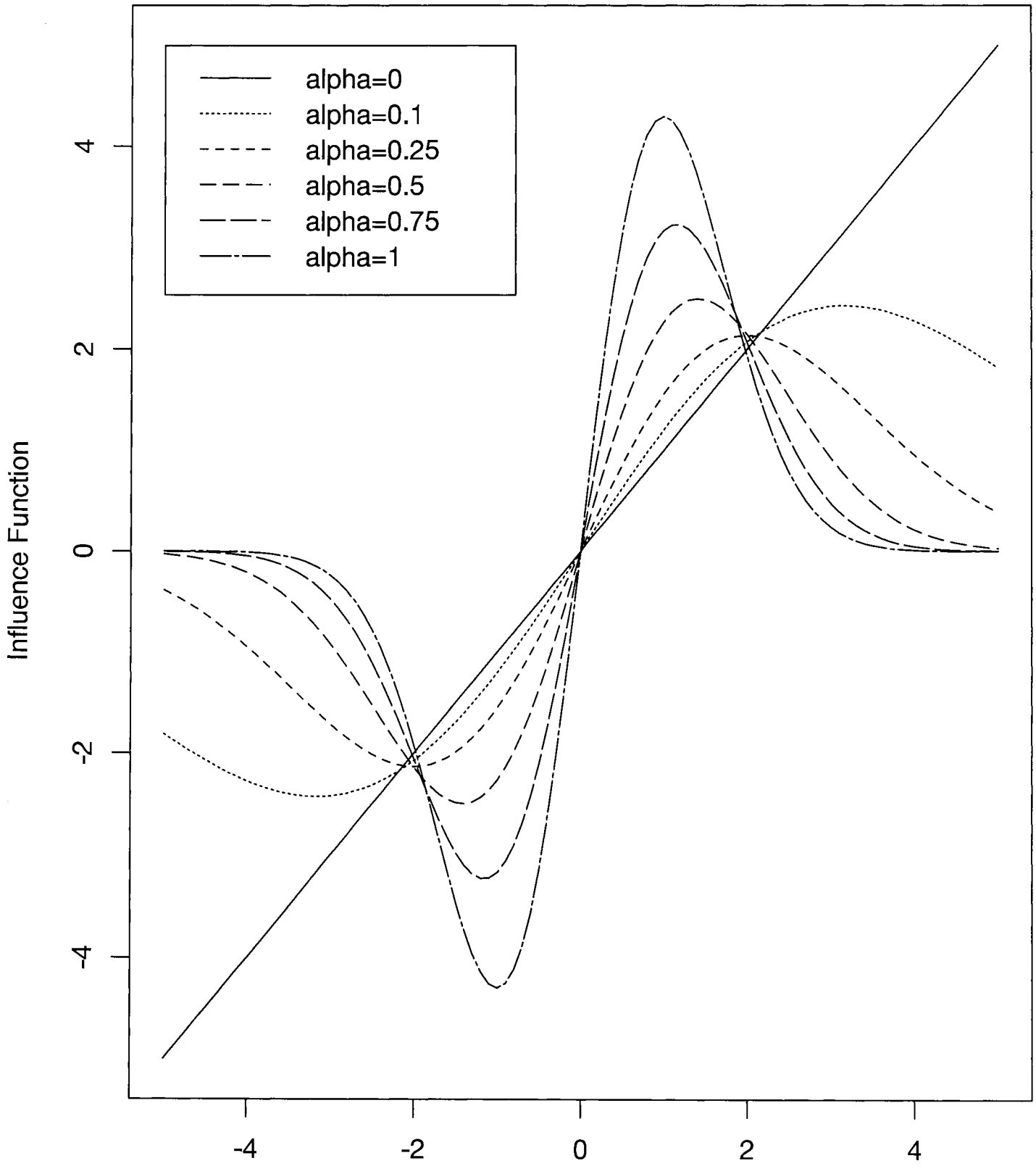
alpha=0
alpha=0.1
alpha=0.25
alpha=0.5
alpha=0.75
alpha=1
Influence Function
4
2
0
-2
-4
-4
-2
0
2
4
x

Figure 1

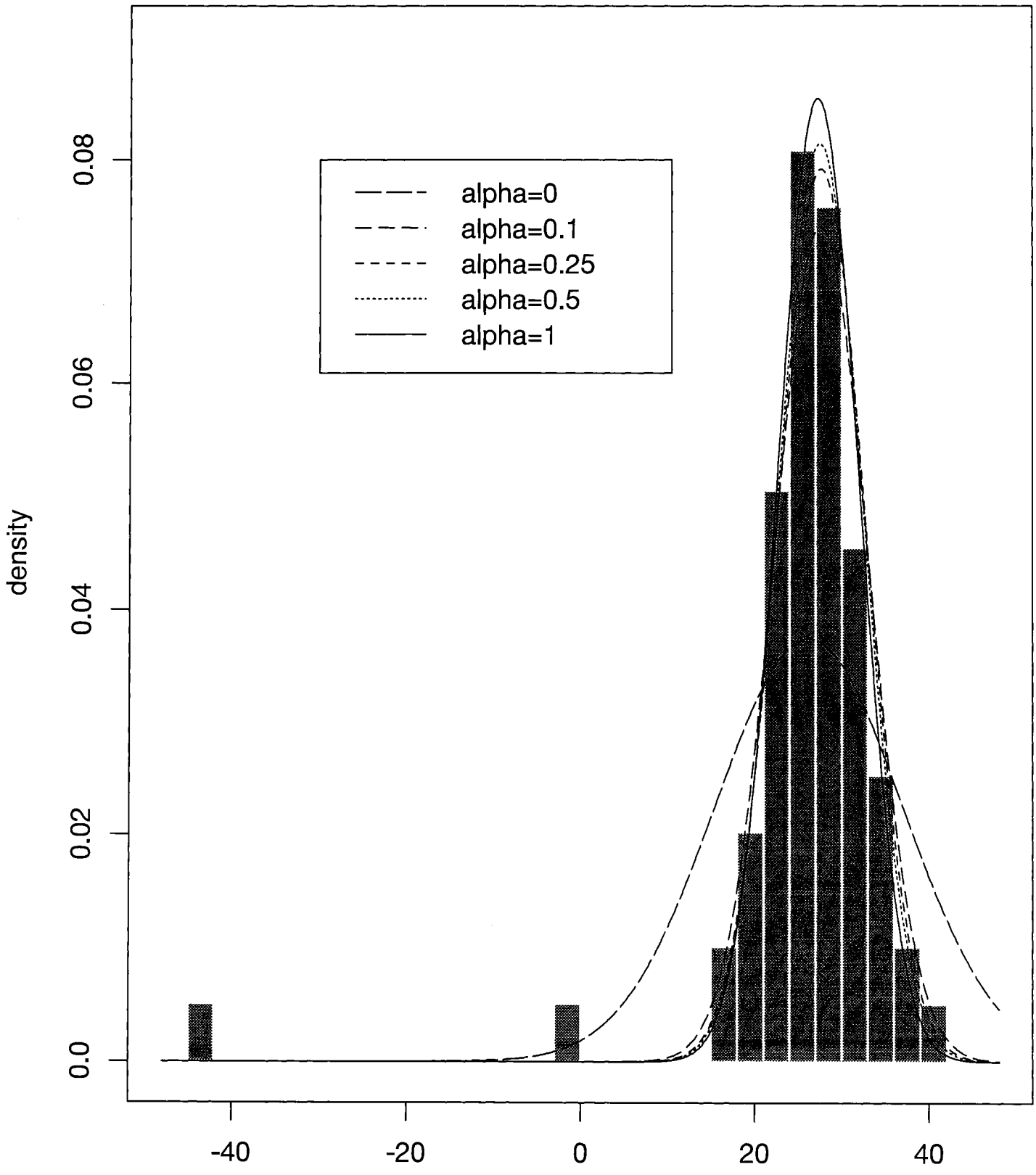
alpha=0
alpha=0.1
alpha=0.25
alpha=0.5
alpha=1
density
0.0
0.02
0.04
0.06
0.08
-40
-20
0
20
40
class intervals

Figure 2