RECent
Center for Economic Research

WORKING PAPER SERIES

Parametric density estimation by minimizing nonextensive entropy

Davide Ferrari

Working paper 16

May 2008

www.recent.unimore.it

RECent: c/o Dipartimento di Economia Politica, Viale Berengario 51, I-41100 Modena, ITALY
Phone +39 059 2056856, Fax +39 059 2056947, E-mail: RECent@unimore.it

Parametric density estimation by minimizing nonextensive entropy

Davide Ferrari

May 26, 2008

Abstract

In this paper, we consider parametric density estimation based on minimizing an empirical version of the Havrda-Charvát-Tsallis ([15], [25]) nonextensive entropy. The resulting estimator, called the Maximum Lq-Likelihood estimator (MLqE), is indexed by a single distortion parameter q, which controls the trade-off between bias and variance. The method has two notable special cases. If q tends to 1, the MLqE is the Maximum Likelihood Estimator (MLE). When $q = 1/2$, the MLqE is a minimum Hellinger distance type of estimator with the perk of avoiding nonparametric techniques and the difficulties of bandwith selection. The MLqE is studied using asymptotic analysis, simulations and real-world data, showing that it conciliates two apparently contrasting needs: efficiency and robustness, conditional to a proper choice of q. When the sample size is small or moderate, the MLqE trades bias for variance, resulting in a reduced mean squared error compared to the MLE. At the same time, the MLqE exhibits strong robustness at expense of a slightly reduced efficiency in presence of observations discordant with the assumed model. To compute the MLq estimates, a fast and easy-to-implement algorithm based on a reweighting strategy is also described.

1 Introduction

In parametric estimation, one approach is to compute the parameters of interest by minimizing some appropriate data-based divergence between an assumed model and the true model density underlying the data. The successfull tradition in this area dates back to Rao [23] and Kullback [19]. Undoubtely, the most popular representative of these methods is the Maximum Likelihood estimator (MLE), whose relationship with the Kullback-Leibler (KL) divergence has been pointed out by Akaike [2]. Despite the MLE is asymptotically efficient, in practice the large sample property is satisfied when two important conditions hold: (i) the assumed model mimics well the empirical distribution of the data and (ii) the sample size is sufficiently large.

In the last few decades, a large body of literature aimed to produce estimators that are not unduly affected by small departures from such requirements. Beran [4] first introduced a density minimum divergence estimator based on Hellinger distance. Similar research lines were embraced by Lindsay [20], Park et al.[22] and Bhandari et al. [5]. Basu et al. [3] considered the use of minimum density power divergences, a class of divergences indexed by a parameter that controls for the trade-off between robustness and efficiency. Their approach avoids kernel smoothing and exhibits robustness properties similar to those of L_2-norm based estimators. In continuous models, these techniques require some degree of nonparametric analysis, with all the complications related to the bandwith choice, which can be hard to handle in high-dimensional problems.

In a different direction, Hu and Zidek [16] proposed the weighted likelihood estimator, derived via minimization of the KL divergence subject to data-dependent constraints. The weighted likelihood approach extends the local likelihood method of Tibshirani and Hastie [24] and it shares its underlying purpose with other methods such as weighted least squares and kernel smoothers which can reduce an estimator's variance while increasing its bias to reduce mean-squared error. In practice, however, the advantages of

weighted likelihood methods rely heavily on a proper selection of the weights, which in many problems can be only performed using "ex-post" data-driven procedures such as cross-validation [27].

In this paper, we consider a family of quasi-logarithmic density divergences. The task of minimizing the proposed family has an information-theoretical flavor, since it amounts to minimization of Tsallis-Havrda-Charvát entropy, sometimes called nonextensive or q-entropy ([15],[25]). Tsallis and collegues have successfully employed such measures in the context of statistical mechanics (e.g., see [25]). More recently, applications have appeared in finance, social, biomedical and environmental sciences (e.g., see Gell-Mann [12]). The underlying goal of our work is to address the statistical usage of the q-entropy for density estimation and explore the properties of the new estimator.

The q-entropy, is indexed by a single parameter of distortion q, which controls the trade-off between asymptotic bias and variance of the parameter estimators which are the minimizers of such a family. The resulting estimator is called Maximum Lq-Likelihood estimator (MLqE) and has been studied by Ferrari and Yang [11] in the context of small tail inference. When q is fixed, the MLqE belongs to the class of M-estimators, but yet representing a novel case motivated by the need of improving upon the efficiency of MLE when the sample size is small or moderate. When q is judiciously chosen and the sample size is moderate or small, the MLqE trades successfully bias for variance, reducing the mean squared error, sometimes dramatically compared to the classical MLE. This phenomenon is confirmed by the asymptotic analysis, computer simulations and real-world examples.

Besides, our approach appears to conciliate both efficiency and robustness aspects, which usually involve distinct techniques: efficiency is prioritized when the model is thought to appropriately describe the data at hand and robustness is stressed when it is not. In our view, these objectives are intertwined as the degree to which an observation is treated as "outlying" de-

pends not only on the probability of its occurrence under the assumed model, but also on the sample size. In presence of outliers or perturbations from the assumed model, the same methodology can be exploited to the purpose of classical robust estimation. The MLqE generalizes other familiar minimum divergence robust estimators depending on the value of the distortion parameter q. For example, when $q = 1/2$ the MLqE can be regarded as equivalent to minimization of Hellinger distance. However, contrarily to other existing methods such as Beran's minimum Hellinger distance estimator (MHDE), our apporach has the perk of not involving any nonparametric analysis, yet mantaining reasonable performances.

In addition to the appealing properties, our methodology answers three important needs of the practicioner: easy implementability, interpretability and computational efficiency. The estimating equations are simply obtained by replacing the logarithm of log-likelihood function in the usual maximum likelihood procedure by the distorted logarithm $L_q(u) = (u^{1-q} - 1)/(1 - q)$. The resulting optimization task can be formulated in terms of a weighted version of the familiar score function, where the weights are proportional to the $(1-q)$th power of the assumed density. Consequently, a simple and fast algorithm is automatically available for coumputing MLq estimates and in many cases the steps of the algorithm reduce to a simple variable transformation.

The rest of the paper is organized as follows. In section 2, we introduce a class of quasi-logarithmic divergences and point out their connection with nonextensive entropies. In section 3, we introduce the MLqE for parametric families. In section 4, convergence in probability and asymptotic normality results of MLqE are provided in light of exisiting M-estimation theory. In addition, we discuss the trade-off between bias and variance for particular families of distributions. In section 5, we present an easy-to-implement procedure for computing the MLq estimates and account for possible strategies for the choice of the distortion parameter q. In section 6, we apply the method to real-world examples and assess the finite-sample performance of

the MLqE via Monte Carlo simulations. In section 7, final remarks are given.

2 Power divergence and nonextensive entropy measure

Consider a family of models $\mathcal{F}$ having densities (or probability mass functions) $\{g\}$ with respect the measure μ on the support space Ω. Denote the true density by f, which does not have to necessarily belong to $\mathcal{F}$. Given a convex function $\varphi : \mathbf{R} \to \mathbf{R}$, the large class of Csiszár divergences [10] is given by $E_f \varphi \{f(x)/g(x)\}$. Commonly, optimization is performed with respect to g and different choices of φ lead to different divergence measures. Pheraps, the most common choice is $\varphi(\cdot) = \log(\cdot)$, which yields the popular KL divergence, or *relative entropy* [19]. Consider instead the following family of divergences.

Definition 2.1. *Define the power divergence between the distributions $g(x)$ and $f(x)$ as*

$$\mathcal{D}_q(g||f) = -\frac{1}{q}\int_\Omega f(x) L_q\left\{\frac{g(x)}{f(x)}\right\} d\mu(x), \tag{2.1}$$

where $L_q(u) = (u^{1-q} - 1)/(1-q)$, and $q \in (-\infty, \infty) \setminus \{1\}$.

When $q = 1$, the integrand is undefined and we set $\log(\cdot) = \lim_{q\to 1} L_q(\cdot)$, recovering the KL divergence. Interestingly, some algebra shows that many common divergences can be recovered as special cases of the power divergence. Namely, one can obtain Neyman's Chi-squared divergence

$$NCS(g||f) = \int_\Omega \frac{[f(x) - g(x)]^2}{2f(x)} d\mu(x); \tag{2.2}$$

Hellinger distance:

$$HD(g||f) = \int_{\Omega} \left[\sqrt{g(x)} - \sqrt{f(x)} \right]^2 d\mu(x); \tag{2.3}$$

KL divergence:

$$KL(g||f) = \int_{\Omega} f(x) \log \left\{ \frac{f(x)}{g(x)} \right\} d\mu(x); \tag{2.4}$$

Pearson's Chi-squared:

$$PCS(g||f) = \int_{\Omega} \frac{[f(x) - g(x)]^2}{2g(x)} d\mu(x); \tag{2.5}$$

by letting $q = -1, 1/2, 1, 2$, respectively. The same type of divergence has been considered by Cressie and Read [9] in relation to goodness-of-fit tests and by Lindsay [20] in the context of robust estimation. Although in general the power divergence is not a distance as it lacks of symmetry, it enjoys the following important discrimination property.

Theorem 2.1. *Let $g(x)$ and $f(x)$ be two density functions on Ω. Then, $\mathcal{D}_q(g||f) \geq 0$ and the equality is attained if and only if $g = f$ almost everywhere.*

Proof. Note that $\frac{1}{q}\partial^2 L_q(u)/\partial u^2 < 0$ for all $-\infty < q < \infty$. Thus, by Jensen's inequality we have

$$-\frac{1}{q} \int_{\Omega} f(x) L_q \left\{ \frac{g(x)}{f(x)} \right\} d\mu(x) \geq -\frac{1}{q} L_q \int_{\Omega} f(x) \frac{g(x)}{f(x)} d\mu(x) = 0. \tag{2.6}$$

□

When $q = 1$, $L_q(\cdot)$ is the usual logarithm and the task of minimizing $\mathcal{D}_1(g||f)$ can be equivalently restated in terms of minimization of Shannon's entropy. In particular, $\mathcal{D}_1(g||f) = \mathcal{H}_1(f||g) - \mathcal{H}_1(f||f)$, where $\mathcal{H}_1$ represents

Shannon's information measure, defined as

$$\mathcal{H}_1(g||f) = -\int_\Omega f(x)\log\{g(x)\}\,d\mu(x). \tag{2.7}$$

The quantity $-\log g(x)$ is interpreted as the information content of the outcome x evaluated at the candidate density $g(\cdot)$ and $\mathcal{H}_1(g||f)$ is the average uncertainty removed after the actual outcome of the random variable X is revealed. When $q \neq 1$, the q-logarithm obeys the following *pseudo-additivity* property:

$$L_q(u_1u_2) = L_q(u_1) + L_q(u_2) + (1-q)L_q(u_1)L_q(u_2),\ u_1, u_2 > 0,\ q > 0 \tag{2.8}$$

and full additivity is recovered as $q \to 1$. Thus, we can write

$$-q\mathcal{D}_q(g||f) = \int L_q(g) + L_q(f^{-1}) + (1-q)L_q(g)L_q(f^{-1})f\,d\mu \tag{2.9}$$

$$= \int \frac{f^{q-1}-1}{1-q} + \frac{g^{1-q}f^{q-1} - f^{q-1}}{1-q} f\,d\mu \tag{2.10}$$

$$= \int L_q(g)f^q d\mu - \int L_q(f)f^q d\mu. \tag{2.11}$$

Since in (2.11) integration is taken with respect a power transformation of the true density, minimization of $\mathcal{D}_q(g||f)$ based on an empirical version of f might be cumbersome. Instead, consider the tranformation $f^{(\alpha)}(x) := f(x)^\alpha / \int f(x)^\alpha d\mu(x)$, $\alpha > 0$. Replacing the true density f in Eq.(2.11) with the transformation $f^{(1/q)}$ gives

$$\mathcal{D}_q(g||f^{(1/q)}) = Z^{-1}(q)\left(\int L_q(f^{(1/q)})f\,d\mu - \int L_q(g)f\,d\mu\right), \tag{2.12}$$

where $Z(q) = q\int f^{1/q}d\mu$. Therefore, minimizing (2.12) is equivalent to min-

imize the *q-entropy*, or nonextensive entropy, defined as

$$\mathcal{H}_q(g||f) = -\int_\Omega f(x) L_q\left\{g(x)\right\} d\mu(x). \tag{2.13}$$

Given observations $X_1, \ldots, X_n$ from f, our program is to minimize $\mathcal{H}_q(g||f)$. Since f is unkown, we replace the above expectation with one taken with respect the empirical distribution of the data F_n and find the minimizer of $-\sum_i L_q\left\{g(X_i)\right\}$, say $\widehat{f}_*$. Of course, $\widehat{f}_*$ is a biased estimate of the target f and one can promptly remedy to this by considering $\widehat{f} = \widehat{f}_*^{(q)}$ instead. However, this does not have to be necessarily the case and later we shall see that retaining the bias can reduce sensibly the variance of the estimates, resulting in an overall gain in terms of mean squared error when the sample size is small.

The transformed density $f^{(1/q)}$ is sometimes referred to as *zooming* or *escort* distribution ([21],[1]) and q provides a tool for accentuating different regions of the untransformed true density f. In particular, note that when $q > 1$, regions with density values close to zero are accentuated, while for $q < 1$ regions with density values further from zero are emphasized.

3 The Maximum Lq-Likelihood method

In the rest of the paper, we consider the parametric familiy $\mathcal{F}(\Theta) = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbf{R}^p\}$. The true parameter vector is denoted by θ_0. In the parametric case, the q-entropy is

$$\mathcal{H}_q(\theta||\theta_0) = -\int_\Omega f(x;\theta_0) L_q\left\{f(x;\theta)\right\} d\mu(x). \tag{3.1}$$

Let $\theta^* = \arg\min_{\theta\in\Theta} \mathcal{H}_q(\theta||\theta_0)$. and note that θ^* depends upon both θ_0 and the distortion parameter q. For q fixed, we make the fundamental assumption that there exists a unique target parameter θ^*. The Maximum Lq-Estimator

of θ_0 is the point that minimizes the q-entropy relative to the probability mass function $F_n(x)$ associated with the empirical distribution of the sample and $f(x;\theta)$.

Definition 3.1. *Let $X_1, ..., X_n$ be a random sample from $f(x;\theta_0)$, $\theta_0 \in \Theta$. The Maximum L_q-Likelihood Estimator (MLqE) of θ_0 is defined as*

$$\widehat{\theta}_{q,n} = \arg\max_{\theta\in\Theta} \ell_q(\theta) := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} L_q\left[f(X_i;\theta)\right], \quad q > 0, \tag{3.2}$$

where L_q is the q-logarithmic function defined in (2.1).

When $q \to 1$, if the estimator $\widehat{\theta}_{n,1}$ exists, it is the maximum likelihood estimator of the parameters, which maximizes $\int \log\{f(x;\theta)\}\, dF_n(x)$. In general, the estimating equations have the form

$$\Psi_n(\theta) := n^{-1} \sum_{i=1}^{n} U(X_i;\theta) f(X_i;\theta)^{1-q} = 0, \tag{3.3}$$

where $U(x;\theta) = \nabla_\theta \log\{f(x;\theta)\}$ is the maximum likelihood score function. When $q \neq 1$, eq.(3.3) provides a relative-to-the-model downweighting. Observations that disagree sensibly with the model receive low weight. In the case $q = 1$, all the observations receive the same weight. The idea of setting weights that are proportional to the family from which the model is to be chosen is not new in literature. Windham [28] and Choi et al. [8] propose similar strategies to robustify esimators. For location models, the MLqE is the same as minimum density power divergence of Basu et al. and the robustified estimator of Windham: in such case (3.3) equals to equation (2.4) in Basu et al. [3], when $q = 1 - \alpha$. Our perspective seeks a contact with these approaches and ultimately highlights the role played by nonextensive entropy measures in downweighting with respect the model rather than the data.

4 Properties and standard errors

When q is fixed, the MLqE is an M-estimator and properties can be derived by applying existing theory (see Huber [17] and Hampel et al. [13]). M-estimators are zeros of equations of the form $\sum_i \psi(X_i;\theta) = 0$; in our case the criterion function ψ is $\psi(x;\theta) = \nabla_\theta L_q\{f(x;\theta)\}$. Let $U(x;\theta)$ and $I(x;\theta)$ denote the score function and the information matrix of $f(x;\theta)$, respectively. Let $J_q(\theta)$ and $K_q(\theta)$ be the following $p \times p$ matrices:

$$K_q(\theta) := \int_\Omega f(x;\theta)^{2(1-q)} U(x;\theta) U(x;\theta)^{\mathtt{T}} f(x;\theta_0) d\mu(x) \tag{4.1}$$

and

$$J_q(\theta) := \int_\Omega f(x;\theta)^{(1-q)} [(1-q) U(x;\theta) U(x;\theta)^{\mathtt{T}} - I(x;\theta)] f(x;\theta_0) d\mu(x) \tag{4.2}$$

In the next section, we shall see that under some conditions: (i) there exists a sequence of MLqE points $\widehat{\theta}_{q,n}$ that is consistent for θ^* and (ii) the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_{q,n} - \theta^*)$ is asymptotically normal with mean 0 and variance $J_q(\theta^*)^{-1} K_q(\theta^*) J_q(\theta^*)^{-1}$.

4.1 Convergence results

The criterion function is $\psi(x;\theta) = f(x;\theta)^{1-q} U(x;\theta)$. Let $\Theta^* \subseteq \Theta$ be the set of points such that $\int |\psi(x,\theta)| f(x;\theta_0) dx < \infty$ and assume that Θ is compact. For $\theta \in \Theta^*$, consider

$$\Psi(\theta) := \int_\Omega L_q\{f(x;\theta)\} f(x;\theta_0) d\mu(x) < \infty \tag{4.3}$$

and set $\Psi(\theta) = -\infty$ if θ is not in Θ^*. Consequently, θ^* is such that $\sup_{\theta\in\Theta} \Psi(\theta)$ is finite. The next theorem establishes consistency of the MLqE for estimating θ^*.

Theorem 4.1. *Assume the following conditions:*
(C.1) For $\theta \in \Theta$, $\psi(x, \theta)$ is continuous almost everywhere;
(C.2) For all suffciently small balls B, $\sup_{\theta \in B} \{\psi(x, \theta)\}$ is measurable and

$$E_{\theta_0} \sup_{\theta \in B} \{\psi(x, \theta)\} < \infty. \tag{4.4}$$

Then, any sequence $\widehat{\theta}_{q,n}$ of MLqE satisfying $\psi_n(\widehat{\theta}_{q,n}) \geq \psi_n(\theta_q^*) - o_p(1)$, is such that for any $\varepsilon > 0$ and every compact set $K \subset \Theta$,

$$P\left(||\widehat{\theta}_{q,n} - \theta^*|| > \varepsilon \wedge \widehat{\theta}_{q,n} \in K\right) \to 0. \tag{4.5}$$

Proof. The proof is given in van der Vaart [26], Theorem 5.14. □

Next, we introduce some additional smoothness conditions needed to obtain asymptotic normatlity of MLqE.

Lemma 4.2. *Suppose that $\psi(x; \theta)$ is differentiable in a neighborhood of θ^* almost everywhere. Assume that there exists an open ball $B \in \Theta$ and a constant c such that $||\nabla_\theta \psi(x; \theta)|| \leq c$ for θ in B. Then, for every $\theta_1, \theta_2 \in B$ almost everywhere, there exists a constant $\gamma(x)$, such that*

$$|\psi(x; \theta_1) - \psi(x; \theta_2)| \leq \gamma(x)||\theta_1 - \theta_2||, \quad \textit{and} \quad E||\gamma(x)||^2 < \infty. \tag{4.6}$$

The lemma is a well-known property of differentiable mappings and states that if $\psi(x; \theta)$ is differentiable mapping, then it satisfies a global Lipschitz condition on a set B in θ if its derivative is bounded on B and if B is convex.

Lemma 4.3. *If the order of integration with respect to x and differentiation with respect to θ can be interchanged in $\Psi(\theta)$ for θ in a neighborhood of θ^*, then $\Psi(\theta)$ is twice continuous differentiable in that neighborhood and its Hessian matrix is $\nabla_\theta^2 \Psi(\theta) = -J_q(\theta)$.*

Proof. Consider the score function as $U(x; \theta) = \nabla_\theta \log f(x; \theta)$ and the information matrix $I(x; \theta) = -\nabla_\theta^2 \log f(x; \theta) = \nabla_\theta U(x; \theta)$. The first derivative of

$\psi(x;\theta)$ is $f(x;\theta)^{(1-q)}U^{\mathrm{T}}(x;\theta)$. The second derivative is

$$\nabla_\theta[f(x;\theta)^{(1-q)}U^{\mathrm{T}}(x;\theta)] = (1-q)\frac{\nabla_\theta[f(x;\theta)]}{f(x;\theta)^q}U^{\mathrm{T}}(x;\theta) + f(x;\theta)^{(1-q)}I(x;\theta) \tag{4.7}$$

$$= f(x;\theta)^{(1-q)}[(1-q)U(x;\theta)U^{\mathrm{T}}(x;\theta) + I(x;\theta)] \tag{4.8}$$

The result follows from the given condition. □

The next theorem states the asymptotic normality of the MLqE.

Theorem 4.4. *Let θ^* be an interior point of Θ, and suppose the conditions of Lemma 4.3 and Lemma 4.3 hold. Moreover, assume that there is an integrable function $a(x)$ such that $|u_{jk}(x;\theta)f(x;\theta)^{2(1-q)}| < a(x)$ for $j,k = 1,\ldots,p$, where u_{jk} denotes the jk-th element of the matrix $U(x;\theta)U(x;\theta)^T$. Then, any sequence $\widehat{\theta}_{q,n}$ that is consistent for θ^* is such that*

$$\sqrt{n}(\widehat{\theta}_{q,n} - \theta^*) \xrightarrow{\mathcal{D}} N\left(0, J_q(\theta^*)^{-1}K_q(\theta^*)J_q(\theta^*)^{-1}\right) \tag{4.9}$$

where K_q and J_q are given in eq.(4.1) and eq.(4.2).

Proof. By Lemma 4.3, we can write the following Taylor expansion of $\psi(\theta)$:

$$\Psi(\theta) = \Psi(\theta^*) + \frac{1}{2}(\theta-\theta^*)^T\nabla^2_\theta\Psi(\theta^*)(\theta-\theta^*) + o(||\theta-\theta^*||^2). \tag{4.10}$$

By the Lipschitz condition (4.6), the desired result follows immediately from applying Theorem 5.21 in Van der Vaart [26]. □

Note that assumption in Lemma 4.3 the interchangability of integration and differentiation, implicitly implies the conditions for the existence of J_q. Such requirements are

$$E_{\theta_0}\left[i_{kj}(x;\theta)f(x;\theta)^{1-q}\right] < \infty, \quad \text{and} \quad E_{\theta_0}\left[u_{kj}(x;\theta)f(x;\theta)^{1-q}\right] < \infty, \tag{4.11}$$

where i_{kj} and u_{kj} are kj-elements of the matrices $I(x;\theta)$ and $U(x;\theta)U(x;\theta)^\intercal$, respectively. Existence of K_q is ensured by the assumptions of the theorem.

4.2 Standard errors

A convenient approach for computing standard errors is to use the influence function, which is shown to be proportional to the criterion function ψ ([17],[13]). For the MLqE, the influence function is $-J_q^{-1}(\theta^*)\left[f(x;\theta^*)^{1-q}U(x;\theta^*)\right]$ and Consistent estimates of the asymptotic variance of $n^{1/2}\widehat{\theta}_{q,n}$ can be obtained using Huber's sandwitch estimator (e.g., see Huber [17]). Let $k(x) = f(x;\widehat{\theta}_{q,n})^{1-q}U(x;\widehat{\theta}_{q,n})$. The variance estimator is

$$\widehat{\mathrm{Var}}(\widehat{\theta}_{q,n}) = (n-1)^{-1}\widehat{J}_q^{-1}\sum_{i=1}^{n} k(X_i)k(X_i)^\intercal \widehat{J}_q^{-1}, \tag{4.12}$$

where $\widehat{J}_q$ is obtained by replacing $\widehat{\theta}_{q,n}$ in the expression of the influence function and taking expectation with repect the empirical distribution F_n. Estimates of the variance of the MLqE and confidence intervals can be computed also using other standard techniques such as bootstrap.

4.3 Exponential Families

In many cases, the target parameter θ^* can be easily computed, as it the case for exponential families. Consider densities of the form $f(x;\theta) = \exp\left\{\theta b(x) - A(\theta)\right\}$, where $\theta \in \Theta$ is the natural parameter and $A(\theta) = \log\int_\Omega \exp\left\{\theta b(x)\right\} d\mu(x)$ is the cumulant generating function (or log normalizer).

Lemma 4.5. *Let $\Psi(\theta)$ be as in eq.(4.3). If $f(x;\theta)$ is an exponential family and the conditions given in Lemma 4.3 are satisfied, then $\theta^* = \theta_0/q$ maximizes $\Psi(\theta)$.*

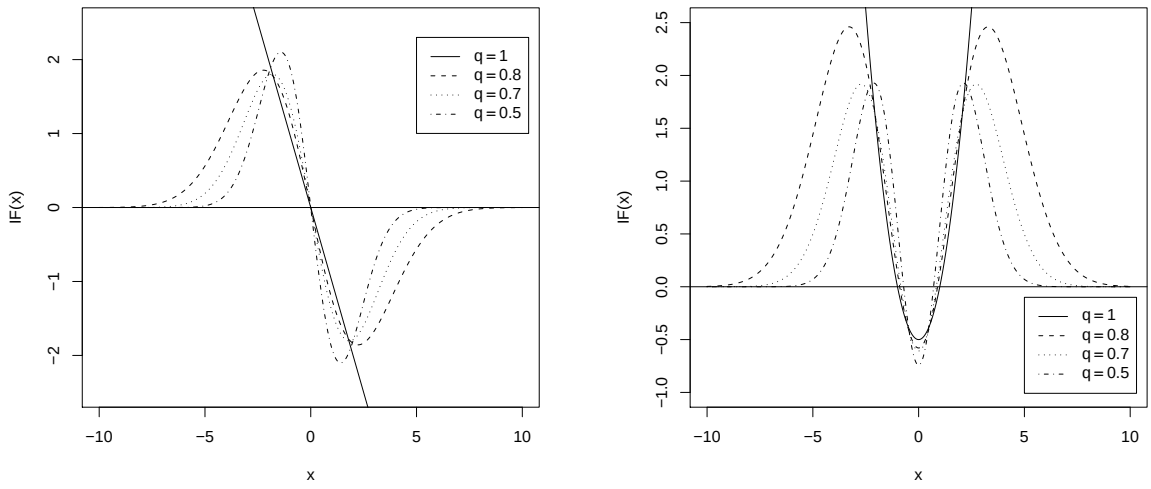

Figure 1: Influence functions for estimating the mean (left) and the standard deviation (right) of a standard normal distribution for various choices of q.

Proof. The first derivative of $\Psi(\theta)$ is

$$\nabla_\theta \Psi(\theta) = \int_\Omega \frac{\nabla_\theta [f(x;\theta)]}{f(x;\theta)^q} f(x;\theta_0) d\mu(x). \tag{4.13}$$

For the families we have that $f(x;\theta_0/q) \propto f(x;\theta_0)^{1/q}$. Thus, by the conditions, we have $\nabla_\theta \Psi(\theta^*) = \nabla_{\theta^*} \int_\Omega f(x;\theta^*) d\mu(x) = 0$. The second derivative is

$$\nabla^2_\theta \Psi(\theta) = \int_\Omega \frac{\nabla^2_\theta f(x;\theta)}{f(x;\theta)^q} f(x;\theta_0) d\mu(x) - q \int_\Omega \frac{[\nabla_\theta f(x;\theta)]^\intercal [\nabla_\theta f(x;\theta)]}{f(x;\theta)^{q+1}} f(x;\theta_0) d\mu(x). \tag{4.14}$$

The first addend of the above expression evaluated at θ^* becomes $\int_\Omega \nabla^2_\theta f(x;\theta) d\mu(x)$. Since differentiation can be passed under integration, we have that $\nabla^2_\theta \int_\Omega f(x;\theta) d\mu(x) = 0$. The second addend is clearly negative semi-definite. Thus, we obtained

that $\nabla^2_\theta \Psi(\theta^*)$ is positive semi-definite. Hence, $\theta^* = \theta_0/q$ is a maximum. □

Since for exponential families the target parameter is just θ_0/q, one can consider $q\widehat{\theta}_{q,n}$, a bias-corrected version of the MLqE. An important example is when $q = 1/2$. Eq. (2.12) points out that such a choice for q corresponds to finding θ that minimizes an empirical version of the Hellinger distance between $f(x;\theta)$ and the zooming transformation $f(x;\theta_0)^{(1/2)}$. Hence, $f(x; 2\widehat{\theta}_{1/2,n})$ gives a Hellinger-type of estimate which does not involve kernel smoothing and all the computational costs related to the bandwith choice. However, simulations for variuous settings of q and n using data from several univariate distributions showed that the mean squared error for the uncorrected MLqE is generally smaller than that of the corrected version. This happens to be the case when the sample size is small or moderate. Insights on this aspect will be given in next sections.

4.4 Trade-off between bias and varicance

4.4.1 Asymptotic calculations

Consider an exponential family and compare the asymptotic mean squared error of MLqE for $q = 1$ with the case when $q \neq 1$. When $q \to 1$, the formula of the asymptotic variance involving J_q and K_q in Theorem 4.4 becomes the inverse of the Fisher information. Thus, the expression the ratio of the asymptotic mean squared errors is computed as

$$\Lambda(q, n; \theta_0) := \frac{aMSE(1, n; \theta_0)}{aMSE(q, n; \theta_0)} = \frac{\text{Tr}\left(J_1(\theta_0)^{-1} K_1(\theta_0) J_1(\theta_0)^{-1}\right)}{n||\theta^* - \theta_0||^2 + \text{Tr}\left(J_q(\theta^*)^{-1} K_q(\theta^*) J_q(\theta^*)^{-1}\right)}, \tag{4.15}$$

where $\text{Tr}(\cdot)$ is the trace operator. The quantity $\Lambda(q, n; \theta_0)$, to be called bias-adjusted relative efficiency, can be used to judge how much is gained/lost relative to the MLE under the model conditions. This is more conveniently explored on a case-by-case basis, as shown in the next two examples.

Example 4.1. Consider the exponential ditribution with density $\theta_0 \exp\{-x\theta_0\}$, $x > 0$, $\theta > 0$. Ferrari and Yang [11] computed the J_q and K_q, obtaining

$$\theta^* = \theta_0/q, \quad J_q(\theta^*) = -\frac{q^3}{\theta_0(\theta_0/q)^q}, \quad K_q(\theta^*) = q\left[\frac{q^2-2q+2}{(2-q)^3}\right]\left(\frac{\theta_0}{q}\right)^{-2q}. \tag{4.16}$$

Thus, the MLqE has squared bias $\theta_0^2(1-1/q)^2$ and asymptotic variance

$$J_q(\theta^*)^{-2}K_q(\theta^*) = \theta_0^2\frac{q^2-2q+2}{q^5(2-q)^3} \tag{4.17}$$

When $q = 1$, we recover the MLE with asymptotic variance θ_0^2 and the bias-adjusted relative efficiency is

$$\Lambda(q,n) = \left[n\left(\frac{1-q}{q}\right)^2 + \frac{q^2-2q+2}{q^5(2-q)^3}\right]^{-1}, \tag{4.18}$$

which turns out to be independent from θ_0.

Example 4.2. Consider a scale normal $N(0,\theta_0^2)$. In this case the target parameter is $\theta^* = \sigma\sqrt{q}$. and the squared asymptotic bias has expression $\theta_0^2(1-\sqrt{q})^2$. The calculation in appendix A shows that the asymptotic variance is

$$J_q(\theta^*)^{-2}K_q(\theta^*) = \theta_0^2\frac{(3-2q+q^2)}{4(2-q)^{5/2}q^{3/2}} \tag{4.19}$$

When $q = 1$ we have the usual MLE with variance $\theta_0^2/2$. Thus, the the bias-adjusted relative efficiency is

$$\Lambda(q,n) = \left(2n(1-\sqrt{q})^2 + \frac{(3-2q+q^2)}{2(2-q)^{5/2}q^{3/2}}\right)^{-1} \tag{4.20}$$

which, as for the case of the exponential distribution does not depend on the

true value of the parameter.

In Fig.2, we represent the relative efficiency between MLE and MLqE corresponding to various choices of the sample size for the previous two examples. When the sample size is small there are values of q that allow for a bias-adjsted efficiency larger than 1.

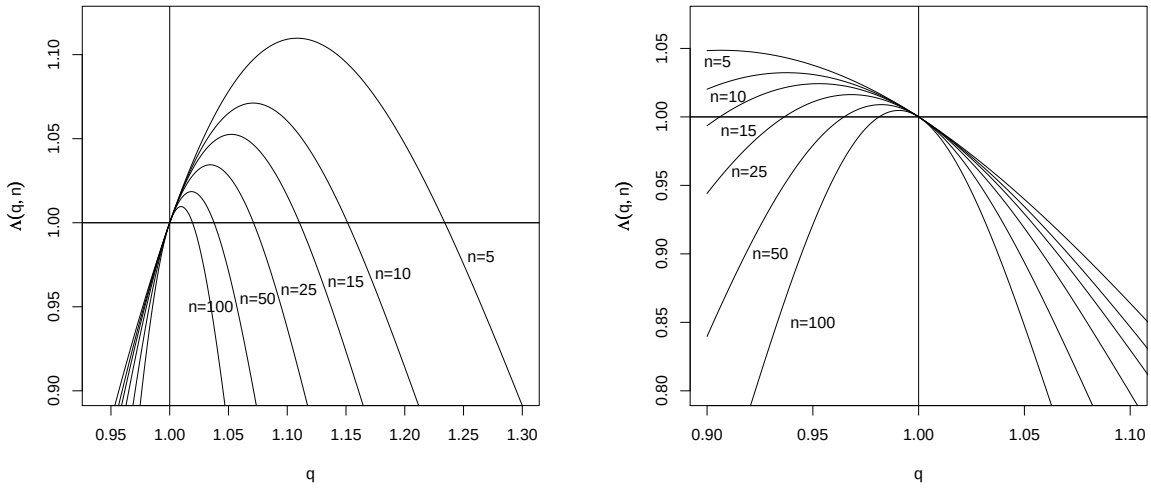

Figure 2: Bias-adjusted relative efficiency betwen MLE and MLqE for different sample sizes as in eq.(4.18) and in eq.(4.20), for an exponential (left panel) and a scale normal (right panel).

4.4.2 Finite sample efficiency

One might ask whether the above asymptotic considerations can actually help to decide the value of the distortion parameter when the sample size is moderate or small. Although we do not provide an analytical answer to such a question at the moment, numerical simulations performed for the scale normal and the exponential distributions indicate that the actual relative efficiency is bounded from below by $\Lambda(q, n)$. A representation of this phenomenon is

given in Fig.4.4.2, where the ratio of the Monte Carlo mean squared errors of the MLE over that MLqE $\widehat{R}(q,n) = \sum_{b=1}^{B}(\widehat{\theta}_{1,n} - \theta_0)^2 / \sum_{b=1}^{B}(\widehat{\theta}_{q,n} - \theta_0)^2$ is compared to the asymptotic relative efficiency $\Lambda(q,n)$ (solid line) for various choices of the sample size. Hence, a choice of q based on maximization of bias-adjusted relative efficiency is expected to be a safe but rather conservative choice.

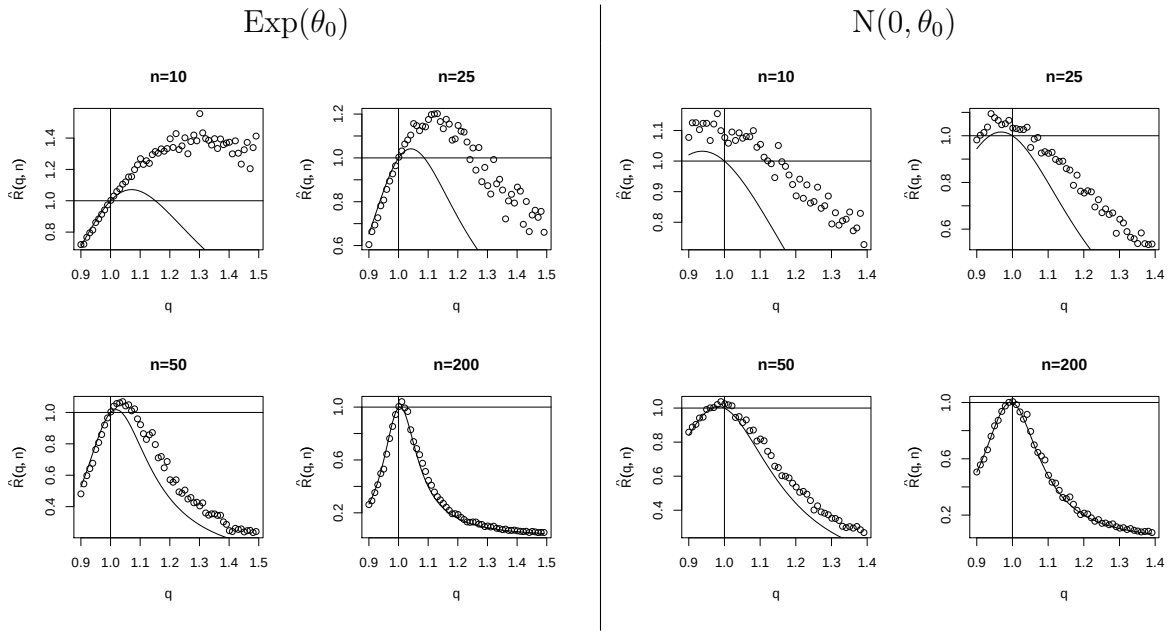

Figure 3: Monte Carlo relative efficiency between MLqE and MLE against q for an exponential (left panel) and a normal for various sample sizes. The solid line is the bias-adjusted relative efficiency as in eq.(4.18) and in eq.(4.20). The Monte Carlo sample size is 1500.

5 Re-weighting algorithm

One of the main perks of the MLqE is that a simple and fast algorithm is automatically available. Fixed q, eq.(3.3) tells us that the estimation problem can be formulated in terms of a weighting process. Let $s \in \{0, 1, \dots\}$ denote

the iteration step.

1. If $s = 0$, set $\theta^{(0)} = \widehat{\theta}_{1,n}$. Note that here $q = 1$ and the initial estimate is set to be the maximum likelihood estimate.

2. For $s > 0$,

$$\theta^{(s+1)} = \left\{ \theta : \sum_{i=1}^{n} w^* \left(X_i; \theta^{(s)} \right) U(X_i; \theta) = 0 \right\}, \tag{5.1}$$

   where $U(x; \theta)$ is the score function and $w^*(X_i; \theta) := f(X_i; \theta)^{1-q} / \sum_{i=1}^{n} f(X_i; \theta)^{1-q}$.

In many important cases, the steps of the algorithm reduce to a straighforward variable transformation, as it illustrated in the following two examples.

Example 5.1. Exponential distribution The initial value is given by $\widehat{\theta}^{(0)} = \overline{X}^{-1}$. The solution at the s-th step is $\widehat{\theta}^{(s)} = (\sum_{i=1}^{n} w_i X_i)^{-1}$, where

$$w_i = \left[\sum_{j=1}^{n} \exp \left\{ -(X_j - X_i) \widehat{\theta}^{(s-1)} (1 - q) \right\} \right]^{-1}. \tag{5.2}$$

Example 5.2. Multivariate normal distribution with unkown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Let $s \in \{0, 1, \ldots, s^*\}$ denote the iteration step.

1. Initialize, setting $\widehat{\boldsymbol{\mu}}^{(0)} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}}^{(0)} = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^{(0)})^{\intercal} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^{(0)})$, i.e., compute the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

2. For $0 < s < s^*$, $\widehat{\boldsymbol{\mu}}^{(s)} = \sum_{i=1}^{n} w_i^{(s-1)} \mathbf{x}_i$ and $\boldsymbol{\Sigma}^{(s)} = \sum_{i=1}^{n} w_i^{(s-1)} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^{\intercal} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$,

where

$$w_i^{(s)} = \frac{f(\mathbf{x}_i; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})^{1-q}}{\sum_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})^{1-q}} \tag{5.3}$$

Finally, if asymptotically unbiased estimates are desired, one can set $\widehat{\boldsymbol{\mu}}^{(s^*)} = \widehat{\boldsymbol{\mu}}^{(s^*-1)}$ and $\widehat{\Sigma}^{(s^*)} = q \widehat{\Sigma}^{(s^*-1)}$.

The algorithm converges quickly, typically in less than 15 iterations. To gain some insight on this behavior, we use an argument analogous to that proposed by Windham [28]. First, note that the reweighting procedure computes a fixed point, which is a solution to $\tau = h(\tau)$. The iterating function is such that $E_{F_n}\left[f(X;\tau)^{1-q}U(X;h(\tau))\right] = 0$, where F_n is the empirical distribution. Differentiating with respect to τ, gives

$$\begin{aligned}\nabla_\tau h(\tau) = (q-1)\left\{E_{F_n}\left[f(X;\tau)^{1-q}I(X;\tau)\right]\right\}^{-1} \\ \times E_{F_n}\left[f(X;\tau)^{1-q}U(X;\tau)^{\intercal}U(X;h(\tau))\right].\end{aligned} \tag{5.4}$$

The above derivative can be restated as

$$\nabla_\tau h(\tau) = (1-q)W(\tau)\left[I + (1-q)W(\tau)\right]^{-1}, \tag{5.5}$$

where

$$\begin{aligned}W(t) = -\left\{E_{F_n}\nabla_\tau\left[f(X;\tau)^{1-q}U(X;h(\tau))\right]\right\}^{-1} \\ \times E_{F_n}\left[f(X;\tau)^{1-q}U(X;\tau)^{\intercal}U(X;h(\tau))\right]\end{aligned} \tag{5.6}$$

Data near the true model, say $dF_n(x) = f(x;\theta_0)$, result in $E_{\theta_0}\left[f(x;\theta^*)^{1-q}U(x;\theta^*)\right] = 0$, where θ^* is the target parameter, depending on θ_0 and q, satisfying $f(x;\theta^*) = f(x;\theta_0)^{(q)}$. One can show that differentiating with respect the parameter at θ_0 leads to $W(\theta_0) = q^{-1}I$. By substituting in eq.(5.5) we obtain a diagonal matrix with diagonal elements equal to $(1-q)$. The local convergence rate is related to the largest eigenvalue of (5.5) at the solution (e.g., see Johson and Riess [18], p. 192). Therefore, if the empirical distribution of the data is close to the true model, we should anticipate a linear convergence rate $r \approx |1-q|$. In addition, the closer is the distortion parameter to 1, the faster is the algorithm.

Figure 5 illustrates the convergence rates of the REMLq algorithm for q ranging from 0.5 to 1.5. The dotted lines correspond to 10 samples from an

Exp(1). As the sample size increases, the empirical distribution of the data approximates better the true model. As a result, the estimated convergence rate of the algorithm gets closer to $|1-q|$.

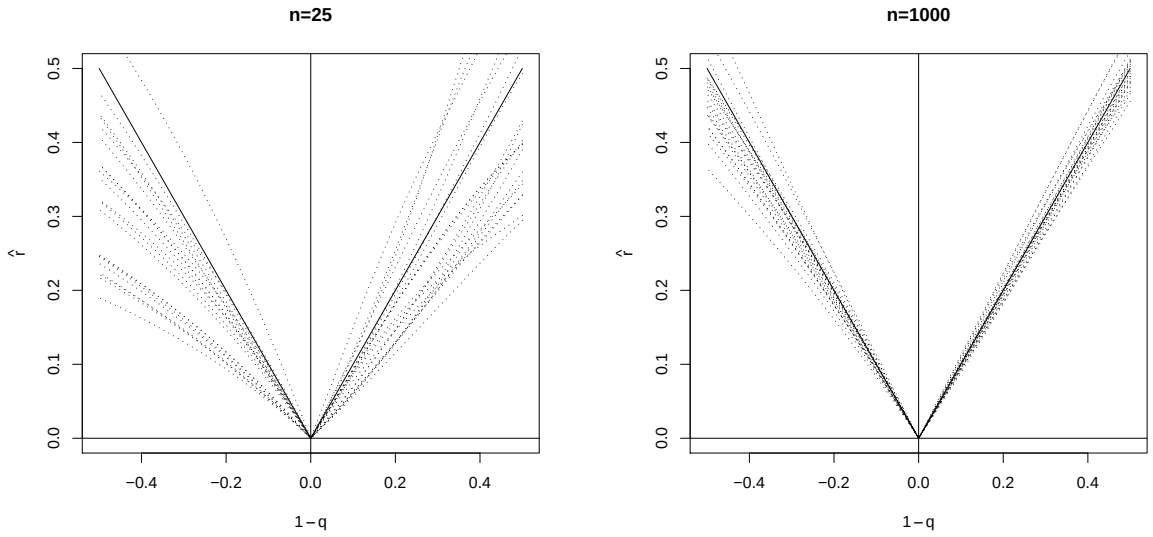

Figure 4: The dotted lines are the estimated convergence rates of the REMLq algorithm for 20 samples of size 25 (left panel) and 1000 (right panel) from an Exp(1). The solid line corresponds to $\widehat{r} = |1-q|$.

5.1 Selecting the distortion parameter q

An important issue in applications is the selection of the distortion parameter, as it leads to different divergence measures and can potentially alterate the trade-off between efficiency and robustness of the estimator. We discuss three possible strategies to be used based on the goals of the experimenter.

Strategy 1: Bias-adjusted efficiency. The first approach takes advantage of the variance reduction properties of the MLqE at the model and its capability to improve upon the MLE by reducing the variance at expenses

of a slightly increased bias, when the sample size is moderate or small. A reasonalbe criterion is to choose q such that $q^* = \arg\min_q \{\Lambda(q, n)\}$. When the asymptotic distribution of the MLqE is available, this method has the advantage to be computationally inexpensive.

Strategy 2: REMLq automatic choice of q by Windham's criterion. Unlike other optimtization methods the convergence rate of the re-weighting algorithm described in section 5 yields a statistical interpretation. Evalulate expression (5.6) at $\widehat{\theta}_{q,n}$, obtaining

$$W(\widehat{\theta}_{q,n})^2 = \widehat{J}_q(\widehat{\theta}_{q,n})^{-1} E_{F_n}\left[f(X;\widehat{\theta}_{q,n})^{1-q} U(X;\tau)^{\intercal} U(X;\widehat{\theta}_{q,n})\right]^2 \widehat{J}_q(\widehat{\theta}_{q,n})^{-1} \tag{5.7}$$

Where $\widehat{J}_q(\widehat{\theta}_{q,n}) = E_{F_n}\nabla_\theta\left[f(X;\widehat{\theta}_{q,n})^{1-q} U(X;\widehat{\theta}_{q,n})\right]$ is an estimate of the matrix $J_q(\theta^*)$. By Schwartz inequality

$$\left(E_{F_n}\left[f(X;\widehat{\theta}_{q,n})^{1-q} U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right)^2 \tag{5.8}$$

$$\leq \left(E_{F_n}\left[f(X;\widehat{\theta}_{q,n}))^{2(1-q)} U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right) \tag{5.9}$$

$$\times \left(E_{F_n}\left[U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right), \tag{5.10}$$

i.e. an estimate of the matrix $K_q(\theta^*)$ times the Fisher information. Therefore,

$$W(\widehat{\theta}_{q,n})^{-2} \geq \widehat{I}_q(\widehat{\theta}_{q,n})\left[\widehat{J}_q(\widehat{\theta}_{q,n})\widehat{K}_q(\widehat{\theta}_{q,n})\widehat{J}_q(\widehat{\theta}_{q,n})\right] \tag{5.11}$$

and $W(\widehat{\theta}_{q,n})^{-2}$ is an empirical upper bound for efficiency. The above calculation enlights that the convergence rate of the algorithm contains information about the efficiency of the estimates through equation (5.5). Windham [28] considered equating diagonal elements of (5.5), say w to $\widehat{r}$, an estimate of the convergence rate. By solving for w one obtains $w = \widehat{r}/[(1-q)(\widehat{r}-1)]$, which holds $w^{-2} = (1-q)^2\left(\widehat{r}^{-1}-1\right)^2$. In practice, for choices of distor-

tion paramters in a grid $Q_k = \{q_1, \ldots, q_k\}$, correponding convergence rates convergence rate can be computed as

$$\widehat{r}_{q_j} = \frac{||\widehat{\theta}_{q_j}^{(S)} - \widehat{\theta}_{q_j}^{(S-1)}||}{||\widehat{\theta}_{q_j}^{(S-1)} - \widehat{\theta}_{q_j}^{(S-2)}||}, \ 1 \leq j \leq k, \tag{5.12}$$

where $\widehat{\theta}^{(S)}$ is the last step of the algorithm. The distortion parameter is then selected according to

$$\widehat{q} = \underset{q \in Q_k}{\arg\min} \ \left\{ (1-q)^2 \left(\widehat{r}_q^{-1} - 1 \right)^2 \right\}. \tag{5.13}$$

Strategy 3: Parametric Bootstrap. Besides the above strategies, data-driven procedures for the estimation of q aimed at the minimization of the generalization error are also viable candidates. In particular, we recommend bootstrap techniques over other methods such as leave-one-out or k-fold cross-validation. In cross-validation, it is customary to divide the original sample in two parts: a training set and a testing set smaller than the total sample size. However, because of the relationship between the sample size and the value of the optimal distortion parameter in MLq estimation, this may cause biased estimates of q. Of course, the situation may be particularly serious when the size of the sample under exam is moderate or small.

6 Numerical studies

6.1 Examples

The following two examples demonstrate the performance of the estimator on two real datasets.

Example 6.1. In this example, we consider $n = 799$ observations of time intervals (in seconds) between successive pulses along a nerve fibre in Hand

et al.[14] (dataset 160). The goal of this example is to show that MLqE is superior to MLE for estimating the exponential rate, when a small or moderate sample size is considered. Inspections on the data shows that an exponential distribution is appropriate. Since there are no evident outliers, the selection of the distortion parameter is based on the bias-adjusted efficiency criterion. Not surprisingly, the ML and MLq estimates for the whole dataset are very close: $\widehat{\theta}_{q^*,n} = 4.37$ (se =.16) with optimal distortion parameter $q^* = 1.05$ and $\widehat{\theta}_{1,n} = 4.58$ (se =.16).

A simple hold-out procedure is then employed for evaluating the performance of the two estimators in small or moderate samples. We draw $B = 250$ subsamples of size $n^* < n$ from the original sample and computed the quadratic error $\mathcal{E}(q, n^*) := B^{-1} \sum_{b=1}^{B} (\widehat{\theta}_{q,n^*} - \widehat{\theta}_{1,n})^2$. The results in table

$n^* =$	10	15	25	50	100	200	400
$\mathcal{E}(1, n^*)$	7.66	7.14	5.13	3.35	2.99	2.76	2.52
$\mathcal{E}(q^*, n^*)$	6.32	5.88	4.39	2.99	2.81	2.66	2.51
q^*	1.071	1.051	1.036	1.021	1.011	1.006	1.001
Gain (%)	17.47	17.72	14.41	10.78	6.13	3.66	0.64

Table 1: Hold-out validation error of MLqE and MLE for estimating Exp(θ) in the nerve pulse data set. The last row indicates the percent gain of MLqE over the MLE.

?? illustrate that setting q slightly larger than one improves the accuracy. The gain is sensible when the sample size is small and persist even for larger samples.

Example 6.2. In this example, we apply our method to Newcomb's dataset, representing 66 measurements of the passage time of light. Among others, Brown and Hwang [7], Basu et al. [3] and Bhandari et al. [6] analyzed this dataset under a normal model $N(\mu, \sigma)$, as it will be the case here. Since the data present strong outliers at -44 and -2, the selection of q is performed by the criterion function based on the estimated convergence rate of

	With outliers				W/o outliers			
	$\widehat{q} = 0.83$	$q = 1$	$q = 1/2(*)$	MHDE	$\widehat{q} = 1.02$	$q = 1$	$q = 1/2(*)$	MHDE
$\widehat{\mu}$	27.65(0.63)	26.21(1.32)	27.25(0.65)	27.46	27.76(0.64)	27.75(0.64)	27.25(0.65)	27.40
$\widehat{\sigma}$	4.63(0.52)	10.66(3.52)	4.34(1.15)	4.98	5.09(0.45)	5.04(0.46)	4.34(1.15)	4.84

Table 2: Estimated parameters for the Newcomb data and their standard errors (in parenthesis). The cases $q = 1$ and $q = 1/2$ correspond to maximum likelihood and Hellinger distance estimates, respectively. The last line shows the bias-adjusted asymptotic efficiency of the estimators compared to that of MLE. $(*)$ Estimates have been obtained by adjusting the MLqE for its asymptotic bias.

the MLqE. Table 6.2 presents the MLqE estimates of μ and σ for different choices of the distortion parameter: $\widehat{q}$ denotes the estimated optimal value of the distortion parameter, $q = 1$ and $q = 1/2$ correspond to maximum likelihood and Hellinger distance esimates. Note that for finding Hellinger distance estimates we adjusted the estimator for its asymptotic bias. Namely, the Hellinger distance estimates of μ and σ are $\widehat{\mu}_{1/2,n}$ and $\sqrt{2}\ \widehat{\sigma}_{1/2,n}$. The analyses were also repeated after leaving out the two evident outliers. With the outliers, MLqE shows remarkable robustness properties compared to the MLE. In particular, the estimates for (μ, σ) of $(27.65, 4.63)$ are very close to to those based on L_2 distance computed by Brown and Hwang ([7], p.254) and Basu et al. ([3], p.557), who found $(27.38, 4.67)$ and $(27.29, 4.67)$ respectively. Bhandari et al. found similar value for the minimum generalized negative exponential density estimator and for the hellinger distance estimator based on kernel smoothing ([6], p. 105). Without the outliers, MLqE adapts well to the data and selects $\widehat{q}$ near 1, resulting in estimates close to the MLEs and giving about the same efficiency. A visual representation of this is supplied in Fig. ??, where fitted normal densities are superimposed to the histograms of Newcomb data. In presence of outliers, the curve corresponding to $\widehat{q} = 0.83$ fits the body of the histogram better than the other cases. When the outliers are left out, the MLqE and MLE are basically identical.

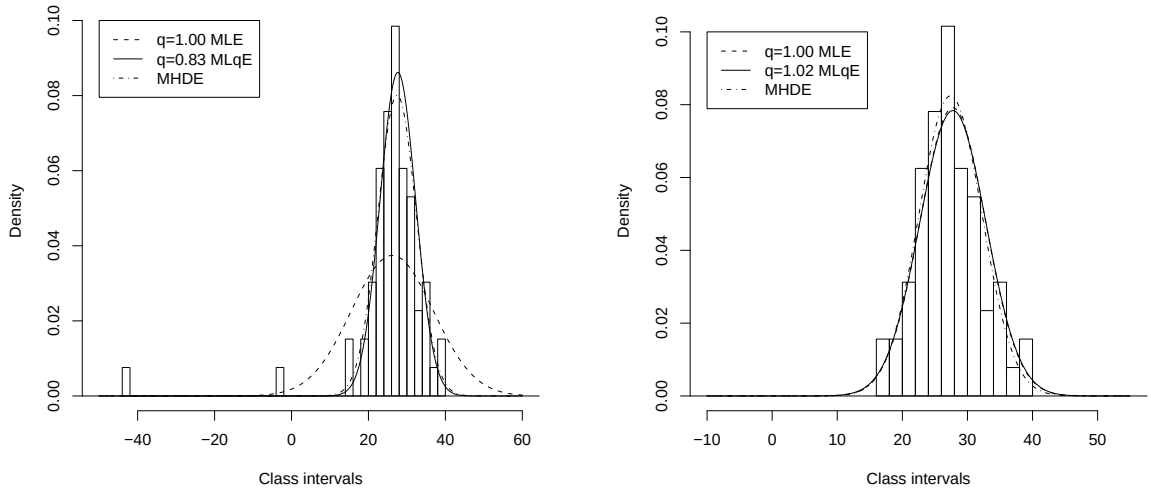

Figure 5: Histograms of the Newcomb data with outliers (left panel) and without (right panel) with normal densities fitted using maximum Lq-likelihood (MLqE), maximum likelihood (MLE) and minimum Hellinger distance (MHD). The distortion parameter for MLqE is computed using Windham's crieterion.

6.2 Simulations

Contaminated normal model. We conducted a simulation study for the $N(\mu,\sigma)$ model and computed the Monte Carlo mean, variance and mean square error for MLE, MLqE and MHDE, under different contaminated models of the form $\tau N(\mu,\sigma) + (1-\tau)N(\mu_c,\sigma)$, where $\mu_c = \mu + 4$, $\sigma = 1$. We considered $\tau = 0$ and 0.05 and $n = 10, 25, 50, 100$. When contaminaticon is included, the samples present nonobvious outliers on the right of the bulk of the data. To decide the value of q, we employ Windham's criterion. The MDHE is implemented using the automatic kernel density function with the Epanechnikov kernel ($w(x) = 0.75(1-x^2)$, if $|x| < 1$, and $w(x) = 0$, otherwise) and bandwith $h = c_n s_n$, where $c_n = 0.5$ and $s_n = (0.6745)^{-1}\text{median}(|X_i - \text{median}(X_i)|)$ (e.g., see Bhandari et al. [5]). We tried other kernels and methods for the bandwidth choices obtaining results similar to those reported here. In our analysis, we also consider the fully nonparametric version of the minimum hellinger distance estimator by computing the MLqE with $q = 1/2$ and adjusting the estimates for the asymptotic bias. In all the experiments, the Monte Carlo sample size is $B = 5000$.

The results in in Table 6.1 suggest that the MLqE performed well whether or not contamination was present. For each simulation setting we the report mean squared error of the estimates, computed as $B^{-1}\sum_b(\widehat{\theta}_b - \theta_0),\theta_0^{\intercal} = (0,1)$, along with its components: the squared bias and the variance. Without contamination, we obtained values of q close to 1. As a consequence, the mean squared error of MLqE occurred to be close to that of MLE. Note that the minimum hellinger distance estimates – both kernel smoothing and MLqE with $q = 1/2$ – tend to be substantially less efficient than the other methods when contamination is absent. When contamination is included, we estimated $1/2 < \widehat{q} < 1$ and the MLqE outperformed not only the MLE but also the MHDE by balancing the trade-off between efficiency and robustness for

all sample sizes. Clearly, both types of minimum hellinger distance estimators do better than MLE in this setting, as the latter is highly nonrobust. It is worth noticing that the $\widehat{q}$ changes towards 1 as the sample size grows in both contaminated and clear data.

Finally, compare the kernel-smoothed MHDE with our fully nonparametric version. The two estimators performed similarly and, as expected, their efficiency tended to be the same for larger samples. Note, that in very small samples, a properly performormed kernel smoothing yields better results, due to the additional flexibility given by the bandwidth selection. However, depending on the choice of the kernel and the bandiwith selection criterion, MHDE can give diverse results when small samples are considered. In most cases, we found that the kernel-smoothed MHDE performed comparably to our method.

Efficiency and choice of q. A second numerical study aimed to explore the behavior of the MLqE when data are sampled from a Exp(1) model. Here, we disregard robustness and focus on assessing the efficiency of MLqE. The performance of MLqE is gauged using $\widehat{R}(q,n)$, the ratio between the Monte Carlo mean squared error of MLE over that of MLqE for sample sizes $n = 5, 15, 25, 50, 100$. When estimating the MLqE, we consider choosing q using both Windham and the bias-adjusted relative efficiency criteria. The standard errors for $\widehat{R}(q,n)$ are computed via the Delta method. The results in Table **??** show that for small or moderate sample sizes, $\widehat{R} > 1$ meaning that the MLqE is more efficient than MLE. However,note that when q is chosen by Whindam's criterion the gain is more modest than the case when the asymptotic criterion is used. Fig.**??** shows $\widehat{R}(q,n)$ corresponding to numerous choices of q on the horizontal axis for various sample sizes. The superimposed solid line represents the bias-asjusted relative efficency between MLqE and MLE in eq.(). One can see that the optimal best values of q based on the Monte Carlo simulations tend to be greater than the maximum for the solid

	$N(0,1)$				$0.95N(0,1)+0.05N(0,4)$			
n	Bias^2	Var	MSE	$\widehat{q}$	Bias^2	Var	MSE	$\widehat{q}$
				MLqE				
15	0.0015	0.0990	0.1005	1.0550	0.0119	0.1714	0.1832	0.8322
25	0.0001	0.0606	0.0608	1.0500	0.0089	0.0988	0.1077	0.8297
50	0.0000	0.0303	0.0303	1.0473	0.0113	0.0442	0.0555	0.8511
100	0.0001	0.0149	0.0151	1.0454	0.0162	0.0207	0.0369	0.8646
				MLE ($q=1$)				
15	0.0031	0.0982	0.1013		0.0993	0.2455	0.3447	
25	0.0009	0.0603	0.0611		0.1142	0.1562	0.2704	
50	0.0002	0.0301	0.0302		0.1245	0.0773	0.2017	
100	0.0001	0.0148	0.0149		0.1401	0.0381	0.1782	
				MLqE*($q=1/2$)				
15	0.0224	0.2214	0.2437		0.0196	0.2524	0.2720	
25	0.0064	0.1301	0.1365		0.0072	0.1451	0.1523	
50	0.0015	0.0595	0.0609		0.0012	0.0643	0.0655	
100	0.0004	0.0274	0.0278		0.0000	0.0305	0.0305	
				MHDE				
15	0.0005	0.1484	0.1489		0.0091	0.2185	0.2276	
25	0.0001	0.0964	0.0965		0.0088	0.1325	0.1413	
50	0.0003	0.0482	0.0485		0.0078	0.0618	0.0696	
100	0.0004	0.0239	0.0243		0.0085	0.0302	0.0387	

Table 3: Monte Carlo squrared Bias, variance and mean square error of the MLqE, MHD and MLE of μ and σ for sample sizes 15,25,50,100 and 250 under clear and contaminated normal model.

6.2

	$n =$	5	15	25	50	100
$\widehat{R}$		1.174(.008)	1.134(.005)	1.103(.004)	1.055(.004)	1.030(.003)
q^* (Adj-Eff)		1.108	1.052	1.034	1.019	1.010
$\widehat{R}$		1.049(.002)	1.068(.002)	1.068(.004)	1.054(.006)	0.988(.009)
$\widehat{q}$ (Windham)		1.071	1.043	1.038	1.034	1.032

Table 4: Monte Carlo relative efficiency between MLqE and MLE for various sample sizes. Asymptotic bias-adjusted relative efficiency (Adj-Eff) and Whindam's criterion are employed for choosing q.

line.

This findings indicate that for smaller samples the asymptotic criterion is too conservative and it can be further improved. Thus, a last set of simulations was devoted to investigate whether the choice of q via bootstrap can improve further the efficiency of MLqE in small or moderate samples. Given a grid of distortion parameters q, we generated Monte Carlo samples from an Exp(1) and for each sample selected the optimal value of q by minimizing a bootstrap estimate of the mean squared error based on 250 bootstrap repetitions. The procedure is repeated for $n = 15, 25, 35, 50, 75, 150, 250$. In Fig., we plot the Monte Carlo estimates of the optimal qs chosen via boostrap along with: (i) the true optima, i.e. the values that minimize the Monte Carlo mean squared error and (ii) optimal q based on minimization of asymptotic mean squared error **??**. Overall, parametric boostrap approximates better the true optima and does sensibly better than the asymptotic criterion for sample sizes of 25 or larger.

Generalized linear models with covariates Our technique for parametric density estimation can be easily extended to generalized linear models (GLM). In the current experiment, we consider a response variable Y, from an exponential distribution with rate $\exp(-X\beta)$, where X denote the

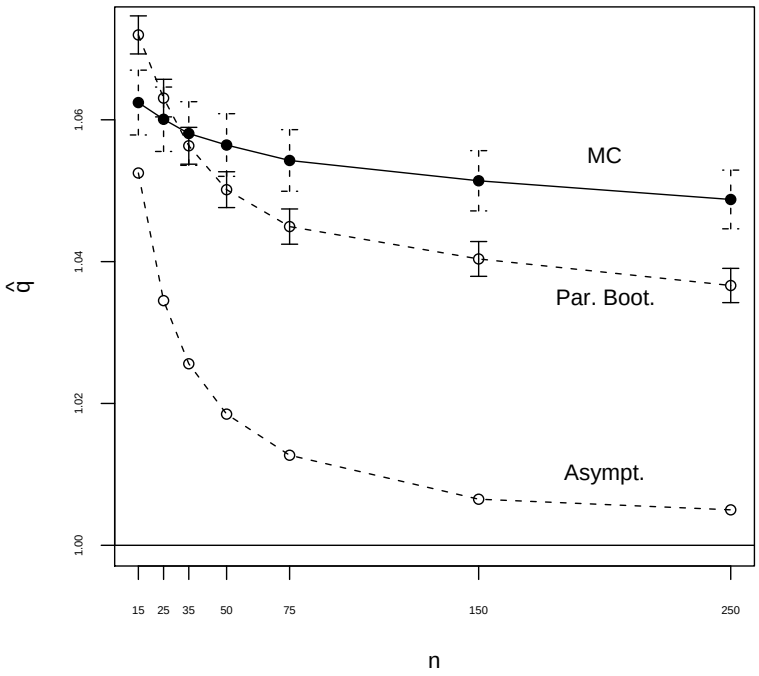

Figure 6: True values of q (MC) along with estimates of the optimal q via parametric bootstrap (Par.Boot) and minimizing eq. (Asympt.). The vertical segments represent 95% cofindence intervals.

p covariates and β is the vector of coefficient to be estimated. The simulations are structured as follows. Initially, we randomly draw n design points $x_1, \ldots, x_n$ from the hypercube $[-1, 1]^p$ and keep them fixed throughout the study. The vector of the true coefficients β_0 is generated from a Unif$[-1, 1]^p$. Then, we cosider 1500 Monte Carlo samples of size n of the response from Exp(exp($X\beta_0$)) and for each Monte Carlo sample we compute the MLqE of β_0. The choice of q is performed minimizing the bias-adjusted relative efficiency. The performance is finally gauged by comparing Monte Carlo estimates of the residuals sum of squares:

$$RSS.ratio = \frac{RSS(n,1)_{MC}}{RSS(n,q)_{MC}} = \frac{\sum_{b=1}^{B}\sum_{i=1}^{n}(\log(y_i) - x_i\widehat{\beta}_{1,n})^2}{\sum_{b=1}^{B}\sum_{i=1}^{n}(\log(y_i) - x_i\widehat{\beta}_{q,n})^2} \tag{6.1}$$

The Monte Carlo standard error of $\widehat{R}(q,n)$ is computed using the Delta method (CITE). The study is repeated for $p = 2, 4, 8, 16$ and $n = 25, 50, 100, 500$.

Finally, we assess the prediction error of the MLqE compared to that of MLE using leave-one-out cross validation. We consider samples of size n of (X, Y) generated as described above. The parameter values β are also generated analogously. Let us denote by (X_{-i}, Y_{-i}), the training data set, obtained by excluding the ith point. From the training set, we calculate the MLqE and MLE of β denoted by $\widehat{\beta}_{q,n}^{(-i)}$. An estimate of the prediction error based on a squared lost is computed as

$$PE(n,q) = n^{-1}\sum_{i=1}^{n}\left(\log(y_i) - x_i\widehat{\beta}_{q,n}^{(-i)}\right)^2. \tag{6.2}$$

We repeat the hold-out validation for different numbers of covariates $p = 2, 4, 8, 10$ and sample sizes $n = 25, 50, 100$.

p	$n =$	25	50	100	500
2		2.4571(0.0149)	2.3461(0.0092)	2.3014(0.0059)	2.3023(0.0026)
4		3.1601(0.0341)	2.7018(0.0142)	2.9942(0.0117)	2.8152(0.0049)
8		6.7402(0.2364)	3.1389(0.0221)	3.3599(0.0156)	2.9757(0.0056)
16		19.9716(0.8626)	22.2592(1.0620)	6.1786(0.0731)	6.0593(0.0233)

Table 5: Monte Carlo estimates of the ratio of the residuals sum of square of MLE over that of MLqE as in (6.1). In parenthesis we report the standard error of the Monte Carlo estimates, computed using the Delta method. The Monte Carlo sample size is 1500.

		MLqE			MLE			Penalty		
p	$n =$	25	50	100	25	50	100	25	50	100
2		1.50	2.95	1.85	3.79	5.85	4.53	3.76	5.82	4.46
4		2.00	1.49	1.90	5.45	4.21	5.12	5.05	4.03	4.97
8		3.04	1.36	1.85	24.88	4.96	5.74	5.45	3.88	5.54
10		3.94	1.60	2.06	24.11	9.46	6.11	6.08	6.07	5.33

Table 6: Prediction errors for MLqE, MLE and penalized likelihood estimation with ridge penalty, computed using leave-one-out validation.

7 Final Remarks

Acknowledgements

Put acknowledgements here

Appendix A

Calculation of the asymptotic variance of the MLqE

In this section we report the main passages for the calculations for the asymptotic variances discussed in section 4.3.

Univariate Normal with known mean

In this case, the score function is

$$\log f(x;\mu,\sigma) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(x-\mu)^2}{2\sigma^2} \tag{7.1}$$

Consider derivatives taken with respect to μ and σ. The gradient vector and hessian matrix are

$$U(x,\sigma) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma} \tag{7.2}$$

and

$$I(x,\sigma) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \tag{7.3}$$

Next, we compute the integrals

$$K_q = \int_{-\infty}^{\infty} U(x,\sigma\sqrt{q})^2 f(x;\mu,\sigma\sqrt{q})^{2(1-q)} f(x;\mu,\sigma)dx \tag{7.4}$$

$$= \frac{\sigma^{2q-4}(2\pi)^{q-1}q^{q-3/2}(3-2q+q^2)}{(2-q)^{5/2}} \tag{7.5}$$

$$J_q^{(A)} = \int_{-\infty}^{\infty} U(x,\sigma\sqrt{q})^2 f(x;\mu,\sigma\sqrt{q})^{(1-q)} f(x;\mu,\sigma)dx \tag{7.6}$$

$$= -2^{(1+q)/2}\pi^{q-1)/2}q^{q/2-1}\sigma^{q-3} \tag{7.7}$$

and

$$J_q^{(B)} = \int_{-\infty}^{\infty} I(x,\sigma\sqrt{q}) f(x;\mu,\sigma\sqrt{q})^{(1-q)} f(x;\mu,\sigma)dx \tag{7.8}$$

$$= -2^{(1+q)/2}\pi^{(q-1)/2}q^{q/2-1}\sigma^{q-3} \tag{7.9}$$

Thus, we using some simple algebra, we compute the diagonal matrix

$$J_q = (1-q)J_q^{(A)} + J_q^{(B)} = -2^{(1+q)/2}\pi^{(q-1)/2}q^{q/2}\sigma^{q-3}. \tag{7.10}$$

Finally, we have

$$J_q^{-1}K_qJ_q^{-1} = \sigma^2\frac{(3-2q+q^2)}{4(2-q)^{5/2}q^{3/2}}. \tag{7.11}$$

Re-weighting algorithm for multivariate normal

The multivariate normal distribution has the following pdf:

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{7.12}$$

where $|\cdot|$ denotes the matrix determinant. The logarithm of the likelihood evaluated at the i-th observation x_i is

$$\ell_i := \log f(\mathbf{x}_i;\boldsymbol{\mu},\Sigma) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{7.13}$$

Define $\mathbf{z}_i = \boldsymbol{\Gamma}\mathbf{x}_i$, $1 \le i \le n$ and $\boldsymbol{\mu}^* = \boldsymbol{\Gamma}\boldsymbol{\mu}$, where $\boldsymbol{\Gamma}$ is such that $\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{\Lambda} = \text{diag}(\lambda_j)$. The determinant of $\boldsymbol{\Sigma}$ can be computed as the product of the latent roots λ_j, i.e. $|\boldsymbol{\Sigma}| = \prod_{j=1}^p \lambda_j$. Next, note that the last summand in (7.13) is

$$(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}(\mathbf{x}_i-\boldsymbol{\mu}) = (\mathbf{z}_i-\boldsymbol{\mu}^*)'\boldsymbol{\Lambda}^{-1}(\mathbf{z}_i-\boldsymbol{\mu}^*). \tag{7.14}$$

Thus, we can rewrite (7.13) as:

$$\ell_i = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^p \log(\lambda_j) - \sum_{j=1}^p \frac{(z_{ij}-\mu_j^*)^2}{2\lambda_j}, \tag{7.15}$$

where μ_j^* and $z_{ij}$are j-th elements of $\boldsymbol{\mu}^*$ and $\mathbf{z}_i$, respectively. Given a vector of constants, $\mathbf{v}' = (v_1, \ldots, v_n)$ such that $\sum_{i=1}^n v_i = 1$, the estimating equations have the form

$$\sum_{i=1}^n v_i \frac{\partial \ell_i}{\partial \mu_k} = \sum_{i=1}^n v_i \frac{(z_{ik} - \mu_j^*)}{\lambda_k}, \ k = 1, \ldots, p \tag{7.16}$$

and

$$\sum_{i=1}^n v_i \frac{\partial \ell_i}{\partial \lambda_k} = -\sum_{i=1}^n \frac{v_i}{2\lambda_k} + \sum_{i=1}^n v_i \frac{(z_{ik} - \mu_j^*)^2}{2\lambda_k}, \ k = 1, \ldots, p. \tag{7.17}$$

Equating (7.16) and (7.17) to zero gives solutions $\widehat{\mu}_k^* = \sum_{i=1}^n v_i z_{ik}$ and $\widehat{\lambda}_k = n^{-1} \sum_{i=1}^n v_i (z_{ik} - \widehat{\mu}_k^*)^2$. Finally, some straightforward algebra shows that the solutions can be written in terms of the untransformed variable $\mathbf{x}_i$ as

$$\widehat{\boldsymbol{\mu}} = \sum_{i=1}^n v_i \mathbf{x}_i \quad \text{and} \quad \widehat{\Sigma} = \sum_{i=1}^n v_i (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})'(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}). \tag{7.18}$$

References

[1] S. Abe. Geometry of escort distributions. *Phys. Rev. E*, 68(3):031101, Sep 2003.

[2] H. Akaike. Information theory and an extension of the likelihood principle, in: 2nd international symposium of information theory. Mar 1973.

[3] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

[4] Rudolf Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5:445–463, 1977.

[5] Subir K. Bhandari, Ayanendranath Basu, and Sahadeb Sarkar. Robust inference in parametric models using the family of generalized negative exponential disparities. *Australian & New Zealand Journal of Statistics*, 48(1):95–114, 2006.

[6] Subir K. Bhandari, Ayanendranath Basu, and Sahadeb Sarkar. Robust inference in parametric models using the family of generalized negative exponential disparities. *Australian & New Zealand Journal of Statistics*, 48(1):95–114, 2006.

[7] L. D. Brown Brown and G. J. T. Hwang. Robust inference in parametric models using the family of generalized negative exponential disparities. *The American Statistician,*, 47(4):251–255, 1993.

[8] E Choi, P Hall, and B Presnell. Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, 87(2):453–465, 2000.

[9] Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B: Methodological*, 46:440–464, 1984.

[10] I. Csiszár. On topological properties of f-divergences. *Studia Math. Hungar*, 2:329339, 1967.

[11] D. Ferrari and Y. Yang. Maximum Lq-likelihood estimation. *submitted to Annals of Statistics*, 2007.

[12] M. Gell-Mann, editor. *Nonextensive Entropy, Interdisciplinary Applications*. Oxford University Press, New York, 2004.

[13] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 2005.

[14] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Data Sets*. Chapman and Hall, London, 1994.

[15] J. Havrda and F. Charvát. Quantification method of classification processes: Concept of structural entropy. *Kibernetika*, 3:30–35, 1967.

[16] H. Hu and J. V. Zidek. The weighted likelihood. *The Canadian Journal of Statistics*, 30(3):347–371, Sept 2002.

[17] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[18] L. W. Johnson and R. D. Riess. *Numerical Analysis*. Addison-Wesley, Reading, 1982.

[19] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[20] Bruce G. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22:1081–1114, 1994.

[21] J. Naudts. Estimators, escort probabilities, and phi-exponential families in statistical physics. *Journal of Inequalities in Applied and Pure Mathematics*, 2004.

[22] A. Park, C.and Basu. The generalized Kullback-Leibler divergence and robust inference. *Journal of Statistical Computation and Simulation*, 73(5):311–332, 2003.

[23] C. Radhakrishna Rao. Criteria of estimation in large samples. In *Selected papers of C.R. Rao, Vol. 2*, pages 331–352. Wiley Eastern Ltd, 1994.

[24] Robert Tibshirani and Trevor Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567, 1987.

[25] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, July 1988.

[26] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, New York, 1998.

[27] X. Wang and J. V. Zidek. Selecting likelihood weights by cross-validation. *The Annals of Statistics*, 33(2):463–500, 2005.

[28] Michael P. Windham. Robustifying model fitting. *Journal of the Royal Statistical Society, Series B: Methodological*, 57:599–609, 1995.

RECent Working Papers Series

The 10 most RECent releases are:

No. 16 PARAMETRIC DENSITY ESTIMATION BY MINIMIZING NONEXTENSIVE ENTROPY (2008)
D. Ferrari

No. 15 THE INTERACTION BETWEEN PARENTS AND CHILDREN AS A RELEVANT DIMENSION OF CHILD WELL BEING. THE CASE OF ITALY (2008)
T. Addabbo, G. Facchinetti, A. Maccagnan, G. Mastroleo and T. Pirotti

No. 14 FORECASTING FINANCIAL CRISES AND CONTAGION IN ASIA USING DYNAMIC FACTOR ANALYSIS (2008)
A. Cipollini and G. Kapetanios

No. 13 ITALIAN DIASPORA AND FOREIGN DIRECT INVESTMENT: A CLIOMETRIC PERSPECTIVE (2008)
M. Murat, B. Pistoresi and A. Rinaldi

No. 12 INTERNATIONAL MIGRATION AND THE ROLE OF INSTITUTIONS (2008)
G. Bertocchi and C. Strozzi

No. 11 TRADE SANCTIONS AND GREEN TRADE LIBERALIZATION (2008)
A. Naghavi

No. 10 MEASURING BANK CAPITAL REQUIREMENTS THROUGH DYNAMIC FACTOR ANALYSIS (2008)
A. Cipollini and G. Missaglia

No. 9 THE EVOLUTION OF CITIZENSHIP: ECONOMIC AND INSTITUTIONAL DETERMINANTS (2007)
G. Bertocchi and C. Strozzi

No. 8 OPENING THE BLACK BOX: STRUCTURAL FACTOR MODELS WITH LARGE CROSS-SECTIONS (2007)
M. Forni, D. Giannone, M. Lippi and L. Reichlin

No. 7 DYNAMIC FACTOR ANALYSIS OF INDUSTRY SECTOR DEFAULT RATES AND IMPLICATION FOR PORTFOLIO CREDIT RISK MODELLING (2007)
A. Cipollini and G. Missaglia

The full list of available working papers, together with their electronic versions, can be found on the RECent website: www.recent.unimore.it/workingpapers