A Comparison of Related Density-Based Minimum Divergence Estimators

M.C. JONES

Department of Statistics, The Open University,
Milton Keynes, MK7 6AA, United Kingdom
m.c.jones@open.ac.uk

NILS LID HJORT

Department of Mathematics, University of Oslo,
P.B. 1053 Blindern, N-0316 Oslo, Norway
nils@math.uio.no

IAN R. HARRIS

Department of Mathematics and Statistics, Northern Arizona University,
Flagstaff, Arizona 86011, U.S.A.
Ian.Harris@nau.edu

AND AYANENDRANATH BASU

Applied Statistics Unit, Indian Statistical Institute,
203 B.T. Road, Calcutta 700 035, India
ayanbasu@isical.ernet.in

SUMMARY

This paper compares the minimum divergence estimator of Basu, Harris, Hjort and Jones (1998) to a competing minimum divergence estimator which turns out to be equivalent to a method proposed from a different perspective by Windham (1995). Both methods can be applied for any parametric model, contain maximum likelihood as a special case, and can be extended to the context of regression situations. Theoretical calculations are given to compare efficiencies under model conditions, and robustness properties are studied and compared. Overall the two methods are found to perform quite similarly. Some relatively small advantages of the former method over the latter are identified.

Some key words: asymptotic relative efficiency; divergences; influence functions; M-estimation; maximum likelihood; robustness

1. Introduction

In Basu, Harris, Hjort & Jones (1998) (henceforth BHHJ), we introduced a new class of minimum divergence parameter estimators. This class, based on 'density power divergences' indexed by $\alpha \geq 0$, includes maximum likelihood estimation as the limiting case as $\alpha \to 0$. As α increases the members of the class exhibit reduced efficiency and increased robustness. Remarkably, quite small values of α were found to afford considerable robustness while retaining very high efficiency relative to maximum likelihood. As their name suggests, these estimators involve divergences between data and model densities (or probability mass functions in the case of discrete random variables) rather than distribution functions. The divergences are estimated without doing any smoothing of the data. The difficult problem of selecting the smoothing parameter, the bane of earlier attempts at density-based minimum divergence estimation, is thus avoided. The methodology is described in §2.1.

In this paper we introduce a natural alternative class of density-based minimum divergence estimators and compare it with the class of estimators introduced in BHHJ. The alternative estimators are essentially the same as those of Windham (1995) who suggested a new approach to robust model fitting. Windham considered weighting the data using weights 'proportional to a power of the density'. In the spirit of method-of-moments estimation, Windham proposed solving for the unknown parameter by equating sample and theoretical moments based on the weighted data. The related class of estimators introduced in the present paper is equivalent to those of Windham when utilising the likelihood score function in place of more arbitrary moment choices. We will show in §2.2 how this class can also be interpreted in minimum divergence terms. This introduction of the divergence function and the interpretation of Windham's method as an optimisation procedure (one that minimises the above divergence) is one of the principal contributions of this paper. Having the divergence then allows one to judge different solutions to Windham's equations. It, too, covers maximum likelihood as a special limiting case and is otherwise more robust but less efficient (at the model). In fact, we will see that the new class of estimators and the one proposed in BHHJ are closely related, yet different (in general). The two divergence families which generate these estimators are shown to be special cases of a larger family of divergences, suggesting in particular that the tuning parameter for the Windham method, say β, can be interpreted as being equivalent to the tuning parameter α for the BHHJ method.

A comparison of the two estimator classes is made in §3. For a given value of α, the BHHJ estimator is seen to be at least as efficient as the Windham estimator. The picture is less clear in terms of robustness. The BHHJ estimator usually is better in terms of breakdown and influence, but typically worse when compared to Windham's using say mean squared error under a contaminated model. The differences tend to be small. Implementation of the two methods is also very similar. In the case of the exponential distribution (discussed in §3.2.3) we recommend the BHHJ estimator, in the other cases we have seen there seems little reason to prefer one over the other.

We indicate in §4 how the estimation methods can be extended from the i.i.d. setting to general regression models. Discussion and some concluding remarks are offered in §5.

2. Density power divergences and corresponding estimators

2·1. *The BHHJ approach*

Consider a parametric family of models $\{F_t\}$, indexed by the unknown finite-dimensional parameter t in an open connected subset Ω of a suitable Euclidean space, possessing densities $\{f_t\}$ with respect to a common dominating measure which we for notational convenience take to be Lebesgue measure. Let G be the distribution underlying the data, having density g with respect to the same measure. BHHJ define the density power divergence between g and f_t to be

$$d_\alpha(g, f_t) = \int \Big\{ f_t^{1+\alpha} - \Big(1 + \frac{1}{\alpha}\Big) g f_t^\alpha + \frac{1}{\alpha} g^{1+\alpha} \Big\} \,\mathrm{d}z \quad \text{for } \alpha > 0 \tag{2.1}$$

and

$$d_0(g, f_t) = \lim_{\alpha\to 0} d_\alpha(g, f_t) = \int g \log(g/f_t) \,\mathrm{d}z. \tag{2.2}$$

Here and in many integral expressions in this paper we omit the variable of integration for convenience. Note that $d_0(g, f_t)$ is the Kullback–Leibler divergence. The version of the parameter which gives the best fit in terms of the density power divergence will be denoted θ_α, assumed to exist and be unique, and given by $d_\alpha(g, f_{\theta_\alpha}) = \min_{t\in\Omega} d_\alpha(g, f_t)$.

Both (2.1) and (2.2) involve g through only (i) a term in g alone which can be dropped because it does not affect minimisation over t, and (ii) a further term which is a linear functional of g, i.e. of the form $\int a(z) g(z) \,\mathrm{d}z$ for known a. Such a linear functional can be estimated via the empirical distribution of the data, as a sample average $\int a(z) \,\mathrm{d}\widehat{G}(z) = n^{-1} \sum_{i=1}^n a(X_i)$, where $\widehat{G}$ is the empirical distribution function. (Indeed, existing theory, see e.g. Silverman & Young, 1987 and de Angelis & Young, 1992, shows little or no advantage in introducing smoothing for such functionals, except perhaps for particular a functions and for small n.) The resulting sample minimum density power divergence estimators are those values $\widehat{\theta}_\alpha$ generated by minimising ($1/\alpha$ times)

$$\alpha \int f_t^{1+\alpha} \,\mathrm{d}z - (1 + \alpha) n^{-1} \sum_{i=1}^n f_t(X_i)^\alpha \tag{2.3}$$

with respect to t, when $\alpha > 0$, and the negative log-likelihood $-n^{-1}\sum_{i=1}^{n}\log f_t(X_i)$ when $\alpha = 0$. Differentiating with respect to t, $\widehat{\theta}_\alpha$ can also be defined by the robustified score equation

$$n^{-1}\sum_{i=1}^{n} f_t(X_i)^\alpha u_t(X_i) - \int f_t^{1+\alpha} u_t \,\mathrm{d}z = 0 \tag{2.4}$$

when $\alpha > 0$, or the ordinary score equation when $\alpha = 0$. Here, $u_t(z) = \partial \log f_t(z)/\partial t$ is the likelihood score function.

It is interesting to note that (2.1) is an example of a so-called Bregman divergence, discussed in Havrda and Charvat (1967), Burbea and Rao (1982), Jones and Trutzer (1989), Jones and Byrne (1990) and Csiszar (1991). From Csiszar (1991), a Bregman divergence is one taking the form

$$\int [H(g(z)) - H(f(z)) - \{g(z) - f(z)\}H'(f(z))]\,\mathrm{d}z,$$

where H is a convex function. Taking $H(f) = f^{1+\alpha}$ gives α times (2.1). It is also interesting to note that no smoothing is needed to implement an estimation strategy based on any Bregman divergence. One can, however, make the following argument to suggest that the only statistically interesting case is given by (2.1). A Taylor series expansion of the Bregman integrand, for f close to g, gives $\frac{1}{2}(f-g)^2 H''(f)$. If we want the divergence to behave like the usual kind of weighted mean integrated squared error which includes Kullback–Leibler and L2 divergences, the case $\alpha = 1$ in (2.1), then we need $H''(f) \propto f^{\alpha-1}$, for some $\alpha \geq 0$. Thus $H(f) \propto f^{\alpha+1}$, leading to the divergence given by (2.1).

2·2. The alternative approach

Windham's (1995) approach is essentially to choose the estimator, say $\widehat{\theta}_\beta^{(0)}$, to solve the following equation in t:

$$\frac{\sum_{i=1}^{n} f_t(X_i)^\beta u_t(X_i)}{\sum_{i=1}^{n} f_t(X_i)^\beta} = \frac{\int f_t^{1+\beta} u_t \,\mathrm{d}z}{\int f_t^{1+\beta}\,\mathrm{d}z}. \tag{2.5}$$

The indexing parameter β equals c in Windham's notation, and we will see later that it is reasonable to take β as being equal to α. This is essentially equation (3) of Windham (1995) except that Windham allows the choice of functional, which we take to be the expectation of the score function, to be an arbitrary expected value. It can be argued quite generally that if one wishes to relate a general parameter estimation method to likelihood estimation, one should incorporate the likelihood score function in some appropriate way, while if one concentrates on generalised method-of-moments estimation, which is what Windham does, the likelihood score is replaced by an appropriate power of z. Of course, for many important models (e.g. the normal) the two approaches coincide. In (2.5), $\beta = 0$

corresponds immediately to maximum likelihood estimation. Equation (2.5) may also be written

$$\int f_t^{1+\beta}\,\mathrm{d}z\, n^{-1}\sum_{i=1}^{n} f_t(X_i)^\beta u_t(X_i) - \int f_t^{1+\beta} u_t\,\mathrm{d}z\, n^{-1}\sum_{i=1}^{n} f_t(X_i)^\beta = 0. \tag{2.6}$$

Clearly, both (2.4) and (2.6) are unbiased estimating equations when $g = f_t$.

If f_t is a location family, then (2.4) and (2.6) both reduce to

$$\sum_{i=1}^{n} f_t(X_i)^\gamma u_t(X_i) = 0, \tag{2.7}$$

where γ is α or β, and thus for this special case both approaches are identical. Equation (2.7) displays an interesting density power downweighting (and hence robustification) of the usual likelihood estimating equation. In general, however, (2.6) does not reduce to (2.4).

We recognise equation (2.6) as being the estimating equation resulting from maximising

$$\Big\{n^{-1}\sum_{i=1}^{n} f_t(X_i)^\beta\Big\}^{1+\beta} \Big/ \Big(\int f_t^{1+\beta}\,\mathrm{d}z\Big)^\beta \tag{2.8}$$

with respect to t. A further informative version of this is that $\widehat{\theta}_\beta^{(0)}$ minimises

$$\beta \log\Big(\int f_t^{1+\beta}\,\mathrm{d}z\Big) - (1+\beta)\log\Big\{n^{-1}\sum_{i=1}^{n} f(X_i)^\beta\Big\}. \tag{2.9}$$

This is a direct analogue of (2.3). What is more, (2.8) is the natural estimator of

$$\Big(\int f_t^\beta g\,\mathrm{d}z\Big)^{1+\beta} \Big/ \Big(\int f_t^{1+\beta}\,\mathrm{d}z\Big)^\beta. \tag{2.10}$$

Thus, as θ_α minimises $\alpha \int f_t^{1+\alpha}\,\mathrm{d}z - (1+\alpha)\int f_t^\alpha g\,\mathrm{d}z$, so the Windham method's best-fitting parameter $\theta_\beta^{(0)}$ maximises (2.10) or equivalently minimises $\beta\log(\int f_t^{1+\beta}\,\mathrm{d}z) - (1+\beta)\log(\int f_t^\beta g\,\mathrm{d}z)$.

It is not difficult to convert the objective function (2.10) into a genuine discrepancy. One version of this is given by

$$\frac{1}{\beta}\left\{\int g^{1+\beta}\,\mathrm{d}z - \frac{\left(\int f_t^\beta g\,\mathrm{d}z\right)^{1+\beta}}{\left(\int f_t^{1+\beta}\,\mathrm{d}z\right)^\beta}\right\}. \tag{2.11}$$

That (2.11) has the required properties of being positive for all $\beta > 0$ with equality if and only if $f_t = g$ follows as a special case of Hölder's inequality. Indeed, alternative versions of this discrepancy also yield the exact same $\theta_\beta^{(0)}$. For instance, one might prefer

$$\frac{1}{\beta}\{1-\rho_\beta(g,f_t)\} = \frac{1}{\beta}\left\{1 - \frac{\int f_t^\beta g\,\mathrm{d}z}{(\int g^{1+\beta}\,\mathrm{d}z)^{1/(1+\beta)}\left(\int f_t^{1+\beta}\,\mathrm{d}z\right)^{\beta/(1+\beta)}}\right\}, \tag{2.12}$$

where the ρ_β measure has a 'correlation' nature, with maximal value 1 only if f_t agrees fully with g. A further form of the divergence which is similar to (2.1) is

$$d_\beta^{(0)}(g,f_t) = \log\Big(\int f_t^{1+\beta}\,\mathrm{d}z\Big) - \Big(1+\frac{1}{\beta}\Big)\log\Big(\int g f_t^\beta\,\mathrm{d}z\Big) + \frac{1}{\beta}\log\Big(\int g^{1+\beta}\,\mathrm{d}z\Big). \tag{2.13}$$

In addition to the above arguments, one may derive both (2.1) and (2.13) from a more general family of divergences, given by

$$\phi^{-1}\Big(\int f_t^{1+\gamma}\,\mathrm{d}z\Big)^\phi - \frac{1+\gamma}{\gamma}\phi^{-1}\Big(\int f_t^\gamma g\,\mathrm{d}z\Big)^\phi + \frac{1}{\gamma}\phi^{-1}\Big(\int g^{1+\gamma}\,\mathrm{d}z\Big)^\phi. \tag{2.14}$$

That (2.14) is a divergence can be shown via the Hölder inequality and some additional analysis; see the Appendix. Selecting $\phi = 1$ immediately gives (2.1) (with γ playing the role of α in (2.1)). To obtain (2.13) (with β the same as γ), take the limit as ϕ goes to zero, and do some simple algebraic manipulations. Since both divergences are embedded in the same family, α in (2.1) and β in (2.13) may be treated as the same parameter. Thus comparisons between (2.1) and (2.13) really represent comparisons between $\phi = 1$ and $\phi = 0$ in (2.14). The remark at the end of this section suggests that these two values of ϕ are the statistically most useful ones.

The calibration $\alpha = \beta$ and the choice $\phi = 0$ or 1 are also supported by the following asymptotic argument. It can readily be shown that for f_t close to g, both (2.1) and (2.11) are close to, for fixed $\gamma = \alpha = \beta$,

$$\tfrac{1}{2}(1+\gamma)\int g^{\gamma-1}(f_t-g)^2\,\mathrm{d}z.$$

This is not true for fixed $\phi \neq 0$ or 1 in (2.14), for which an extra term is present. Additionally, for fixed f_t, it can be shown that for $\gamma \to 0$, both (2.1) and (2.11) take the form

$$\mathrm{KL}(g,f_t) + \gamma\left[\int\left\{(f_t-g)\left(\log f_t + \frac{1}{2f_t}\right) + \tfrac{1}{2}g\left(\log^2 g - \log^2 f_t\right)\right\}\mathrm{d}z\right] + O(\gamma^2),$$

where $\mathrm{KL}(g,f_t)$ is the Kullback–Leibler divergence from g to f_t. The approximations in this paragraph hold provided the integrals are finite.

There are further correspondences between the two approaches in this paper. Windham (1995) considers the weighted density

$$\mathrm{d}G_{\beta,t}(x) = \frac{f_t^\beta(x)}{\mathrm{E}_G\{f_t^\beta(x)\}}\,\mathrm{d}G(x) \tag{2.15}$$

and takes $\widehat{\theta}_\beta^{(0)}$ to satisfy $T((F_\theta)_{\beta,\theta}) = T(\widehat{G}_{\beta,\theta})$. With T the expectation of the score function, this yields (2.5). If instead we use

$$\mathrm{d}G_{\alpha,t}(x) = \frac{f_t^\alpha(x)}{\mathrm{E}_{F_t}\{f_t^\alpha(x)\}}\mathrm{d}G(x) \tag{2.16}$$

with T the expectation of the score function, then $T((F_\theta)_{\alpha,\theta}) = T((\widehat{G})_{\alpha,\theta})$ yields (2.4) (although we note that the right-hand side of (2.16) is not a density). Windham (1995) uses a fixed point algorithm motivated by (2.15), and shows that the convergence rate at the model and for large n is $\beta/(1+\beta)$. In the same way, using (2.16) we can define a fixed point algorithm for the BHHJ estimator which has convergence rate $\alpha/(1+\alpha)$.

Both (2.4) and (2.6) are special cases of M-estimation methodology, with estimating equations of the form $\sum_{i=1}^n \psi(X_i, \theta) = 0$ for appropriate functions ψ. This observation does not diminish the novelty and attractiveness of these minimum distance-based estimation procedures, but it does make some of the theory in the next section flow more or less directly from existing general M-estimation theory; see for example Hampel, Ronchetti, Rousseeuw & Stahel (1986).

In summary, in the same sense that BHHJ suggested a class of smoothing-free minimum density-based divergence estimators based on (2.1) by minimising (2.3), so, by the minimisation of (2.8) or equivalently (2.9), does (2.11) (or (2.12) or (2.13)) yield an alternative class of smoothing-free minimum density-based divergence estimators. In reference to the superfamily of divergences (2.14), we will refer to the BHHJ divergence (2.1) as type 1 (since $\phi = 1$) and the Windham-derived divergence (2.13) as type 0 (since $\phi = 0$). The corresponding estimators obtained from minimising these will then be type 1 and type 0 estimators.

Remark. One may turn the divergence (2.14) into an estimation method for any nonnegative value ϕ, by selecting $\widehat{\theta}$ to minimise

$$\Big(\int f_t^{1+\gamma}\,\mathrm{d}z\Big)^\phi - (1+1/\gamma)\Big\{n^{-1}\sum_{i=1}^n f_t(X_i)^\gamma\Big\}^\phi.$$

Taking the derivative with respect to the parameter and simplifying gives the estimating equation

$$\Big\{n^{-1}\sum_{i=1}^n f_t(X_i)^\gamma\Big\}^{\phi-1}\Big\{n^{-1}\sum_{i=1}^n f_t(X_i)^\gamma u_t(X_i)\Big\} = \Big(\int f_t^{1+\gamma}\,\mathrm{d}z\Big)^{\phi-1}\int f_t^{1+\gamma}u_t\,\mathrm{d}z.$$

One sees that (2.4) and (2.5) are indeed the special cases when ϕ is equal to 1 and 0. For other values of ϕ the above estimating equation is not unbiased for finite n, but is so for the asymptotic case. We have focussed on the two cases 1 and 0 only, however. For other values the method can not be represented as an M-estimation method, but consistency and asymptotic normality can be established under mild regularity conditions. See the appendix for the necessary tools.

3. Comparing the two methods

3·1. *Asymptotic efficiencies*

The two classes of estimators considered in this paper are the BHHJ method $\widehat{\theta}_\alpha$ (type 1) and Windham's method $\widehat{\theta}^{(0)}_\beta$ (type 0). Let $\theta_\alpha = T_\alpha(G)$ and $\theta^{(0)}_\beta = T^{(0)}_\beta(G)$ be the minimum disparity functionals obtained by minimising the divergence between the true density g and f_t for the two methods; g may not necessarily belong to the model family. Under regularity conditions, it can be shown that the $\widehat{\theta}_\alpha$ and $\widehat{\theta}^{(0)}_\beta$ are consistent for θ_α and $\theta^{(0)}_\beta$, respectively, and are asymptotically normal as n grows. In fact, for both methods, $n^{1/2}$ times estimator minus estimand is asymptotically a zero-mean multivariate normal with variance matrix of the form $J_\alpha^{-1} K_\alpha J_\alpha^{-1}$ for type 1 and $(J^{(0)}_\beta)^{-1} K_\beta (J^{(0)}_\beta)^{-1}$ for type 0. Formulae for these are given below. These results hold even if the data distribution G is not equal to F_t for any t.

3·1·1. *Formulae for variances*

Introduce

$$L_{j,\gamma} = \int g^j f_\theta^\gamma \,\mathrm{d}z, \quad M_{j,\gamma} = \int g^j f_\theta^\gamma u_\theta \,\mathrm{d}z \quad \text{and} \quad N_{j,\gamma} = \int g^j f_\theta^\gamma u_\theta u_\theta^{\mathrm{t}} \,\mathrm{d}z$$

for $j = 0, 1$ and positive γ, evaluated at $\theta = \theta_\alpha$ when type 1 is being discussed and at $\theta = \theta^{(0)}_\beta$ when type 0 is discussed. Under model conditions, $L_{0,1+\gamma} = L_{1,\gamma}$, and so on.

For the type 1 estimator, K_α and J_α are given in BHHJ and may be expressed as $K_\alpha = N_{1,2\alpha} - M_{1,\alpha} M^{\mathrm{t}}_{1,\alpha}$ and

$$J_\alpha = \int f_\theta^\alpha (g - f_\theta)(i_\theta - \alpha u_\theta u_\theta^{\mathrm{t}}) \,\mathrm{d}z + N_{0,1+\alpha},$$

where $i_t(x) = -\partial\{u_t(x)\}/\partial t$ is the (positive definite) information function of the model. Under model conditions, these matrices reduce to $J_\alpha = N_{0,1+\alpha}$ and $K_\alpha = N_{0,1+2\alpha} - M_{0,1+\alpha} M^{\mathrm{t}}_{0,1+\alpha}$.

To analyse behaviour of the type 0 estimator one may appeal to general M-estimation methodology. In the present case it is perhaps as easy and illuminating to derive the necessary result, along with expressions for the limiting variance matrix, directly. Such arguments are presented in the Appendix, and make it actually easier to check, for any parametric model at hand, whether there is sufficient regularity to secure the validity of the separate steps in the approximation arguments.

The results are as follows. Write ξ_β for the vector $M_{1,\beta}/L_{1,\beta}$, which is also identical to $M_{0,1+\beta}/L_{0,1+\beta}$; cf. eq. (2.5), and the appendix. One finds that

$$K_\beta^{(0)} = \frac{1}{L_{1,\beta}^2}\int g f_\theta^{2\beta}(u_\theta-\xi_\beta)(u_\theta-\xi_\beta)^{\mathrm{t}}\,\mathrm{d}z,$$

which may also be expressed as $(1/L_{1,\beta}^2)(N_{1,2\beta}-\xi_\beta M_{1,2\beta}^{\mathrm{t}}-M_{1,2\beta}\xi_\beta^{\mathrm{t}}+\xi_\beta\xi_\beta^{\mathrm{t}}L_{1,2\beta})$. Under model conditions,

$$K_\beta^{(0)} = (1/L_{0,1+\beta}^2)(N_{0,1+2\beta}-\xi_\beta M_{0,1+2\beta}^{\mathrm{t}}-M_{0,1+2\beta}\xi_\beta^{\mathrm{t}}+\xi_\beta\xi_\beta^{\mathrm{t}}L_{0,1+2\beta}).$$

Secondly,

$$\begin{aligned} J_\beta^{(0)} \;=\; & \frac{N_{0,1+\beta}}{L_{0,1+\beta}}-\frac{M_{0,1+\beta}}{L_{0,1+\beta}}\Big(\frac{M_{0,1+\beta}}{L_{0,1+\beta}}\Big)^{\mathrm{t}}+\beta\Big(\frac{N_{0,1+\beta}}{L_{0,1+\beta}}-\frac{N_{1,\beta}}{L_{1,\beta}}\Big)\\ & +(1/L_{1,\beta})\int g f_\theta^{\beta} i_\theta\,\mathrm{d}z-(1/L_{0,1+\beta})\int f_\theta^{1+\beta} i_\theta\,\mathrm{d}z. \end{aligned}$$

Under model conditions this reduces to $J_\beta^{(0)}=N_{0,1+\beta}/L_{0,1+\beta}-\xi_\beta\xi_\beta^{\mathrm{t}}$.

Note that as $\alpha=\beta\to 0$, writing K for K_α or $K_\beta^{(0)}$ and J for J_α or $J_\beta^{(0)}$,

$$K\to N_{1,0}=\int g u_\theta u_\theta^{\mathrm{t}}\,\mathrm{d}z,\quad J\to N_{0,1}+\int(g i_\theta - f_\theta i_\theta)\,\mathrm{d}z=\int g i_\theta\,\mathrm{d}z,$$

in agreement with traditional results about the limiting behaviour of maximum likelihood methods outside model conditions.

Statistical inference can be carried out in the form of tests, confidence statements, and so on as long as there is a consistent estimator of the variance matrix of the limiting distribution. Such can be arrived at in various ways. Model-robust estimates emerge for $J_\beta^{(0)}$ and $K_\beta^{(0)}$ when one replaces $L_{0,\gamma}$ and $L_{1,\gamma}$ in the formulae above with $\int f(x,\widehat\theta_\beta^{(0)})^\gamma\,\mathrm{d}x$ and $n^{-1}\sum_{i=1}^n f(X_i,\widehat\theta_\beta^{(0)})^\gamma$, respectively, and similarly with the other M- and N-quantities. The resulting variance matrix estimator may also be written

$$n^{-1}\sum_{i=1}^n \widehat I_i\widehat I_i^{\mathrm{t}},\quad \text{where } \widehat I_i=(\widehat J_\beta^{(0)})^{-1}\widehat L_{1,\beta}^{-1} f(Y_i,\widehat\theta_\beta^{(0)})^\beta\big\{u(Y_i,\widehat\theta_\beta^{(0)})-\widehat L_{1,\beta}^{-1}\widehat M_{1,\beta}\big\}.$$

These $\widehat I_i$ variables have separate interpretation as influences of the data points, and may be used for model-checking purposes. Yet other options exist for estimating the variance matrix, including bootstrapping and jackknifing.

3·1·2. Comparisons for small γ

In this subsection we investigate the large-sample variance matrices for $\gamma\equiv\alpha=\beta$ small, for general models. We work under model conditions, so that $g=f_\theta$, say.

For a general model for which the integrals below exist, consider

$$A_\gamma = \int f_\theta^{1+\gamma}\,\mathrm{d}z, \quad B_\gamma = \int f_\theta^{1+\gamma} u_\theta\,\mathrm{d}z, \quad C_\gamma = \int f_\theta^{1+\gamma} u_\theta u_\theta^{\mathrm{t}}\,\mathrm{d}z \tag{3.1}$$

for nonnegative γ, in terms of which

$$\begin{aligned} J_\gamma &= C_\gamma, \\ K_\gamma &= C_{2\gamma} - B_\gamma B_\gamma^{\mathrm{t}}, \\ J_\gamma^{(0)} &= (1/A_\gamma)(C_\gamma - B_\gamma B_\gamma^{\mathrm{t}}/A_\gamma), \\ K_\gamma^{(0)} &= (1/A_\gamma^2)\{C_{2\gamma} - \xi_\gamma^{(0)} B_{2\gamma}^{\mathrm{t}} - B_{2\gamma}(\xi_\gamma^{(0)})^{\mathrm{t}} + \xi_\gamma^{(0)}(\xi_\gamma^{(0)})^{\mathrm{t}} A_{2\gamma}\}, \end{aligned}$$

where $\xi_\gamma^{(0)} = B_\gamma/A_\gamma$. For small γ, the quantities in (3.1) may be approximated via $f_\theta^\gamma \doteq 1 + \gamma \log f_\theta + \frac{1}{2}\gamma^2(\log f_\theta)^2$, as Taylor expansions to the second order. Thus $C_\gamma \doteq J_0 + \gamma D + \frac{1}{2}\gamma^2 E$ for $D = \int f_\theta \log f_\theta u_\theta u_\theta^{\mathrm{t}}\,\mathrm{d}z$ and $E = \int f_\theta (\log f_\theta)^2 u_\theta u_\theta^{\mathrm{t}}\,\mathrm{d}z$, where J_0 is the Fisher information matrix of the model, with similar expansions to second order of A_γ and B_γ. There are consequent approximations for the variance matrices of the two limiting distributions. Interestingly, after some matrix algebra and analysis work one finds that both $J_\gamma^{-1} K_\gamma J_\gamma^{-1}$ (for type 1) and $(J_\gamma^{(0)})^{-1} K_\gamma (J_\gamma^{(0)})^{-1}$ (for type 0) are equal to

$$J_0^{-1} + \gamma^2 J_0^{-1}[\mathrm{Var}\{u_\theta(X)\log f_\theta(X)\} - 3DJ_0^{-1}D]J_0^{-1} + O(\gamma^3) \tag{3.2}$$

(the first order term vanishes). The matrix inside square brackets is always positive definite.

Equation (3.2) illustrates the relatively small loss of efficiency of both estimation methods, and also that the methods can be expected to perform very similarly, when their tuning parameters are equal and small. The third order terms (for γ^3) differ for the two methods. In all the examples we have investigated, the BHHJ method (type 1) has asymptotic variances smaller than or equal to those of the Windham method (type 0) for the same value of the tuning parameter $\alpha = \beta$, but we have not attempted to prove that this always holds.

3·1·3. Comparisons for some simple models

We next go on to inspect efficiencies for some simple models, again working under model conditions. First, for pure location models, recall that the two regimes agree. Results applicable to both estimators are therefore given for the normal mean model in §4.1(a) of BHHJ. For the normal standard deviation σ, the formula for the asymptotic variance of $n^{1/2}$ times $\widehat{\sigma}_\alpha$ is given in §4.1(b) of BHHJ, and that for $\widehat{\sigma}_\beta^{(0)}$ becomes

$$\frac{(1+\beta)^3(3\beta^2+4\beta+2)}{4(1+2\beta)^{5/2}}\sigma^2.$$

For the exponential distribution with mean θ, the formula for the asymptotic variance of $\widehat{\theta}_\alpha$ is given in §4.1(c) of BHHJ, and that for $\widehat{\theta}^{(0)}_\beta$ is after lengthy calculations found to be

$$\frac{(1+\beta)^4(2\beta^2+2\beta+1)}{(1+2\beta)^3}\theta^2.$$

Explicit expressions for the Poisson distribution are not available for either method but they can easily be evaluated numerically. Exact formulae can be found for the geometric distribution, where $f_\theta(x) = (1-\theta)^{x-1}\theta$ for $x = 1, 2, \ldots$. We omit giving these here, but illustrate their use in the table.

*** Table 1 about here ***

Numerical versions of the resulting efficiencies are presented in Table 1, given as ratios v_0/v, where v_0 is the minimum possible limiting variance (that of the maximum likelihood procedure) and v is the limiting variance for the estimator in question. It is striking that for small values of α, the two methods give almost identical efficiencies, as suggested by the (3.2) approximation. For larger α, the two typically diverge and the type 1 estimator becomes progressively more efficient than the type 0 estimator, although each is becoming very robust at the expense of rather considerable loss of efficiency.

One may note that there are situations where the limiting variance for $\widehat{\theta}_\alpha$ does not increase everywhere for increasing α. For the geometric distribution, for example, with $\theta > \frac{1}{2}$, the efficiency of the type 1 method first decreases with α and then increases; this is not quite visible from the last rows of Table 1 in that values are only displayed for a few α values ≤ 1.

3·2. Robustness

3·2·1. Influence and breakdown

As described in BHHJ, the influence function of $\widehat{\theta}_\alpha$ is

$$\mathrm{IF}(G, y) = J_\alpha^{-1}\Big\{u_{\theta_\alpha}(y) f_{\theta_\alpha}^\alpha(y) - \int u_{\theta_\alpha} f_{\theta_\alpha}^{1+\alpha}\,\mathrm{d}z\Big\}.$$

From the manipulations in §3.1, the influence function for the type 0 estimator can be expressed as

$$\begin{aligned}\mathrm{IF}^{(0)}(G, y) &= (J^{(0)}_\beta)^{-1}\frac{1}{L_{1,\beta}}\Big[\{f_\theta(y)^\beta u_\theta(y) - M_{1,\beta}\} - \frac{M_{1,\beta}}{L_{1,\beta}}\{f_\theta(y)^\beta - L_{1,\beta}\}\Big] \\ &= (J^{(0)}_\beta)^{-1} L_{1,\beta}^{-1} f_\theta(y)^\beta\{u_\theta(y) - \xi_\beta\}.\end{aligned}$$

These are typically bounded functions in y, in contrast to the influence function $J_0^{-1}u_{\theta_0}(y)$ for the maximum likelihood estimator, which is unbounded for most of the popular models.

For the normal $(0,\sigma^2)$ model, somewhat long calculations give $J_\beta^{(0)} = 2/(1+\beta)^2$, and

$$\mathrm{IF}^{(0)}(G,y) = \sigma\tfrac{1}{2}(1+\beta)^{5/2}\exp(-\tfrac{1}{2}\beta y^2/\sigma^2)\{y^2/\sigma^2 - 1/(1+\beta)\}$$

under model conditions. The corresponding influence function for the BHHJ method can be shown to take the form

$$\mathrm{IF}(G,y) = \sigma(1+\alpha)^{5/2}(2+\alpha^2)^{-1}\{\exp(-\tfrac{1}{2}\alpha y^2/\sigma^2)(y^2/\sigma^2-1) + \alpha/(1+\alpha)^{3/2}\}.$$

Figure 1 displays pairs of influence curves for the BHHJ (full line) and Windham methods (dotted line) for estimating σ in the normal $(0,\sigma^2)$ model, computed under model conditions and for $\sigma = 1$, for tuning parameters $\alpha = \beta$ equal to 0, 0.10, 0.25, 0.50. We note that the influence curves for the two methods are nearly identical for $\alpha = \beta \le 0.25$ and not very different for bigger tuning parameters either. For such moderate or larger tuning parameters, the Windham method's influence curve redescends slightly more quickly towards zero than does the BHHJ influence curve for values of the argument outside the region of values most probable under model conditions.

*** Figure 1 about here ***

Fig. 1. Pairs of influence curves for the BHHJ (full line) and Windham methods (dotted line) for estimating σ in the normal $(0,\sigma^2)$ model, for tuning parameter equal to 0, 0.10, 0.25, 0.50. The curves are computed under model conditions and for $\sigma = 1$.

BHHJ showed that for the normal case with unknown mean and variance, the breakdown point associated with $\widehat{\mu}_\alpha$ is $\alpha/(1+\alpha)^{3/2}$. Manipulations analogous to those in §3.2 of BHHJ show that the breakdown point associated with $\widehat{\mu}_\beta^{(0)}$ (in the normal case) is, on the other hand, zero. (This was indicated as a possibility by Windham, 1995.) This gives the BHHJ method a slight robustness edge over its competitor.

3·2·2. Mean squared error under contamination

In this section we simultaneously investigate efficiency and robustness by examining the mean squared error of each estimator under given contaminated models. Consider the mixture distribution $g(z) = (1-\epsilon)f_\theta(z) + \epsilon h(z)$, where $h(z)$ is some contaminating distribution. There is an asymptotic bias and an asymptotic variance for each of the type 0 and type 1 estimators under this true distribution if we fit the model family $\{f_t\}$ and θ is the target parameter. Alternatively, one can numerically calculate the mean squared error of each estimator in finite samples using a Monte Carlo approach. This can be done for various different choices of θ, contaminating distribution h, proportion of contamination ϵ and choice of γ. The purpose is to compare the two types of estimators at the same γ value and to compare different γ values for each estimator in an attempt to suggest which type of estimator to use and possibly which γ to pick.

We investigated three families; geometric, exponential and Poisson. For each family we considered two types of contamination, a point mass at a large value, and a contamination by the same type of distribution (i.e. Poisson with Poisson, geometric with geometric, etc.), with a substantially larger mean. We typically used ϵ around 0.1 and small sample sizes of about 20.

The results were mixed. For most of the simulations the optimal value of $\gamma = \alpha = \beta$ occurred around 0.5, although in one case (an exponential) the optimal value was about 0.95. Also in most cases the type 0 estimator tended to slightly outperform the type 1 estimator, although the difference was typically small, about 5 percent. Only in one case was the difference very large, again the exponential, where for γ near 1 the type 1 estimator produced mean squared errors almost 30 percent smaller than the type 0 mean squared error.

These results do not give very precise advice as to which γ value to use for any of the methods, unless one to some extent can assess the degree of contamination of one's data relative to the chosen model. The main points are (i) that the two methods again are seen to perform rather similarly, for the same value of the tuning parameter, and (ii) that there seems to be a reasonable range of close-to-optimal values of the tuning parameter where the results vary little.

3·2·3. The exponential distribution

On rare occasions the type 0 estimator may exhibit unexpected behaviour, when the parametric family used gives increased probability for data landing in a 'corner' of the sample space, and there are one or more data points extremely close to this corner. Thus the method may be robust for large outliers but not always for small outliers. We illustrate this phenomenon for the case of the exponential distribution.

Suppose that $X_1, \ldots, X_n$ are to be fitted by the $(1/\theta)\exp(-x/\theta)$ family. The estimator $\widehat\theta_\beta^{(0)}$ is the one minimising expression (2.9). It aims at and converges for growing n to the parameter value $\theta_\beta^{(0)}$ which minimises $Q(\theta) = \beta\log(\int f_\theta^{1+\beta}\,\mathrm{d}z) - (1+\beta)\log(\int f_\theta^\beta\,\mathrm{d}G)$, where G is the real mechanism generating data. Suppose that G has a point mass p at a small value x_0 with the remaining $1-p$ part being a unit exponential. The Windham type estimator then aims at minimising

$$Q(\theta) = \beta\log\Bigl(\int f_\theta^{1+\beta}\,\mathrm{d}z\Bigr) - (1+\beta)\log\Bigl\{pf_\theta(x_0)^\beta + (1-p)\int f_\theta(z)^\beta\exp(-z)\,\mathrm{d}z\Bigr\}.$$

In the present situation this simplifies to minimising

$$\beta \log \theta - (1+\beta) \log\{p \exp(-x_0\beta/\theta) + (1-p)\theta/(\beta+\theta)\}.$$

It is now easy to study this curve in θ for some combinations of p and β for a fixed small $x_0 = 0.001$, say. For β reasonably small there are no problems, and there is only one global minimiser, not far from $px_0 + 1 - p$, the mean of the true G (which would be the limit value of the maximum likelihood estimator). However, if the tuning parameter β as well as the contamination parameter p are a little larger than zero, problems may occur. For example, when $\beta = 0.5$ and $p = 0.15$ there is a global 'silly' minimum at $\theta_\beta = 0.0016$, and a more sensible local minimum at $\theta = 0.645$.

This behaviour translates to some rare categories of problematic finite-sample situations. If $\beta = 0.5$, say, and a significant proportion of the data values are very small, then $Q_n(\theta)$ may have two local minima, with the global one being unreasonably close to zero. We observed this in about 1 of 250 random samples of size $n = 20$, for example, with a single extremely small data value. With increasing n the problem goes away, unless a fixed proportion of data values remain very small.

What we learn is that in some cases the criterion function (2.9) may have two local minima, one extremely small, the other moderate. This may also happen in models other than the exponential. Interestingly, the type 1 estimator of BHHJ appears to be free of such problems, and could be preferred on this ground. A simple modification of Windham's type 0 method is to look into both minima and then pick the reasonable one. This may be formalised for example in terms of closeness to a crude, robust, $n^{1/2}$-consistent estimator (which in the one-parameter models could be tied to the median, for example).

3·3. Limitations when integrals are infinite

Both estimation methods under consideration involve the quantity $\int f_\theta^{1+\gamma}\,dz$, cf. the criterion functions (2.3) and (2.9). For some parametric models this and related integrals are infinite for a region of parameter values θ, depending on the value of γ. This may in particular happen when the density $f_\theta(x)$ is unbounded in x, as for certain parameter combinations in the gamma, beta and Weibull families, for example.

To illustrate the point, consider the simple family $f_\theta(x) = \theta x^{\theta-1}$ on the unit interval, where θ is positive and unknown. Here $\int f_\theta^{1+\gamma}\,dz$ is finite only when $\theta > \gamma/(1+\gamma)$. When $\theta > 1$ there are no problems, and both methods are consistent and behave according to results discussed above. For smaller θ, the methods work only when $\gamma < \theta/(1-\theta)$, and will err otherwise.

Calculations for this example illustrate that both methods may lose rather a lot in efficiency to the maximum likelihood method, and that the BHHJ method loses significantly less than the Windham one, when $\theta < 1$, unless γ is quite small compared to $\theta/(1-\theta)$. When the model holds for $\theta = 0.25$, for example, then the two methods have a chance of behaving reasonably only for γ values below 1/3, and the efficiencies are already as low as 48.6 and 46.1 percent, respectively, when $\gamma = 0.10$, and as unacceptably low as 6.7 and 5.3 percent, respectively, when $\gamma = 0.15$.

3·4. Examples

The first example is the Newcomb light speed data, analysed by Brown and Hwang (1993) and BHHJ. The normal density $\mathrm{N}(\mu, \sigma^2)$ is fitted to the data. Table 2 gives the estimates $\widehat{\mu}_\alpha, \widehat{\sigma}_\alpha$ and $\widehat{\mu}^{(0)}_\beta, \widehat{\sigma}^{(0)}_\beta$ for $\gamma = \alpha = \beta = 0, 0.02, 0.05, 0.1, 0.25, 0.5, 1$. As one can see from the table the methods appear virtually identical for small γ. For larger γ, between say 0.5 and 1, there is some discrepancy, but typically one would not want to use this large a γ as the efficiency loss becomes serious. The most important aspect here is that the estimate of σ quickly goes down in size as $\gamma \geq 0.05$.

*** Table 2 about here ***

The second example is also considered in BHHJ and is an analysis of data on fruit flies presented by Simpson (1987). Male flies are exposed to different doses of a chemical and are then mated with unexposed females. 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. In Simpson (1987) and BHHJ, Poisson models with mean λ were fitted to the data.

Table 3 gives the results of fitting the Poisson (λ) model to these data, both with and without the outlier. Again for $\gamma < 0.5$ there is virtually no difference between $\widehat{\lambda}_\alpha$ and $\widehat{\lambda}^{(0)}_\beta$. The most important aspect of both methods is that for tuning parameter as small as 0.02, the estimates are drastically shifted away from the maximum likelihood result, and actually making them quickly similar to results obtained when the outlier is removed. This reflects the methods' effective robustness towards outliers.

We might add that our treatment of this example is primarily meant to illustrate the use of our estimation methods; in the particular situation at hand the single prolific male fly might well be of real significance for some aspects of the analysis, and the simple Poisson model might be too naive.

*** Table 3 about here ***

The third example is based on an analysis of telephone line fault data presented in Welch (1987), also analysed by Simpson (1989). The data represent the difference of inverse fault rates between the test and the control in 14 matched pairs. The observations are –988, –135, –78, 3, 59, 83, 93, 110, 189, 197, 204, 229, 269, 310, in values test minus control, multiplied by 10^5. Here we carry out a parametric test under the $\mathrm{N}(\mu,\sigma^2)$ model of the hypothesis $\mu=0$ versus $\mu>0$, where σ is unspecified. We perform the analogue of a one-sided Wald test, comparing

$$W_\gamma = n^{1/2}\widehat\mu_\gamma\Big/\Big(1+\frac{\gamma^2}{1+2\gamma}\Big)^{3/4}\widehat\sigma_\gamma$$

to the $\mathrm{N}(0,1)$ distribution. The underlying fact used here is that $n^{1/2}(\widehat\mu-\mu)$ has a limiting $\mathrm{N}(0,\tau^2)$ distribution, where $\tau^2=\{1+\gamma^2/(1+2\gamma)\}^{3/2}\sigma^2$, for both estimation methods; see the BHHJ paper.

Table 4 presents the parameter estimates, Wald statistics and p-values using both the type 0 and type 1 estimators for several values of γ. Again, for small values of γ, the two sets of estimates as well as p-values are very similar. The most important aspect of the example is to illustrate that the ordinary maximum likelihood-based method can be 'fooled' by outliers making the σ estimates too big; with the effective robust methods of this paper the message comes through that there is a significant difference between the test and control objects.

*** Table 4 about here ***

In these examples we used a fixed point algorithm for the two-parameter problems, and a bisection method for the one-parameter problem.

4. Robust regression

A parametric estimation method is much more valuable if it can be used not only in settings with independent and identically distributed data but also in general regression models. How the type 1 methodology can be extended to regression contexts was briefly indicated in §3.5 of the BHHJ paper. The following brief arguments show how the type 0 estimation method can also be extended to regression cases.

Consider a situation with a model $f_\theta(y\,|\,x)$ for some true density $g(y\,|\,x)$. Look at the x-conditional distance

$$\begin{aligned} D_\beta[g(\cdot\,|\,x),f_\theta(\cdot\,|\,x)] \;=\; &\frac{1}{\beta}\Big[\Big(\int g(y\,|\,x)^{1+\beta}\,\mathrm{d}y\Big)^{1/(1+\beta)} \\ &-\Big(\int g(y\,|\,x)f_\theta(y\,|\,x)^\beta\,\mathrm{d}y\Big)\Big/\Big(\int f_\theta(y\,|\,x)^{1+\beta}\,\mathrm{d}y\Big)^{\beta/(1+\beta)}\Big], \end{aligned} \tag{4.1}$$

defined here as a close relative to the Windham-related distance function (2.11), and consider making an estimation method for θ that aims at minimising

$$D_\beta[g(\cdot\,|\,\cdot),f_\theta(\cdot\,|\,\cdot)] = \int D_\beta[g(\cdot\,|\,x),f_\theta(\cdot\,|\,x)]\,R(\mathrm{d}x), \tag{4.2}$$

where R is the distribution of covariates. Using the form (2.11) itself leads if followed by the arguments below to a less fortunate problem form; the point of the transformation (4.1) is to arrive at a criterion function of the type $\mathrm{E}_{X,Y} H_\theta(X,Y)$, for a suitable function H, where the expectation is with respect to the simultaneous distribution of (x_i, Y_i) pairs. And minimising (4.2) means maximising $\mathrm{E}_{X,Y} f_\theta(Y \mid X)^\beta / v_\theta(X)^{\beta/(1+\beta)}$ where $v_\theta(x) = \int f_\theta(y \mid x)^{1+\beta}\,\mathrm{d}y$. This leads to the following proposal: let $\widehat\theta = \widehat\theta^{(0)}_\beta$ maximise

$$n^{-1}\sum_{i=1}^n f_\theta(Y_i \mid x_i)^\beta / v_\theta(x_i)^{\beta/(1+\beta)} = n^{-1}\sum_{i=1}^n \exp\{\beta \log f_\theta(Y_i \mid x_i) - (\beta/(1+\beta)) \log v_\theta(x_i)\}.$$

This succeeds in the sense of giving an estimator that is consistent for the θ_β minimising the overall d_β distance above. One might also derive a $n^{1/2}(\widehat\theta - \theta_\beta)$ limit result, and so on. Under model conditions, θ_β is the value of the true parameter.

This procedure amounts to a robust generalisation of the maximum likelihood method (which emerges when $\beta \to 0$), and has very little efficiency loss under model conditions if $\beta \le 0.10$, say. It can easily be used to robustify linear, Poisson and gamma regression, and so on. In the absence of covariates, the method reduces to that of Windham.

One may also develop robust model choice criteria in the style of Akaike's information criterion (or use cross validation). These methods can help making a flexible robust estimation and inference package for practical statistics in a wide range of regression contexts.

5. Discussion

This paper has shown how the estimator of Windham (1995) can be recast as a minimum divergence estimator closely related to the estimator introduced by Basu, Harris, Hjort and Jones (1998). It is shown that both divergences are members of a larger family of divergences, and furthermore that these two appear to be the most interesting ones from a statistical viewpoint. Both minimum divergence estimators include maximum likelihood as a special limiting case and both are examples of M-estimators. We also stress that the methods work in any sample space, for example when the data are vectors. In particular, minimising the appropriate versions of (2.3) and (2.9) will provide robust estimation of mean vector and variance matrix for the multinormal model.

We have no universal way of selecting the tuning parameters α for type 1 or β for type 0 estimators. These fine-tune the underlying discrepancy measures and act to balance loss of efficiency under model conditions (compared to maximum likelihood methods) versus increased robustness. See the parallel discussion of this point in BHHJ.

Taken as a whole the examples and calculations suggest there is little difference between the BHHJ and Windham methods in practical circumstances, at least for the simple models we have investigated. For a given $\gamma = \alpha = \beta$ the type 1 estimator is slightly more efficient than the type 0 estimator. Implementation of the two procedures is quite similar. The situation with regards to robustness is a little more mixed. The type 0 estimator tended to slightly outperform the type 1 estimator in terms of mean squared error in many cases. The breakdown in the $\mathrm{N}(\mu, \sigma^2)$ problem is zero for the type 0 estimator, which is a disadvantage, but one that may not be too worrisome; Maguluri and Singh (1997) give other examples of robust estimators with zero breakdown. The type 0 method of Windham may also have occasional problems with very small observations for some life-time distributions, as indicated in §3.3, problems apparently not encountered for the BHHJ method.

Both estimation methods dealt with here have extensions to general regression models, and may be supplemented further with robust model choice criteria. All in all they make up valuable versatile robust and nearly efficient alternatives to traditional statistical inference in all parametric models.

Appendix

Some rather technical matters have been placed here, in order not to interrupt the natural flow of discussion in the main part of our paper.

A1. The two-parameter divergence family

We prove here that (2.14) defines a divergency (being nonnegative for all densities f and g with equality only when they are equal). The starting point is that of inequalities

$$\frac{1}{1+\gamma}x^{-\gamma} + \frac{\gamma}{1+\gamma}x \geq 1 \qquad \text{for all } \gamma > 0,\ x > 0, \tag{A.1}$$

and

$$\int f^\gamma g \leq \Big(\int f^{1+\gamma}\Big)^{\gamma/(1+\gamma)} \Big(\int g^{1+\gamma}\Big)^{1/(1+\gamma)} \qquad \text{for all } \gamma > 0. \tag{A.2}$$

The first of these is proved using basic calculus, while the second follows from the Hölder inequality. Now raise both sides of (A.2) to the power of $1+\gamma$, and rearrange a little, to get

$$\left(\frac{\int f^\gamma g}{\int f^{1+\gamma}}\right)^\gamma \leq \frac{\int g^{1+\gamma}}{\int f^\gamma g}.$$

This implies

$$\frac{1}{1+\gamma}\left(\frac{\int g^{1+\gamma}}{\int f^\gamma g}\right)^\phi + \frac{\gamma}{1+\gamma}\left(\frac{\int f^{1+\gamma}}{\int f^\gamma g}\right)^\phi \geq \frac{1}{1+\gamma}\left(\frac{\int f^\gamma g}{\int f^{1+\gamma}}\right)^{\gamma\phi} + \frac{\gamma}{1+\gamma}\left(\frac{\int f^{1+\gamma}}{\int f^\gamma g}\right)^\phi$$

for each positive ϕ. Applying (A.1) shows that the right hand side is at least 1, which upon a little further transport of symbols gives

$$\Big(\int g^{1+\gamma}\Big)^\phi + \gamma\Big(\int f^{1+\gamma}\Big)^\phi \geq (1+\gamma)\Big(\int f^\gamma g\Big)^\phi.$$

This leads to the required conclusion.

A2. Large-sample behaviour of type 0 estimators

The following provides a direct derivation of the limiting distribution result for type 0 estimators, rather than deducing it as a consequence of more general results for M-estimation. The advantage is simplicity and the possibility of checking regularity conditions for each step in the argument, for the parametric model at hand. Also, similar arguments become necessary for handling estimators of the type considered in the remark ending Section 2, for values of ϕ outside 0 and 1.

The type 0 estimator solves

$$\begin{aligned} U_n(\theta) &= \frac{n^{-1}\sum_{i=1}^n f_\theta(X_i)^\beta u_\theta(X_i)}{n^{-1}\sum_{i=1}^n f_\theta(X_i)^\beta} - \frac{\int f_\theta^{1+\beta}u_\theta}{\int f_\theta^{1+\beta}} \\ &= \frac{A_n(\theta)}{B_n(\theta)} - \frac{M_{0,1+\beta}}{L_{0,1+\beta}} = 0. \end{aligned}$$

The least false parameter is also the solution to $u(\theta)=0$, where u is the limit in probability of U_n; that is,

$$\xi_\beta = \frac{M_{1,\beta}}{L_{1,\beta}} = \frac{M_{0,1+\beta}}{L_{0,1+\beta}}.$$

To prove the limit theorem, we use

$$\begin{aligned} n^{1/2}(\widehat\theta_\beta^{(0)}-\theta) &\doteq \{-U_n'(\theta)\}^{-1}\,n^{1/2}U_n(\theta) \\ &\doteq \{-U_n'(\theta)\}^{-1}\frac{1}{L_{1,\beta}}\Big[n^{1/2}\{A_n(\theta)-M_{1,\beta}\} - \frac{M_{1,\beta}}{L_{1,\beta}}n^{1/2}\{B_n(\theta)-L_{1,\beta}\}\Big] \\ &\to_d (J_\beta^{(0)})^{-1}\mathrm{N}(0,K_\beta^{(0)}) = \mathrm{N}(0,(J_\beta^{(0)})^{-1}K_\beta^{(0)}(J_\beta^{(0)})^{-1}). \end{aligned}$$

Here $U_n'(\theta)$ is the $p\times p$ matrix derivative of the p-vector $U_n(\theta)$, and the notation $C_n \doteq D_n$ is used to indicate that their difference tends to zero in probability.

Take $K_\beta^{(0)}$ first. We have

$$\begin{aligned} K_\beta^{(0)} &= \frac{1}{L_{1,\beta}^2}\mathrm{Var}\Big[f_\theta(Y)^\beta u_\theta(Y) - M_{1,\beta} - \frac{M_{1,\beta}}{L_{1,\beta}}\{f_\theta(Y)^\beta - L_{1,\beta}\}\Big] \\ &= \frac{1}{L_{1,\beta}^2}\mathrm{Var}\{f_\theta(Y)^\beta u_\theta(Y) - \xi_\beta f_\theta(Y)^\beta\} \\ &= \frac{1}{L_{1,\beta}^2}\int g f_\theta^{2\beta}(u_\theta-\xi_\beta)(u_\theta-\xi_\beta)^{\mathrm t}\,\mathrm{d}y. \end{aligned}$$

This can also be expressed as $(1/L_{1,\beta}^2)(N_{1,2\beta} - \xi_\beta M_{1,2\beta}^{\mathrm t} - M_{1,2\beta}\xi_\beta^{\mathrm t} + \xi_\beta\xi_\beta^{\mathrm t}L_{1,2\beta})$. Under model conditions,

$$K_\beta^{(0)} = \frac{1}{L_{0,1+\beta}^2}\int f_\theta^{1+2\beta}(u_\theta-\xi_\beta)(u_\theta-\xi_\beta)^{\mathrm t}\,\mathrm{d}y$$

$$= \frac{1}{L_{0,1+\beta}^2}(N_{0,1+2\beta} - \xi_\beta M_{0,1+2\beta}^{\rm t} - M_{0,1+2\beta}\xi_\beta^{\rm t} + \xi_\beta\xi_\beta^{\rm t} L_{0,1+2\beta}).$$

Next let us work with $J_\beta^{(0)}$, the limit in probability of $-U_n'(\theta)$. One finds $A_n'(\theta) \to_p \beta N_{1,\beta} - \int g f_\theta^\beta i_\theta \,{\rm d}y$ and $B_n'(\theta) \to_p \beta M_{1,\beta}$. Some further manipulations lead to

$$\begin{aligned} J_\beta^{(0)} &= \frac{N_{0,1+\beta}}{L_{0,1+\beta}} - \frac{M_{0,1+\beta}}{L_{0,1+\beta}}\Big(\frac{M_{0,1+\beta}}{L_{0,1+\beta}}\Big)^{\rm t} + \beta\Big(\frac{N_{0,1+\beta}}{L_{0,1+\beta}} - \frac{N_{1,\beta}}{L_{1,\beta}}\Big) \\ &\qquad + (1/L_{1,\beta})\int g f_\theta^\beta i_\theta \,{\rm d}y - (1/L_{0,1+\beta})\int f_\theta^{1+\beta} i_\theta \,{\rm d}y. \end{aligned}$$

Under model conditions this reduces to $J_\beta^{(0)} = N_{0,1+\beta}/L_{0,1+\beta} - \xi_\beta\xi_\beta^{\rm t}$.

A3. Variance formulae for the geometric family

Study $f_\theta(x) = (1-\theta)\theta^x$ for $x = 1, 2, \ldots$, which has score function $u_\theta(x) = (x-1)/\theta - 1/(1-\theta)$. From $G_0(a) = \sum_{x=1}^\infty a^x = 1/(1-a)$ it is elementary to deduce

$$\begin{aligned} G_1(a) &= \sum_{x=1}^\infty x a^x = a/(1-a)^2, \\ G_2(a) &= \sum_{x=1}^\infty x^2 a^x = (a+a^2)/(1-a)^3. \end{aligned}$$

With definitions as per equation (3.1), one finds

$$\begin{aligned} A_\gamma &= \theta^{1+\gamma} G_0((1-\theta)^{1+\gamma}), \\ B_\gamma &= \theta^{1+\gamma}\Big\{\frac{1}{\theta}G_0((1-\theta)^{1+\gamma}) - \frac{1}{1-\theta}G_1((1-\theta)^{1+\gamma})\Big\}, \\ C_\gamma &= \theta^{1+\gamma}\Big\{\frac{1}{\theta^2}G_0((1-\theta)^{1+\gamma}) - \frac{2}{\theta(1-\theta)}G_1((1-\theta)^{1+\gamma}) + \frac{1}{(1-\theta)^2}G_2((1-\theta)^{1+\gamma})\Big\}. \end{aligned}$$

These are then sufficient to lead to explicit and easily programmable formulae for limiting variances for the two estimation methods, as per equations following (3.1).

Acknowledgements

The authors would like to thank Michele Basseville for references and Brent Burch and Roy St Laurent for helpful conversations.

References

Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

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

BURBEA, J. & RAO, C. R. (1982). Entropy differential metric and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis* **12**, 575–596.

CSISZAR, I. (1991). Why least-squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics* **19**, 2032–2066.

DE ANGELIS, D. & YOUNG, G. A. (1992). Smoothing the bootstrap. *International Statistical Review* **60**, 45–56.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986). *Robust Statistics: the Approach Based on Influence Functions.* New York: Wiley.

HAVRDA, M. E. & CHARVAT, F. (1967). Quantification method of classification processes: concept of structural alpha-entropy. *Kybernetica* **3**, 30–35.

JONES, L. K. & BYRNE, C. L. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Transactions on Information Theory* **36**, 23–30.

JONES, L. K. & TRUTZER, V. (1989). Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method. *Inverse Problems* **5**, 749–766.

MAGALURI, G. & SINGH, K. (1997). On the fundamentals of data robustness. In *Handbook of Statistics*, **15**, *Robust Inference*, eds. G. S. Maddala and C. R. Rao, pp. 537–550. Amsterdam: North-Holland.

SILVERMAN, B. W. & YOUNG, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74**, 469–479.

SIMPSON, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association* **82**, 802–807.

SIMPSON, D. G. (1989). Hellinger deviance tests: efficiency, breakdown points, and examples. *Journal of the American Statistical Association* **84**, 107–113.

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

WINDHAM, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society, Series B* **57**, 599–609.

Table 1: Asymptotic relative efficiencies of the minimum divergence estimators, given as ratios of limiting variances, in percent, for various values of $\alpha = \beta$.

Model	Estimator	0.00	0.02	0.05	0.10	0.25	0.50	1.00
Normal μ	both	100	99.9	99.7	98.8	94.1	83.8	65.0
Normal σ	type 1	100	99.9	99.3	97.6	88.8	73.1	54.1
	type 0	100	99.9	99.3	97.5	88.5	70.6	43.3
Exponential	type 1	100	99.8	99.1	96.8	85.8	68.4	50.9
	type 0	100	99.8	99.1	96.7	85.1	63.2	33.8
Poisson ($\lambda = 3$)	type 1	100	99.9	99.7	98.8	94.4	85.0	67.9
	type 0	100	99.9	99.7	98.8	94.2	84.2	65.3
Poisson ($\lambda = 10$)	type 1	100	99.9	99.7	98.8	94.1	84.0	65.6
	type 0	100	99.9	99.7	98.8	94.0	83.8	64.9
Geometric ($\theta = 0.1$)	type 1	100	99.8	99.1	96.8	85.9	68.4	51.1
	type 0	100	99.8	99.1	96.7	85.1	63.3	33.9
Geometric ($\theta = 0.9$)	type 1	100	99.9	99.4	98.0	92.0	84.1	82.2
	type 0	100	99.9	99.4	98.0	91.7	81.7	71.3

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

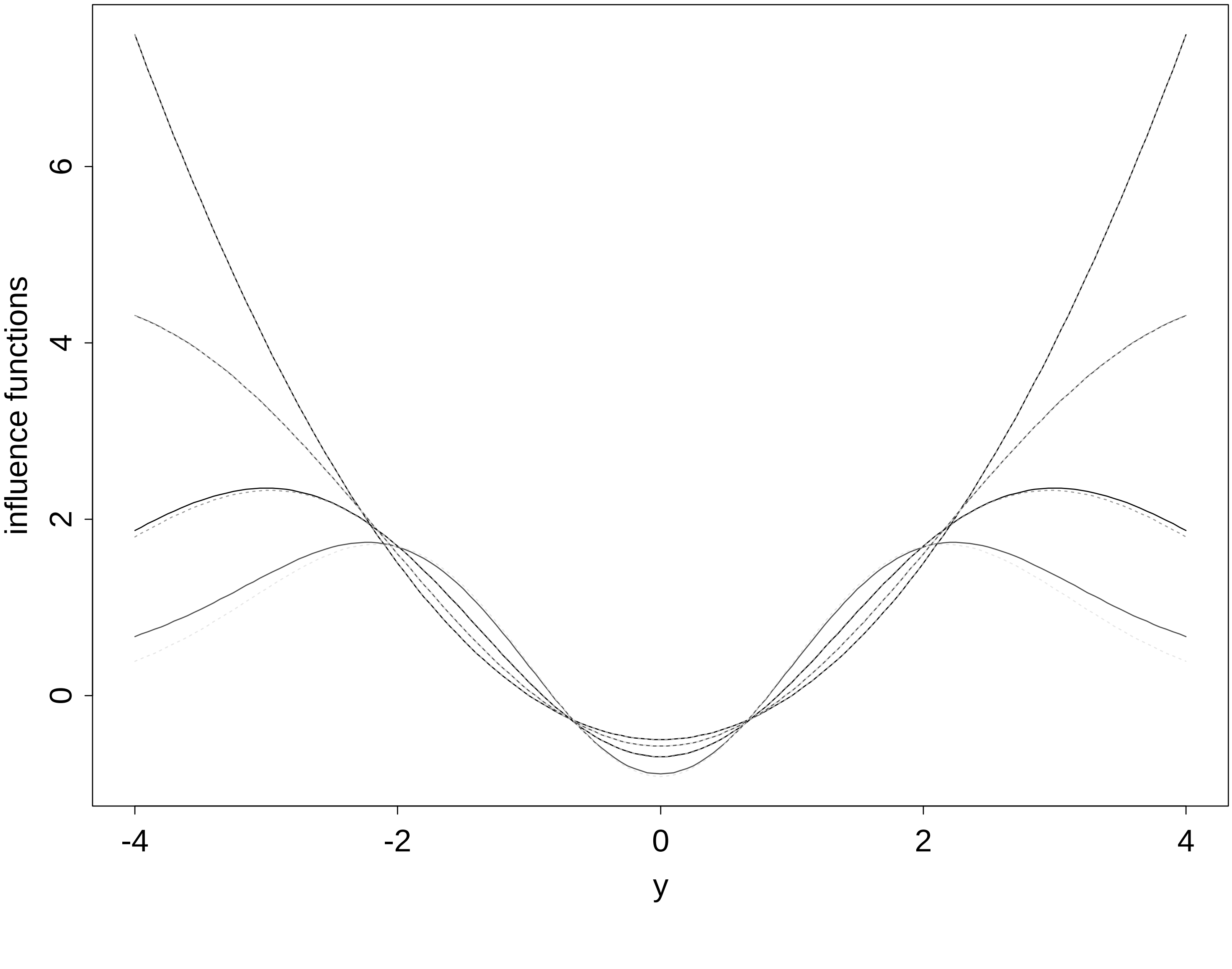
$\gamma = \alpha = \beta$	0.00	0.02	0.05	0.10	0.25	0.50	1.00
$\widehat{\mu}_\alpha$	26.21	26.74	27.44	27.60	27.64	27.52	27.29
$\widehat{\mu}_\beta^{(0)}$	26.21	26.74	27.44	27.60	27.64	27.51	27.25
$\widehat{\sigma}_\alpha$	10.66	8.92	5.99	5.39	5.04	4.90	4.67
$\widehat{\sigma}_\beta^{(0)}$	10.66	8.92	5.99	5.38	5.01	4.82	4.34

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

$\gamma = \alpha = \beta$	0.00	0.001	0.01	0.02	0.05	0.10	0.25	0.50	1.00
$\widehat{\lambda}_\alpha$ (all data)	3.059	2.506	0.447	0.394	0.393	0.392	0.386	0.375	0.365
$\widehat{\lambda}_\beta^{(0)}$ (all data)	3.059	2.506	0.447	0.394	0.392	0.390	0.381	0.362	0.330
$\widehat{\lambda}_\alpha$ (outlier deleted)	0.394	0.394	0.394	0.393	0.392	0.390	0.382	0.366	0.350
$\widehat{\lambda}_\beta^{(0)}$ (outlier deleted)	0.394	0.394	0.394	0.393	0.392	0.390	0.381	0.362	0.330

Table 4: Test statistics and p-values for the Wald type test for the telephone fault data.

Estimator used	$\gamma = \alpha = \beta$	0.01	0.10	0.25	0.50	1.00
Type 1	$\widehat{\mu}$	42.8	96.0	124.7	131.1	142.2
	$\widehat{\sigma}$	305.6	209.2	133.4	136.9	139.5
	W statistic	0.52	1.71	3.39	3.28	3.07
	p-value (normal)$\times 10^5$	29998	4405	35	52	105
Type 0	$\widehat{\mu}^{(0)}$	42.8	96.4	124.9	131.7	144.3
	$\widehat{\sigma}^{(0)}$	305.6	207.8	132.0	133.8	132.4
	W statistic	0.52	1.73	3.43	3.37	3.29
	p-value (normal)$\times 10^5$	29998	4222	30	38	51


influence functions
0
2
4
6
-4
-2
0
2
4
y