*Biometrika* (2012), **99**, 1, *pp.* 238–244
doi: 10.1093/biomet/asr061
Advance Access publication 1 December 2011


# On robust estimation via pseudo-additive information

By DAVIDE FERRARI

*University of Modena and Reggio Emilia, via Berengario 51, 41100 Modena, Italy*
davide.ferrari@unimore.it

and DAVIDE LA VECCHIA

*Universita' della Svizzera italiana, via G. Buffi 13, 6904 Lugano, Switzerland*
davide.la.vecchia@usi.ch

Summary

We consider a robust parameter estimator minimizing an empirical approximation to the $q$-entropy and show its relationship to minimization of power divergences through a simple parameter transformation. The estimator balances robustness and efficiency through a tuning constant $q$ and avoids kernel density smoothing. We derive an upper bound to the estimator mean squared error under a contaminated reference model and use it as a min-max criterion for selecting $q$.

*Some key words*: Change-of-variance; Minimum divergence estimation; Power divergence; $q$-entropy; Robustness.

## 1. Introduction

Let $\mathscr{F}_\Theta = \{F_t, t \in \Theta \subseteq R^p\}$ $(p \geqslant 1)$ be a family of parametric distributions with densities $f_t$ and let $\mathscr{G}$ be the class of all distributions $G$ having a density $g$ with respect to Lebesgue measure. Assume $f_t$ and $g$ have common support $\mathcal{X} \subseteq R^k$ $(k \geqslant 1)$. Our suggestion is to replace the usual loglikelihood function for a set of independent observations $X_1, \ldots, X_n$ from $G$ by the surrogate likelihood function

$$S_r(t) = \sum_{i=1}^{n} f_t(X_i)^r, \tag{1}$$

where $0 < r < 1$ is a tuning constant. We consider the surrogate estimator $\hat{\theta}_n^*$ obtained by maximizing (1). We shall show that as $r$ tends to 0 the usual maximum likelihood estimator is recovered and as $r$ increases the estimator becomes more robust. An important example is the multivariate normal distribution $N_p(\mu, \Sigma)$. However, there are two problems for this and similar examples. First, the global maximum of (1) is singular, regardless of the data, occurring for $|\Sigma| = 0$. To fix this problem, we look for a local maximum of (1). The population version of (1) is often better-behaved, with a unique stationary point, which is also the global optimum, lying in the interior of the parameter set. In such cases, the local maximum of (1) converges to this population value as $n \to \infty$. Second, even if $G$ lies in $\mathscr{F}_\Theta$, the estimator $\hat{\theta}_n^*$ will not be consistent. This problem can be fixed, assuming a closure condition on $\mathscr{F}_\Theta$, by introducing a calibration function to rescale the parameter estimate. Given a density $g$ and an index $0 < \alpha < \infty$, the power transformation is defined by

$$g^{(\alpha)}(x) = g^\alpha(x) \left\{ \int g^\alpha(x) dx \right\}^{-1}, \tag{2}$$

provided the integral in the denominator converges. We assume that $\mathscr{F}_\Theta$ is closed under (2), for all $0 < \alpha < 1$, and define the continuous function $\tau_\alpha \colon \Theta \mapsto \Theta$ satisfying $f_{\tau_\alpha(t)}(x) = f_t^{(\alpha)}(x)$, for all $x \in \mathcal{X}$. The calibration function $\tau_\alpha$ maps a parameter on the power-transformed scale to a parameter on the original

scale and ensures that the procedure is Fisher consistent; see § 2. The final estimator is computed as $\hat{\theta}_n = \tau_{1-r}(\hat{\theta}_n^*)$, where $\hat{\theta}_n^*$ is the maximizer of (1). For common families of distributions, $\tau_\alpha(t)$ has a closed form.

For later use it is more convenient to parameterize the surrogate likelihood in terms of $q = 1 - r$ rather than $r$, so that $0 < q < 1$. The population version of (1) is proportional to a generalized information measure, sometimes referred to as $q$-entropy (Tsallis, 1988), and defined by

$$H_q(f_t||g) = -\int L_q\{f_t(x)\}g(x)dx,$$

where

$$L_q(u) = \begin{cases} (u^{1-q} - 1)/(1-q) & (q \neq 1), \\ \log(u) & (q = 1). \end{cases}$$

Minimizing the sample version $\ell_q(t) = -\sum_{i=1}^n L_q\{f_t(X_i)\}$, for a given $0 < q < 1$ is the same as maximizing (1), since $S_r(t)$ is linearly related to $\ell_q(t)$, for $r = 1 - q$. As $q \to 1$, we recover the usual negative loglikelihood $\ell_1(t)$, which can be viewed as the empirical counterpart of the additive Shannon entropy (Akaike, 1973). Minimizing $\ell_q(t)$ is equivalent to solving the estimating equations

$$\sum_{i=1}^n u_q(X_i, t) = \sum_{i=1}^n u(X_i, t) f_t(X_i)^{1-q} = 0, \tag{3}$$

where $u(x, t) = \nabla_t \log\{f_t(x)\}$ is the score function. In equation (3), the score function receives weights depending on the model itself and $q$. Choices of $q$ smaller than 1 define a robust M-estimation procedure. If $q$ is near 0, the procedure gains robustness because observations that are inconsistent with the target model receive small weights.

## 2. Power divergences, q-entropies and Fisher consistency

We consider the family of power divergences of $f_t$ with respect to $g$, defined by

$$D_q(f_t||g) = -\frac{1}{q}\int L_q\left\{\frac{f_t(x)}{g(x)}\right\} g(x)dx. \tag{4}$$

Notable divergences are special cases of (4) such as the Kullback–Leibler divergence, obtained when $q \to 1$, and the Hellinger distance, $q = 1/2$. In our setting, if $q \to 1$, $L_q(\cdot) \to L_1(\cdot) = \log(\cdot)$. The additivity of $\log(\cdot)$ implies $D_1(f_t||g) = H_1(f_t||g) - H_1(g||g)$ and minimization over $\Theta$ depends only on the term $H_1(f_t||g) = -E_G[\log\{f_t(X)\}]$. Hence, given independent observations with common distribution $G$, the expectation in $H_1(f_t||g)$ is approximated by $-n^{-1}\sum_{i=1}^n \log\{f_t(X_i)\}$ and the minimizer of such a quantity is the maximum likelihood estimator. For $q \neq 1$, one cannot proceed as for $q = 1$, because $L_q(\cdot)$ is not additive. Therefore, traditional estimators based on (4) minimize $D_q(f_t||\hat{g}_h)$, where $\hat{g}_h$ is a nonparametric estimate of the true density $g$ (Beran, 1977; Lindsay, 1994). This, however, leads to complications in multivariate problems: choosing the bandwidth can be difficult, and the accuracy of such parameter estimators rests on the convergence rate of the density smoother, which suffers from the curse of dimensionality. We now identify a strategy alternative to minimization of (4) by kernel smoothing.

Lemma 1. *Assume that* $0 < \int g(x)^{1/q}dx < \infty$. *Then* $D_q(f_t||g^{(1/q)}) = \xi^{-1}\{H_q(f_t||g) - H_q(g^{(1/q)}||g)\}$, *where* $\xi = qE_G\{g(X)^{1/q-1}\}$.

Lemma 1 shows that, up to a constant not depending on $f_t$, a power divergence can be split into the difference between the $q$-entropy for $f_t$ and that for $g^{(1/q)}$. In the proof of Lemma 1, Jensen's inequality and the convexity of $L_q(\cdot)$ imply that $D_q(f_t||g) \geqslant 0$, where $D_q(f_t||g) = 0$ if and only if $f_t = g$ almost everywhere. This property and Lemma 1 lead to Fisher consistency of the minimizer of $H_q(f_t||g)$, when the parameter is properly rescaled.

Let $\theta^* = T_q^*(G)$ be the minimizer of $H_q(f_t||g)$ and define $T_q(G) = \tau_q(\theta^*)$. Throughout the paper, we assume the existence of an open set $\Theta^* \subset \Theta$, such that $\theta^*$ is an interior point of $\Theta^*$. We also assume that $E_G\{\sup_{t\in\Theta^*} f_t^{1-q}(X)\} < \infty$, which ensures the existence of $\theta^*$. Usually, $\theta^*$ is unique, and we will assume this to be the case. Finally, we assume that $\tau_q(t)$ is defined for all $t \in \Theta^*$, which is equivalent to requiring that the power transformation $f_t^{(q)}$ is a nondegenerate density in the original family $\mathscr{F}_\Theta$, for all $t \in \Theta^*$.

PROPOSITION 1 (Fisher consistency). *Let $\tau_q: \Theta \mapsto \Theta$ be the transformation defined following (2). Then, $\theta = \arg\min_{t\in\Theta} D_q(f_t||f_\theta) = \tau_q\{T_q^*(F_\theta)\}$.*

The closure of $\mathscr{F}_\Theta$ under (2) holds automatically for canonical exponential families with density $f_t(x) = \exp\{\eta(t)^\top a(x) - b(t) + c(x)\}$, when $c(x)$ is identically 0, and implies a closed form for $\tau_q(\cdot)$. For the exponential density $f_\lambda(x) = \lambda\exp(-\lambda x)(x > 0, \lambda > 0)$ we have $\tau_q(\lambda) = q\lambda$. For the multivariate normal distribution with mean $\mu$ and covariance $\Sigma$, we obtain $\tau_q(\mu^\top, \mathrm{vech}^\top\Sigma) = (\mu^\top, q^{-1}\mathrm{vech}^\top\Sigma)$.

To clarify the role played by $\tau_q(\cdot)$, consider differentiating $H_q(f_t||g)$ under $F_\theta$. If the order of differentiation and integration can be exchanged, $\nabla_t H_q(f_t||f_\theta) = -\int \nabla_t f_t(x) f_\theta(x) f_t(x)^{-q} dx$. If $t$ is such that $f_t(x) = f_\theta(x)^{1/q}$, or equivalently $t = \tau_q^{-1}(\theta) = \tau_{1/q}(\theta)$, then

$$\nabla_t H_q(f_t||f_\theta) = -c_q(\theta)\int \nabla_t f_t(x)dx = -c_q(\theta)\nabla_t \int f_t(x)dx = 0,$$

where $c_q(\theta) = \{\int f_\theta^{1/q}(x)dx\}^q$, i.e., $\tau_q^{-1}(\theta)$ is the root of $\nabla_t H_q(f_t||f_\theta) = 0$. Lemma 1 points out the role played by the power transformation $g^{(1/q)}$, which is the target density when minimizing $H_q(f_t||g)$. For $0 < q < 1$, $g^{(1/q)}$ enhances parts of $g$ with higher density values. Thus, the relevance of the majority of the data is increased and the importance of the tails, which are usually the most affected by the presence of contamination, is reduced.

## 3. ESTIMATION AND INFINITESIMAL ROBUSTNESS

### 3·1. *Computational aspects*

To compute the estimates, standard optimization methods can be considered, and the form of equation (3) suggests exploring iterative reweighting strategies. Regardless of the computational approach, some care is needed because the objective function $\ell_q(t)$ could have its global minimum on the boundary of $\Theta$, besides having at least one local minimum in the interior of $\Theta$. For example, if $f_\sigma(x)$ is the density of a normal distribution $N_1(0, \sigma)$, then $\ell_q(\sigma) \propto \sum_{i=1}^n f_\sigma(X_i)^{1-q}$, which diverges to $\infty$ when $X_i = 0$ for some $i$ and $\sigma \to 0$. To avoid singular solutions, one can build on and improve a preliminary robust estimator, say $\tilde\theta_n$. For example, the one-step estimation method, takes a preliminary root-$n$ consistent M-estimator of $\theta = \tau_q(\theta^*)$ and computes a solution to (3) as $\hat\theta_{q,n}^* = t - \{\sum_{i=1}^n \nabla_t u_q(X_i, t)\}^{-1}\sum_{i=1}^n u_q(X_i, t)$, evaluated at $t = \tau_q^{-1}(\tilde\theta_n)$, e.g., see van der Vaart (1998, Theorem 5.48). Valid choices for the preliminary estimator $\tilde\theta_n$ are the M-estimators proposed by Kent & Tyler (1996) or other estimators such as those discussed in § 4.

### 3·2. *Asymptotics, influence and change-of-variance functions*

Equation (3) defines an M-estimator, so the asymptotics of $T_q^*(G_n)$ and $T_q(G_n)$ can be treated using existing theory. Define $p \times p$ matrices $K_q(t, G) = E_G\{u_q(x, t)u_q(x, t)^\top\}$, $J_q(t, G) = E_G\{\nabla_t u_q(x, t)\}$ and write $K_q(t) = K_q(t, F_t)$, $J_q(t) = J_q(t, F_t)$, if $G = F_t$. One can show that $n^{1/2}\hat\theta_{q,n}$ converges to a multivariate normal variate with mean $\theta$ and variance matrix

$$V_q(\theta, G) = \bar J_q(\theta, G)^{-1}\bar K_q(\theta, G)\bar J_q(\theta, G)^{-\top}, \tag{5}$$

where $\bar J_q(t, G) = J_q\{\tau^{-1}(t), G\}\{\nabla_t\tau_q(t)\}^{-1}$ and $\bar K_q(t, G) = K_q\{\tau^{-1}(t), G\}$. In the rest of the paper, we use the notation $V_q(t) = V_q(t, F_t)$, $\bar J_q(t) = \bar J_q(t, F_t)$ and $\bar K_q(t) = \bar K_q(t, F_t)$. For $F_\theta \in \mathscr{F}_\Theta$, we consider deviations $F_\epsilon = (1-\epsilon)F_\theta + \epsilon W$, $W \in \mathscr{G}$ $(0 \leqslant \epsilon \leqslant 1/2)$. We denote the estimating functional and the asymptotic

variance under the misspecified model by $T_q(\epsilon) = T_q(F_\epsilon)$ and $V_q(\epsilon) = V_q\{T_q(\epsilon), F_\epsilon\}$, respectively. As it is customary in the literature of M-estimation, the calculations for the next results are carried out using the worst-case contamination $W = \delta_x$ $(x \in \mathcal{X})$ where $\delta_x$ is Dirac's delta.

A standard calculation shows that the influence function for $T_q(\cdot)$ is $\mathrm{IF}_q(x, \theta) = \nabla_\theta \tau(\theta) \mathrm{IF}_q^*(x, \theta)$, where $\mathrm{IF}_q^*(x, \theta) = -J_q^{-1}(\theta^*, F_\theta) u_q(x, \theta^*)$ is the influence function for $T_q^*(\cdot)$. If $0 < q < 1$, then $\mathrm{IF}_q(x, \theta)$ is proportional to $f_{\theta^*}^{1-q}(x) u(x, \theta^*)$, where the term $f_{\theta^*}^{1-q}(x)$ usually corrects for the unboundedness of the score function and implies a redescending estimator. The influence function provides a first order approximation to bias for the M-functional $T_q(\cdot)$, since $T_q(\epsilon) - \theta \approx \epsilon \partial T_q(\epsilon)/\partial\epsilon|_{\epsilon=0} = \epsilon \mathrm{IF}_q(x, \theta)$. The gross-error sensitivity is defined by $\gamma_q(\theta) = \sup_x \|\mathrm{IF}_q(x; \theta)\|$; if $\gamma_q(\theta) < \infty$, we say that $T_q(\cdot)$ is B-robust.

The influence function alone does not provide direct information on the stability of the asymptotic variance of $T_q(\cdot)$. For this reason, we use the change-of-variance function for $T_q(\cdot)$ defined by the mapping $\mathrm{CVF}_q \colon \mathcal{X} \times \Theta \mapsto R^{p \times p}$ such that $\partial\{V_q(\epsilon)\}_{ij}/\partial\epsilon|_{\epsilon=0} = \{\mathrm{CVF}_q(x, \theta)\}_{ij}$ $(i, j = 1, \ldots, p)$. The change-of-variance sensitivity is defined by $\kappa_q(\theta) = \sup_x \mathrm{tr}\{\mathrm{CVF}_q(x, \theta)\}/\mathrm{tr}\{V_q(\theta)\}$. The change-of-variance function measures the influence of a small amount of contamination on the asymptotic variance and $\kappa_q(\theta)$ represents the worst variability change under infinitesimal contamination. If $\kappa_q(\theta) < \infty$, the estimator is said to be variance robust, or V-robust.

PROPOSITION 2 (Change-of-variance function). *Define $u_q$ as in (3) and let $V_q(\theta)$ be as in (5). Assume that $E_{F_\theta}\{(u_q)_i\} < \infty$, $E_{F_\theta}\{(\nabla_\theta u_q)_{ij}\} < \infty$, and $E_{F_\theta}\{\partial(\nabla_\theta u_q)_{ij}/\partial\theta_k\} < \infty$ $(i, j, k = 1, \ldots, p)$. Then $CVF_q(x, \theta)$ is equal to*

$$\begin{aligned} &\bar{J}_q(\theta)^{-1} \tilde{K}_q(x, \theta) \bar{J}_q(\theta)^{-\mathrm{T}} + \bar{J}_q(\theta)^{-1} \tilde{K}_q(x, \theta)^{\mathrm{T}} \bar{J}_q(\theta)^{-\mathrm{T}} + \tilde{J}_q(x, \theta)^{-1} \bar{K}_q(\theta) \bar{J}_q(\theta)^{-\mathrm{T}} \\ &\quad + \bar{J}_q(\theta)^{-1} \bar{K}_q(\theta) \tilde{J}_q(x, \theta)^{-\mathrm{T}} + IF_q(x, \theta) IF_q(x, \theta)^{\mathrm{T}} - V_q(\theta), \end{aligned} \tag{6}$$

*where*

$$\begin{aligned} \tilde{K}_q(x, \theta) &= E_{F_\theta}\{u_q(X, \theta) IF_q^{\mathrm{T}}(x, \theta) \nabla_\theta \tau_q^{-\mathrm{T}} \nabla_\theta u_q(X, \theta)^{\mathrm{T}}\}, \\ \tilde{J}_q(x, \theta)^{-1} &= Q_q(x, \theta) J_q - \bar{J}_q^{-1}(\theta)\{D_q(x, \theta) + \nabla_\theta u_q(x) - J_q(\theta)\} J_q^{-1}(\theta) \end{aligned}$$

*and $Q_q(x, \theta)$, $D_q(x, \theta)$ have elements $\{Q_q(x, \theta)\}_{i,j} = \sum_{k=1}^p \partial(\nabla_\theta \tau_q)_{i,j}/\partial\theta_k \{IF_q(x, \theta)\}_k$, $\{D_q(x, \theta)\}_{i,j} = E_{F_\theta}[\sum_{k=1}^p \partial\{\nabla_\theta u_q(X)\}_{i,j}/\partial\theta_k \{IF_q^*(x, \theta)\}_k]$ $(i, j = 1, \ldots, p)$.*

Now, $E_{F_\theta}\{\mathrm{IF}_q(X, \theta)\} = 0$ and, if all expectations in (6) are well-defined, we also have $E_{F_\theta}\{\mathrm{CVF}_q(X, \theta)\} = 0$, since $E_{F_\theta}\{\mathrm{IF}_q(X, \theta)\mathrm{IF}_q(X, \theta)^{\mathrm{T}}\} = V_q(\theta)$ and for the mixed terms $E_{F_\theta}\{\tilde{K}_q(X, \theta)\} = 0$ and $E_{F_\theta}\{\tilde{J}_q(X, \theta)^{-1}\} = 0$. If $u_q(x, t)$ is replaced by a generic M-functional, Proposition 2 generalizes known results derived for the one-parameter case, see Hampel et al. (1986, Ch. 2.5) and Genton & Rousseeuw (1995) for scale and location. Finally, from an inspection of the expressions of the influence and change-of-variance functions, sufficient conditions for B- and V-robustness are boundedness of $u_q$ and its first derivative. These are satisfied for common distributions such as those in the exponential family.

### 3·3. *Worst-case mean squared error and min-max selection of q*

In this section, we study the mean squared error of $\hat{\theta}_n$, under $\epsilon$-contamination. The approximate worst-case bias is $\epsilon\gamma_q(\theta)$. An extrapolation of the asymptotic variance is

$$\mathrm{tr}\{\mathrm{CVF}_q(x, \theta)\}/\mathrm{tr}\{V_q(\theta)\} = \partial \log[\mathrm{tr}\{V_q(\epsilon)\}]/\partial\epsilon|_{\epsilon=0} \approx \epsilon^{-1}(\log[\mathrm{tr}\{V_q(\epsilon)\}] - \log[\mathrm{tr}\{V_q(\theta)\}]).$$

From the above expression, we obtain $\mathrm{tr}\{V_q(\theta)\}\exp\{\epsilon\kappa_q(\theta)\}$ as an approximation of the worst-case variance. By combining the information about the worst-case bias with that on the worst-case variance,

we obtain an approximate upper bound for the mean squared error:

$$\text{MMSE}(q, \theta; n, \epsilon) = \epsilon^2 \gamma_q(\theta)^2 + n^{-1} \text{tr}\{V_q(\theta)\} \exp\{\epsilon \kappa_q(\theta)\}. \tag{7}$$

This can be used as a criterion for choosing $q$. For given $n$ and $0 \leqslant \epsilon \leqslant 1/2$, we set a grid of tuning parameters and compute the corresponding estimates. Then, we choose the value of $q$ minimizing the maximal mean squared error. The selected value of $q$ will automatically take care of the interplay between bias and variance, as a function of the $\epsilon$-contamination and $n$.

## 4. LINK WITH OTHER PROCEDURES

### 4·1. *Related M-estimators*

The strategy of setting weights proportional to the assumed model has appeared from different motivations in various contexts. In a setting different from the current paper, Ferrari & Yang (2010) consider (3) for estimation of the tail probability under the correct model when the sample size is small. Basu et al. (1998) propose the minimum power density divergence estimator, which shares some appealing features with the procedure described here. Both approaches are fully parametric, as they do not require kernel smoothing and are applicable to a wide range of models. When $\mathscr{F}_\Theta$ is a location family, the estimation equations (2.4) in Basu et al. (1998) are basically the same as equation (3). In general, however, the two methods rely on two different families of divergences, which overlap only for the special case of the Kullback–Leibler divergence for pure location models. Basu et al. (1998) consider a Bregman divergence which generalizes the integrated square error; instead, our information theory approach leads to a generalization of the Hellinger distance. Consequently, outside the location family, the trade-off between robustness and efficiency is not necessarily the same for the two estimators and depends both on the form of $F_\theta$ and on the degree of contamination. This is illustrated in § 4·2. The approach of Basu et al. (1998) preserves the Fisher consistency using the typical recentring of the estimating function, by computing $\int f_t(x)^c u(x, t) dx$, $c > 0$. The computation of this quantity can be cumbersome, especially for multivariate models with many parameters. Instead, typically the rescaling transformation $\tau_q$ has closed form and is easy to compute.

In some specific instances, the approach presented here coincides with known redescending M-estimators of location and scatter, say $\mu$ and $\Sigma$, of an elliptic density $\phi(s)$, where $s = (x - \mu)^\text{T} \Sigma^{-1} (x - \mu)$. For the Gaussian density, the solutions of the weighted likelihood (3) satisfy the equations $\hat{\mu} = E_{G_n}\{w(S)X\}$ and $\hat{\Sigma} = E_{G_n}\{v(S)(X - \mu)(X - \mu)^\text{T}\}$, where $E_{G_n}$ stands for the arithmetic mean over $i = 1, \ldots, n$ and $w(s) = \phi(s)^{1-q} / E_{G_n}\{\phi(S)^{1-q}\}$, $v(s) = 2q^{-1} \exp\{-(1 - q)s\} / E_{G_n}[\exp\{-(1 - q)S\}]$. Kent & Tyler (1996) proposed constrained M-estimates by minimizing $E_G\{\rho(S)\} + \log\{\det(\Sigma)\}/2$, subject to $E_G\{\rho(S)\} \leqslant \epsilon \rho(\infty)$. When $\rho$ is the exponentially weighted function $\rho(s) = 1 - \exp\{-(1 - q)s\}$, minimizing (3) is equivalent to their approach for estimating $\mu$. Their scatter matrix estimate, however, differs from ours as their weights in $\hat{\Sigma}$ take the form $v(s) = 2(1 - q) \exp\{-(1 - q)s\}$. For the normal distribution, this different weight specification affects the trade-off between robustness and efficiency: while our redescending estimator can be more efficient, Kent and Tyler's estimator implies higher breakdown.

### 4·2. *Trade-off between robustness and efficiency for particular families*

Consider estimating the parameter $\lambda$ of an exponential distribution with density $\lambda \exp(-\lambda x)$ $(x > 0, \lambda > 0)$. The asymptotic variance and the gross-error sensitivity are

$$V_q(\lambda) = \lambda^2 \frac{(2q - 2 - q^2)}{(q - 2)^3 q^3}, \quad \gamma_q(\lambda) = \lambda q \left\{ 1 + \frac{1}{(1 - q)^{1/2}} \right\}.$$

For small values of $q \to 0$, $\gamma_q$ is small and the estimator is expected to be remarkably robust. This advantage, however, comes with large efficiency losses compared with maximum likelihood estimation. Conversely, if $q \to 1$, then we obtain the maximum likelihood variance $V_q \to V_1 = \lambda^2$, but $\gamma_q \to \infty$. Intermediate choices of $q$ balance those two limit scenarios. Figure 1 shows the maximal mean squared error

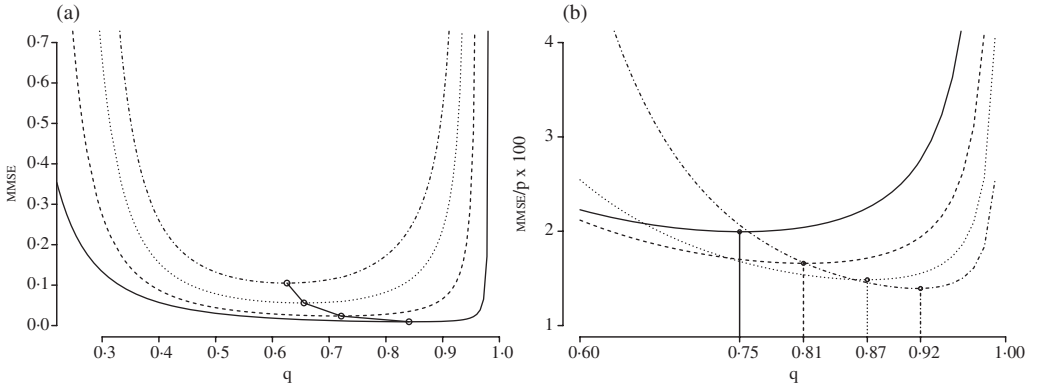


Fig. 1. Maximal mean squared error. (a) Exponential distribution with rate $\lambda = 1$, $n = 150$ and $\epsilon = 1\%$ (solid), $\epsilon = 5\%$ (dashed), $\epsilon = 10\%$ (dotted) and $\epsilon = 15\%$ (dot-dashed). (b) Elementwise maximal mean squared error for the mean of $N_p(0, I)$, $\epsilon = 0{\cdot}05$, $n = 100$, for $p = 1$ (solid), $p = 2$ (dashed), $p = 4$ (dotted) and $p = 8$ (dot-dashed). The circles show optimal values of $q$.

Table 1. *Percent asymptotic relative efficiency of our estimator with respect to maximum likelihood for $N_p(\mu, I)$. The optimal values of $q$ obtained via maximal mean squared error minimization are in brackets*

| | | $\epsilon = 0{\cdot}05$ | | | | $\epsilon = 0{\cdot}15$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $p$ | 1 | 5 | 15 | 30 | 1 | 5 | 15 | 30 |
| | | 90·8 | 95·8 | 97·9 | 98·6 | 86·0 | 93·3 | 96·9 | 98·6 |
| 100 | | [0·75] | [0·89] | [0·95] | [0·97] | [0·69] | [0·86] | [0·94] | [0·97] |
| | | 85·9 | 93·3 | 96·9 | 98·6 | 84·1 | 93·3 | 96·9 | 98·6 |
| 1000 | | [0·69] | [0·86] | [0·94] | [0·97] | [0·67] | [0·86] | [0·94] | [0·97] |

against $q$, for $\lambda = 1$ and $n = 150$. For a small $\epsilon$-contamination, a wide interval for $q$, from about 0·30 to 0·95, ensures small errors. Choices of $q$ close to 1 in that range are preferred, since they provide high efficiency. When $\epsilon$ increases, the interval of safe choices for $q$ narrows and moves away from 1.

For estimating the mean $\mu$ of a multivariate normal $N_p(\mu, \Sigma)$ with known $\Sigma$, the asymptotic variance is $V_q = \{q(2-q)\}^{-p/2-1}\Sigma$, when no contamination occurs. In the presence of contamination, both influence and change-of-variance functions are bounded for $0 < q < 1$. The former exhibits the typical shape of a redescending estimator. Figure 1 shows the maximal mean squared error for estimating a mean component of $N_p(0, I)$ for $p = 1, 2, 4$ and 8, when $\epsilon = 0{\cdot}05$ and $n = 100$. We also report the corresponding optimal values of $q$. When $p$ increases, two simultaneous effects occur: the optimal value of the tuning constant $q$ gets closer to 1 and the global maxima of both influence and change-of-variance functions decrease. Interestingly, the values of $q$ minimizing the worst-case error also correspond to the highest efficiency. In Table 1, we report the asymptotic relative efficiencies, at the model, with respect to the maximum likelihood estimator corresponding to optimal values of $q$ selected by (7), for $\epsilon = 0{\cdot}05, 0{\cdot}15$, and $n = 100, 1000$. For estimating $\mu$, the procedure is the same as that of Basu et al. (1998) when their tuning parameter is $\alpha = 1 - q$. Thus, the trade-off between robustness and efficiency illustrated above holds equally for both estimators. However, outside the location family this is not generally the case and the two estimators yield different efficiencies for a given robustness level. In Table 2, we compare the asymptotic relative efficiencies of the two estimators, for different choices of $\alpha$ and $q$ and for estimating the scale $\sigma$ of $N_1(0, \sigma^2)$. For any $\alpha$, we compute the gross-error sensitivity of Basu's estimator, $\gamma_\alpha$, and set $q$ such that $\gamma_q = \gamma_\alpha$. When $q$ approaches 1, the two estimators behave similarly: both are almost fully efficient. For $q$ larger than 0·85, the estimator of Basu et al. (1998) is slightly better, while other choices imply a better trade-off between robustness and efficiency of our estimator. Finally, we considered the estimator of Kent & Tyler (1996), with weights as discussed in § 4. In general, Kent and Tyler's estimator offers greater robustness

Table 2. *Asymptotic relative efficiency in percent for $N_1(0, \sigma^2)$ of our estimator, $ARE_q$, and the Basu et al. (1998) estimator, $ARE_\alpha$, computed for different tuning constants $q$ and $\alpha$ yielding the same gross-error sensitivity $\gamma_\alpha = \gamma_q$*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma_\alpha = \gamma_q$ | $\infty$ | 8·12 | 4·47 | 2·35 | 1·74 | 1·58 | 1·55 | 1·51 | 1·46 | 1·44 |
| $\alpha$ | 0·00 | 0·05 | 0·10 | 0·25 | 0·50 | 0·75 | 0·85 | 1·00 | 1·25 | 1·35 |
| $\text{ARE}_\alpha$ | 100 | 99·3 | 97·5 | 88·8 | 73·0 | 61·5 | 58·1 | 54·1 | 49·4 | 48·0 |
| $q$ | 1·00 | 0·98 | 0·95 | 0·90 | 0·85 | 0·82 | 0·81 | 0·80 | 0·79 | 0·78 |
| $\text{ARE}_q$ | 100 | 96·1 | 92·2 | 81·1 | 69·3 | 63·3 | 61·7 | 59·8 | 56·9 | 55·8 |

with gross-error sensitivity values as low as 0·77. However, the minimal efficiency loss compared with maximum likelihood is about 35%.

## ACKNOWLEDGEMENT

The authors thank the editor, an associate editor, and a referee for their insightful comments and suggestions that significantly improved the paper. The work is supported by the Swiss National Science Foundation Pro*Doc Program and NCCR FinRisk.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of Lemma 1 and Propositions 1 and 2.

## REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory*, Eds. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

BASU, A., HARRIS, I. R., HJORT, N. L. & JONES, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–59.

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445–63.

FERRARI, D. & YANG, Y. (2010). Maximum $L_q$-likelihood estimation. *Ann. Statist.* **38**, 753–83.

GENTON, M. G. & ROUSSEEUW, P. J. (1995). The change-of-variance function of $M$-estimators of scale under general contaminations. *J. Comp. Appl. Math.* **64**, 69–80.

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

KENT, J. T. & TYLER, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Ann. Statist.* **24**, 1346–70.

LINDSAY, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–114.

TSALLIS, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *J. Statist. Phys.* **52**, 479–87.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

[*Received June* 2009. *Revised August* 2011]