MAXIMUM LQ-LIKELIHOOD ESTIMATION

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

DAVIDE FERRARI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

May, 2008

Acknowledgments

I would like to thank my thesis advisor, Prof. Yuhong Yang: his inestimable help made possible for me to complete this work. Certainly, I will treasure all I learned under his competent guidance and hope I can bring some of his dedication, fairness and patience into my future professional life.

Thanks must also go to, in no particular order:

Michelle, for always being next to me through this;

my parents, who always encouraged and supported my decisions, although they are still wondering what statistics is about;

Prof. Jones, for helpful conversations and encouragement;

Prof. Jiang and Prof. Cook for helpful suggestions;

Liliana, who shared with me some of her genuine enthusiasm for mathematics and statistics, many meals and coffees and gave me priceless math lectures;

my classmates James, Joo and Aaron, who helped me to survive the first years of the Ph.D. program;

Seongho, for always "opening his door" to me and sharing so many exciting discussions about machine learning, shrinkage and penalization;

Kofi, Brian and Matt for patiently listening to me on so many occasions ... especially when it was about the fact that information is not additive;

Alessandro, for many meals, gym sessions, coffees, discussions on computational chemistry and for sharing some of the "chemistry of statistics";

Andrea, for being a positive presence in my life and for all our discussions, even about statistics;

my dear friends Dondi, Massimo, Carlo and Franco in Casinalbo, just because they exist.

To my family: Zinaida, Giorgio, Nina, Michelle and Andrea.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information: the negative reciprocal value of probability.

C. E. Shannon

When Claude Elwood Shannon wrote his *Mathematical Theory of Communication* (Shannon, 1948), he showed how something seemingly as intangible as information could be described quantitatively. His ideas had important consequences. Perhaps, the impact and universality of the concept of information on scientific thought can be compared only with that of energy. It should not be surprising, however, that the mathematical characterization of information is not unique: the underlying concept is too broad to be fully described by one definition. Nevertheless, it is possible to define measures that agree with intuitive requirements stating how information should be represented.

Usually, information measures have a probabilistic nature. Consider a random variable X with probability distribution p and let $\mathcal{I}$ be function of p. Because of the relationship between X and p, we can use the notation $\mathcal{I}(X)$, instead of $\mathcal{I}(p)$. For two random variables X and Y, a "suitable" information measure should obey the following properties (Gell-Mann and Lloyd, 1996):

(i) $\mathcal{I}(X) \geq 0$ (Nonnegativity)

(ii) $\mathcal{I}(X,Y) = \mathcal{I}(Y,X)$ (Symmetry)

(iii) $\mathcal{I}(X,Y) \geq \mathcal{I}(X)$ (Accumulation)

(iv) $\mathcal{I}(X) + \mathcal{I}(Y) \geq \mathcal{I}(X, Y)$ (Convexity),

where $\mathcal{I}(X, Y)$ denotes the joint information, i.e. the information associated with the joint distribution $p(x, y)$. Any function $\mathcal{I}$ that enjoys the above properties, is called an *information measure*. The interested reader is referred to Ebank et al. (1998) and Arndt (2001) for a detailed treatment on the characterization of information measures.

The choice of the functional form of $\mathcal{I}$ can further specify properties (i-iv) in various ways. For example, Shannon information has the logarithmic form $\mathcal{H}_1(X) = -E_p \log p(X)$. One of the most important features attached to Shannon information concerns the specification of the accumulation property (iii). In particular, the logarithm implies the so-called *strong additivity* property of information, i.e. $\mathcal{H}_1(X, Y) = \mathcal{H}_1(X) + \mathcal{H}_1(Y|X)$, which details how information from the additional variable Y contributes to increase the overall information provided by X and Y. Later, we shall see that $\mathcal{H}_1$ is uniquely defined. In particular, the Shannon-Kinchin Theorem establishes that $\mathcal{H}_1$ is the only measure that satisfies the strong additivity (see Chapter 2).

Not surprisingly, the concept of information – particularly the notion introduced by Shannon – is relevant and universal also in statistics. A large number of inferential techniques depend on it, often implicitly. A highly successful application is the divergence measure between two distributions introduced by Kullback and Leibler. The Kullback-Leibler (KL) divergence, or *relative entropy*, between two distributions $p(x)$ and $p^*(x)$ is defined as

$$E_p \log \frac{p(X)}{p^*(X)} = E_p \log p(X) - E_p \log p^*(X), \tag{1.0.1}$$

where p is often treated as the "true" distribution generating the data (Kullback and Leibler, 1951; Kullback, 1959). The Kullback-Leibler (KL) divergence had provided a cornerstone for statistical estimation, hypothesis testing and later on as a criterion for model selection. For example, when p^* has a parametric form, minimizing the KL divergence between p^* and the empirical measure is equivalent to Maximum Likelihood estimation, i.e. find the values of the parameters that minimize $-\sum_{i=1}^{n} \log p^*(X_i)$, where $X_1, \ldots, X_n$ is an i.i.d. vector of observations

from p^*. Furthermore, Shannon entropy and KL divergence provide the theoretical justification of methods such as empirical likelihood (Owen, 1990, 1991, 2001; Bertail, 2006), Weighted Likelihood (WL) (Hu and Zidek, 2002; Agostinelli, 2002; Wang and Zidek, 2005a,b) and kernel smoothers. Instances of t WL are also seen in prior work in a number of specific contexts (e.g., see Newton and Raftery (1994); Prakasa Rao (1991)) including the local likelihood estimation of Tibshirani and Hastie (1987). Extensions of the local likelihood in the domain of nonparametric regression can be found in Fan and Gijbels (1996).

Often the foundational issues of information theory have been overlooked in statistics. As a result, measures of information other than Shannon entropy (and its relative version, the KL divergence), which are equally reasonable but also convey new and useful statistical properties, do not receive sufficient attention. In the past five decades, a wealth of literature has been devoted to identifying alternatives to Shannon information. Sometimes, these measures were obtained by relaxing some of its underlying assumptions. This is the case for the generalized measure proposed by Havrda and Charvát (1967) and re-discovered later by Tsallis (1988) in the context of statistical mechanics. Havrda-Charvát-Tsallis entropy (or q-entropy) replaces the logarithm in $\mathcal{H}_1$, with a more general function indexed by a constant q. The q-entropy is defined as

$$\mathcal{H}_q(X) = -E_p L_q p(X), \tag{1.0.2}$$

where $L_q(u) = (u^{1-q}-1)/(1-q)$, $u > 0$, $q \neq 1$. This functional form retains all the salient traits of Shannon entropy. It has an additional appealing feature: it allows a more flexible representation of the property of accumulation. It is appropriate to say that $\mathcal{H}_q(X)$ is "almost" additive and the degree of additivity depends on the constant q, which we will refer to as the *distortion parameter*. In Chapter 2, we articulate in more detail the properties of q-entropy in comparison to Shannon entropy.

The underlying spirit of this dissertation is to challenge the prevailing paradigm in statistics that regards information as ubiquitously characterized by strong additivity. We argue that, in many situations, a broader definition of information can

lead to more accurate inferential methodologies. In this work, we start to address these issues by considering the usage of Havrda-Charvát-Tsallis entropy with the goal of statistical inference ultimately in mind. In particular, we are concerned with the estimation setting where data $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$, $p \geq 1$. Within this standard framework, our objectives can be summarized as follows:

- To design a general estimating procedure based on an empirical version of the q-entropy function. Due to the choice of the estimation setting, the procedure results in a new estimator of the parameters.

- To explore the behavior of the new estimator and establish useful statistical properties. In particular, we are interested to gain understanding of how the accuracy of the estimator is affected by the distortion parameter q in relation to the:

  - total amount of information at hand (sample size);
  - presence of information discordant with the assumed model (outliers);
  - complexity of the problem (dimension of the parameter space).

- To gauge the effectiveness of the new estimator in various estimation settings in comparison to well-established methodologies.

Evidence of worthiness of the new methodology is collected from multiple viewpoints: asymptotic analysis, numerical studies and exploration of various real-world datasets.

In what follows, we describe how our study is articulated throughout the chapters. In Chapter 3, we propose the Maximum Lq-Likelihood Estimator (MLqE), a new parameter estimator based on an empirical version of the q-entropy functional $\mathcal{H}_q(X)$. We focus on exponential families and study the statistical properties of the MLqE by means of asymptotic derivations and Monte Carlo simulations. In this context, we aim to gain insight about the relationship between the distortion parameter q and the sample size. In addition to parameter estimation, another

important aspect is investigated: the estimation of a tail probability via the plug-in approach.

Interestingly, proper tuning of the distortion parameter q is connected to an improvement in the accuracy – in terms of mean square error – of the resulting estimator. We remark that the tuning of q can be deterministic, without necessarily requiring an additional estimation of such a parameter (which would imply increased complexity of the model). One of the most important features of the proposed estimator is that when the sample size is small or moderate, the MLqE successfully trades bias for precision, resulting in a substantial reduction of the mean squared error compared to the classic Maximum Likelihood Estimator (MLE). When the sample size is large, a necessary and sufficient condition on q to ensure asymptotic normality and efficiency of MLqE is established. Also, numerical studies on various models show that the MLqE can be beneficial in situations where the number of parameters to be estimated is large compared to the sample size.

The strategy of paying some price in terms of bias for reducing the variance in order to improve the mean square error is common in statistical literature and several methodologies share a similar underlying purpose. This idea dates back to the James-Stein estimator (James and Stein, 1961). Important examples include the local likelihood of Tibshirani and Hastie (1987), nonparametric regression based on kernel smoothers. Hu and Zidek (2002) unify these and other approaches under the umbrella of weighted likelihood estimation theory. In Chapters 3 and 5 we emphasize similarities with these methodologies. In particular, we observe that the MLqE can be regarded as a weighted likelihood estimator, where the weights are set to be proportional to a power tranformation of the pdf of the chosen model the model. Related estimation strategies can be found in Windham (1995), Markatou et al. (1998) Choi et al. (2000) and Park (2003) in the context of robust estimation.

In Chapters 4 and 5, we make a connection between Lq-likelihood estimation and the literature concerning divergence-based methods. Divergence statistics (i.e., those obtained by replacing either one or both arguments of a properly chosen divergence by suitable estimators) have become a good alternative to methodologies based on the likelihood ratio in both continuous and discrete models, as well as to

the classical Pearson-type statistic in discrete models. The importance of and the interest in divergence-based statistics is stressed by Pardo (2006), who provides a systematic treatment of the existing literature on the topic. An overview of some relevant literature on divergence estimators is given in Chapter 4, with some attention to robust estimation. In Chapter 5, we observe that the MLqE minimizes an interesting family of divergences: the so-called power divergences (see Cressie and Read (1984)). The family of power divergences has various notable special cases for specific values of q: (i) KL divergence ($q \to 1$), (ii) Hellinger distance ($q = 1/2$) and (iii) Pearson Chi-Square divergence ($q = 2$). This observation has relevant consequences. For example, when $q = 1/2$ the MLqE is actually a minimum Hellinger distance estimator with the perk of avoiding nonparametric techniques and the difficulties of bandwidth selection.

In Chapter 5, we make two main contributions. First, for fixed q, we extend the theory developed in Chapter 3 and derive asymptotic properties of the MLqE for general families of distributions. In this general context, our asymptotic calculations, simulations and real-world examples show that the MLqE reconciles two apparently contrasting needs: efficiency and robustness, conditional on a proper choice of q. When the sample size is small or moderate, the MLqE trades bias for variance, resulting in an overall improvement in accuracy. At the same time, the MLqE exhibits strong robustness at the expense of a slightly reduced efficiency in the presence of observations discordant with the assumed model.

The second main contribution of Chapter 5 concerns the computational issues of optimization involved in our estimation method. Usually, computing the MLqE is not trivial as it entails maximizing a nonlinear function of the parameters. Consequently, having a reliable *ad-hoc* algorithm is valuable, especially when the dimension of Θ is large. In order to compute the MLq estimates, a fast and easy-to-implement algorithm based on a re-weighting strategy is provided. The algorithm is shown to have approximately linear convergence rate when the empirical distribution of the data is close enough to the model generating the sample. Further, the convergence rate of the algorithm hides some useful information: it is actually an approximate upper bound for efficiency. Therefore, in certain situations, reasonable criteria for the choice of the distortion parameter q can be

constructed based on the convergence rate.

In Chapter 6, we move away from methodological considerations and examine a practical application of the new methodology for estimating *financial risk*. Financial risk (or market risk) represents the hazard that the value of an investment will decrease due to moves in market factors. Clearly, this issue is critical for banks and insurance companies. Recently, quantile estimation based on Extreme Value Theory (EVT) has found a successful domain of application in such a context, outperforming other methods. Although a natural approach for estimating the parametric models provided by the EVT is Maximum Likelihood estimation, the small sample size usually available make the asymptotic properties of MLE untrustworthy. In this context, our goal is to demonstrate that our methodology is an improvement. In particular, our findings show that MLqE outperforms the standard MLE when estimating tail probabilities and quantiles of the Generalized Extreme Value (GEV) and the Generalized Pareto (GP) distributions. The performance of MLqE is assessed using both real-world financial data and Monte Carlo simulations.

Finally, in Chapter 7 we conclude by discussing the main results of our work. A set of open questions and possible future research directions on this new topic are also provided.

Chapters 3, 4 and 6 of this work represent self-standing manuscripts in review for publication or already published. Therefore, occasional repetitions of definitions and concepts may be found without compromising the overall consistency of the work. When appropriate, proofs of theorems or other results can be found in appendices to the chapters.

Chapter 2

Havrda-Charvát-Tsallis entropy

The fact that information can be measured is, by now, generally accepted. How and with what expressions of measure this should be done, is not an open and shut question, however.

A. Rényi

In this chapter, Shannon entropy and some of its generalizations are presented. First, we consider the Shannon-Kinchin axiomatic characterization of information, which naturally leads to the logarithmic form of Shannon entropy. Then, we discuss how relaxing the strong additivity property of Shannon entropy brings a more flexible class of information measures, the Havrda-Charvát-Tsallis entropies, or q-entropies (Havrda and Charvát, 1967; Tsallis, 1988). The objective of this chapter is to summarize the main features of q-entropy in comparison to Shannon information. In particular, we establish the chain rule for q-entropy, give the entropy rate for a sequence of i.i.d. random variables and discuss the property of asymptotic equipartition. Finally, we argue in favor of q-entropy for building new inferential procedures. In particular, q-entropy appears to be suitable for inferential problems involving (i) nontrivial structures of dependence among observations and (ii) small/moderate samples. The second issue will be one of the main objects of this thesis.

2.1 Shannon entropy

The first axiomatic characterization of entropy is due to Shannon (1948). Later Kinchin (1953) made his statement more exact. Let X be a discrete random variable, taking values on the finite support $\Omega = \{x_1, \ldots, x_m\}$ and let $p_i = P(X = x_i)$. Moreover, $p_i \geq 0$, $1 \leq i \leq m$ and $\sum_i p_i = 1$. Let h be a function on $(0, 1]$ and let $h(p_i)$ be a function that expresses the uncertainty removed after observing x_i (or equivalently, the information obtained by revealing that X has taken the value x_i). Further, define

$$\Delta_m = \left\{ p := (p_1, \ldots, p_m) \text{ s.t. } p_i \geq 0, \ \sum_i p_i = 1, i = 1, \ldots, m \right\}. \tag{2.1.1}$$

and consider a function $\mathcal{H} : \Delta_m \mapsto \mathbb{R}$, $(m = 1, 2, \ldots)$

$$\mathcal{H}(p) = \sum_{i=1}^{m} p_i h(p_i). \tag{2.1.2}$$

We also write $\mathcal{H}(X)$ for the above quantity. The function $\mathcal{H}$ is to be interpreted as the *average* amount of uncertainty removed after observing the sample. A functional form for $\mathcal{H}(X)$ is obtained by the axiomatic characterization stated in the following theorem.

Theorem 2.1.1. *Let $\mathcal{H} : \Delta_m \mapsto \mathbb{R}$ be a function satisfying the following properties*

(i) Continuity: The function $\mathcal{H}$ is continuous in its argument;

(ii) Maximality:

$$\mathcal{H}(1/m, \ldots, 1/m) = \max \{\mathcal{H}(p_1, \ldots, p_m) : p_i \in \Delta_m\} > 0; \tag{2.1.3}$$

(iii) Strong additivity: For $p_{ij} \geq 0$, $p_i = \sum_{j=1}^{k_i} p_{ij}$ $(i = 1, \ldots, m;\ j = 1, \ldots, k_i)$;

$$\mathcal{H}(p_{11}, \ldots, p_{mK_m}) = \mathcal{H}(p_1, \ldots, p_n) \tag{2.1.4}$$

$$+ \sum_{i=1}^{m} p_i \mathcal{H}\left(\frac{p_{i1}}{p_i}, \ldots, \frac{p_{im_i}}{p_i}\right); \tag{2.1.5}$$

(iv) Expansibility: $\mathcal{H}(p_1, \ldots, p_m) = \mathcal{H}(p_1, \ldots, p_m, 0)$.

Then,

$$\mathcal{H}(p) = -k \sum_{i=1}^{m} p_i \log(p_i), \tag{2.1.6}$$

where $k > 0$, $b > 1$, *with* $0 \log 0 = 0$.

The theorem is often referred to as the *Shannon-Kinchin characterization* of the Shannon entropy. A proof of this version of the Uniqueness Theorem is given by Aczél and Daróczy (1975), p.67. Other versions can be found in Shannon (1948) and Aczél and Daróczy (1975).

The result is important as it states that Shannon measure based on the logarithmic function is uniquely determined by some rather natural postulates. Continuity states that for any m, $\mathcal{H}$ should be a continuous and symmetric function of p. This can be phrased as saying that nearby probabilities give nearby uncertainties. Another important property emerging from the postulates (ii) and (iv) is *monotonicity*. Namely, if X is uniformly distributed, then $\mathcal{H}$ should be a nondecreasing function of m, i.e., the knowledge of more outcomes corresponds to smaller uncertainty.

The following is a direct consequence of the postulates. Given two variables X and Y, strong additivity implies that the joint information about X and Y can be split into two parts: information about X alone and information about Y given the information already known about X:

$$\mathcal{H}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y|X). \tag{2.1.7}$$

Although there is more than one version of this theorem based on slightly different articulation of the axioms, the underlying idea of additivity is always present in some sense. However, the requirement of additive accumulation of information is the most severe as it shapes completely the functional form of (2.1.6).

Finally, we remark that the convention $0 \log 0 = 0$ is easily justified by continuity since $u \log u \to 0$ as $u \to 0$.

Example 2.1.1 (Bernoulli r.v.)**.** Let X be defined on $\Omega = \{0, 1\}$ and let $\theta = P(X = 1)$. Then, $\mathcal{H}(X) = -\theta \log(\theta) - (1 - \theta) \log(1 - \theta)$. Fig. 2.1(a) (solid curve) shows that $\mathcal{H}(X)$ is a concave function of the distribution of X and takes value 0 if $\theta = 0$ or 1, which is reasonable because X becomes deterministic and there is no uncertainty involved.

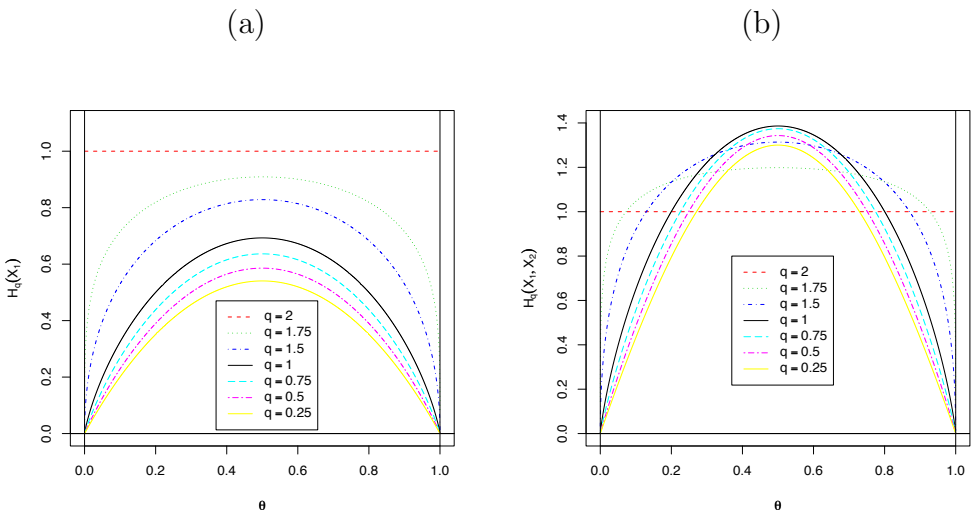


Figure 2.1: (a) $\mathcal{H}_q(X)$ versus θ for a Bernoulli random variable. (b) Joint entropy $\mathcal{H}_q(X_1, X_2)$ versus θ for two independent Bernoulli random variables. The solid curve ($q = 1$) corresponds to Shannon entropy.

Although Shannon entropy is defined for a discrete random variable, it can be easily extended to the continuous case (see Cover and Thomas (2006)). Entropy for continuous random variables is sometimes called *differential entropy*. Let X be a continuous random variable defined on Ω, with probability density function

$p(x)$. The entropy in this case is defined as

$$\mathcal{H}(X) = -\int_{\Omega} p(x) \log p(x) dx, \tag{2.1.8}$$

whenever the integral exists. As in the discrete case, the differential entropy depends only on the probability density function of the random variable, and hence the differential entropy is sometimes written as $\mathcal{H}(p)$ rather than $\mathcal{H}(X)$. Note that unlike the entropy of a discrete random variable, the above quantity may be infinitely large, negative or positive. Furthermore, although discrete entropy is invariant under a one-to-one change of variable, this is not necessarily true in the continuous case.

Example 2.1.2 (Uniform Distribution)**.** Consider a random variable X distributed uniformly on the real interval $[a, b]$, so that its density is $1/(b-a)$ from a to b and 0 elsewhere. Then, its entropy is

$$\mathcal{H}(X) = \int_a^b (b-a)^{-1} \log(b-a)^{-1} dx = -\log(b-a). \tag{2.1.9}$$

Note that if $b - a < 1$, the entropy is negative.

Example 2.1.3 (Multivariate normal distribution)**.** Let $(X_1, \ldots, X_p)^{\intercal} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The multivariate normal distribution has pdf

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \tag{2.1.10}$$

Then,

$$\mathcal{H}(\mathbf{X}) = -\int f(\mathbf{x};; \boldsymbol{\mu}, \boldsymbol{\Sigma}) [-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\intercal} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{2.1.11}$$

$$- \frac{n}{2} \log (2\pi |\boldsymbol{\Sigma}|)] d\mathbf{x}, \tag{2.1.12}$$

where $|\cdot|$ denotes the matrix determinant. A straightforward calculation (see Cover

and Thomas (2006), p.230) shows that

$$\mathcal{H}(\mathbf{X}) = \frac{1}{2}\log(2\pi e)^n|\mathbf{\Sigma}|. \tag{2.1.13}$$

Since Shannon entropy was introduced, other and more general measures of information have been proposed and a wealth of mathematical literature has been devoted to relax the postulates of Shannon entropy. The information measure proposed by Rényi (1961) will be briefly discussed in the next section and in the remainder of this chapter we will examine in more detail the entropy first proposed by Havrda and Charvát (1967).

2.2 Rényi entropy

Any argument that Shannon entropy is the only possible measure of information is valid only within the restricted scope of the coding problems considered by Shannon. In a famous paper, Rényi (1961) pointed out that more general information measures in other types of problems may serve just as well, or even better, than Shannon entropy. He argued that the choice of one measure over another should be supported either by their operational significance or by a set of natural postulates characterizing them, and preferably by both. Rényi proposed a scalar parametric entropy which includes Shannon entropy as a limiting case.

Rényi generalized entropy, sometimes called entropy of order r, keeps the main scheme of the postulates of Shannon entropy, but introduces a more general definition of mean. The discrete version of Rényi entropy has the form

$$\mathcal{H}_r(p) = \frac{1}{1-r}\log\sum_{i=1}^{n} p_i^r,\ r \neq 1,\ r > 0, \tag{2.2.1}$$

and its continuous counterpart can be written as

$$\mathcal{H}_r(p) = \frac{1}{1-r}\log\int_{\Omega} p(x)^r dx,\ r \neq 1,\ r > 0. \tag{2.2.2}$$

One can easily see that $\lim_{r\to 1}\mathcal{H}_r = \mathcal{H}$, where $\mathcal{H}$ is Shannon entropy.

Although the definition of mean is modified, Rényi entropy enjoys the following additivity property. Let X and Y be two random variables. The joint information of two independent variables X and Y is

$$\begin{aligned}\mathcal{H}_r(X,Y) &= \frac{1}{1-r}\log\int\int p(x,y)^r dydx \\ &= \frac{1}{1-r}\log\int p(x)^r dx + \frac{1}{1-r}\log\int p(y)^r dy \\ &= \mathcal{H}_r(X) + \mathcal{H}_r(Y), \end{aligned} \tag{2.2.3}$$

i.e., the overall information given by X and Y is understood as the sum of the information of the individual random variables. Note that if X and Y are not independent, by the probability chain rule, we can factor the joint distribution as $p(x,y) = p(y|x)p(x)$, but still cannot separate $\mathcal{H}_r(X,Y)$ into the contribution given by $\mathcal{H}_r(X)$ and $\mathcal{H}_r(Y|X)$, unless $r = 1$.

2.3 Havrda-Charvát-Tsallis entropy

In a different direction, Havrda and Charvát (1967) proposed a generalized entropy measure that keeps the usual definition of mean, but relaxes the postulate of strong additivity (iii) in Theorem 2.1.1. Havrda and Charvát entropy, sometimes referred to as entropy of degree s, is

$$\mathcal{H}_s(p) = (2^{1-s}-1)^{-1}\left[\sum_{i=1}^{n} p_i^s - 1\right]. \tag{2.3.1}$$

More recently, a slightly modified version of the above function has been of increasing interest in various scientific domains. Tsallis and colleagues (Tsallis, 1988; Tsallis et al., 1998) have successfully exploited such an information measure in physics in relation to non-equilibrium phenomena. The generalization is considered to be one of the most viable candidates for generalizing Boltzmann-Gibbs thermodynamics theory. Since Tsallis' seminal paper (Tsallis (1988)), a growing number of applications have appeared in various disciplines such as finance, biomedical sciences, environmental sciences and linguistics (*e.g.*, see Gell-Mann

(2004)). The q-entropy, or Havrda-Charvát-Tsallis entropy is defined as

$$\mathcal{H}_q(p) = (1-q)^{-1} \sum_{i=1}^{m} p_i[p_i^{-q} - 1]. \tag{2.3.2}$$

One way to characterize Havrda-Charvát-Tsallis entropy for discrete random variables is provided by the generalized Shannon-Kinchin conditions (Suyari, 2004).

Theorem 2.3.1. *Let $\mathcal{H}_q : \Delta_m \mapsto \mathbb{R}$ satisfy*

(i) Continuity: The function $\mathcal{H}_q$ is continuous in its arguments;

(ii) Maximality:

$$\mathcal{H}_q(m^{-1}, \ldots, m^{-1}) = \max\{\mathcal{H}_q(p_1, \ldots, p_m) : p_i \in \Delta_m\} > 0; \tag{2.3.3}$$

(iii) Generalized Shannon additivity: For $p_{ij} \geq 0$, $p_i = \sum_{j=1}^{k_i} p_{ij}$ ($i = 1, \ldots, m$; $j = 1, \ldots, k_i$);

$$\mathcal{H}_q(p_{11}, \ldots, p_{mk_m}) = \mathcal{H}_q(p_1, \ldots, p_m) \tag{2.3.4}$$

$$+ \sum_{i=1}^{m} p_i^q \mathcal{H}_q\left(\frac{p_{i1}}{p_i}, \ldots, \frac{p_{ik_i}}{p_i}\right); \tag{2.3.5}$$

(iv) Expandability: $\mathcal{H}_q(p_1, \ldots, p_m, 0) = \mathcal{H}_q(p_1, \ldots, p_m)$.

Then $\mathcal{H}_q$ has the form

$$\mathcal{H}_q = \frac{\sum_{i=1}^{n} p_i(p_i^{1-q} - 1)}{\phi(q)}, \tag{2.3.6}$$

where (a) $\phi(q)$ is continuous and such that $\phi(q)(q-1) > 0$ (for $q \neq 1$), (b) $\lim_{q\to 1} \phi(q) = 0$ and $\phi(q) \neq 0$ for $q \neq 1$, (c) there exists $(a, b) \in \mathbb{R}^+$ such that $a < 1 < b$ and ϕ is differentiable on $(a, 1)$ and $(1, b)$ and (d) there exists a positive constant k such that $\lim_{q\to 1} \partial\phi(q)/\partial q = 1/k$.

Proofs of the Generalized Uniqueness Theorem are given by Suyari (2004) and

Furuichi (2005). In particular, Furuichi (2005) extends the result to the case of Havrda-Charvát-Tsallis relative entropy.

Note that in condition (ii), the strict positivity of the maximum of the entropy functional is imposed. Examples of $\phi(q)$ satisfying the requirements stated above are: (1) $\phi(q) = 1 - 2^{1-q}$, which appeared in the formulation proposed by Havrda and Charvát (1967) and (2) $\phi(q) = 1 - q$, which was considered by Tsallis (1988). Analogously to Shannon's entropy, Tsallis' measure can be extended to continuous random variables as

$$\mathcal{H}_q(p) = (1-q)^{-1} \int_\Omega p(x)[p(x)^{1-q} - 1]dx. \tag{2.3.7}$$

The connection of the q-entropy with Shannon information can be seen more transparently by introducing some additional notation in the following definition.

Definition 2.3.1. *Let p be the pdf (or pmf) of the random variable X. The q-entropy is defined as*

$$\mathcal{H}_q(X) = -E_p L_q p(X), \tag{2.3.8}$$

where

$$L_q(u) = \begin{cases} \dfrac{u^{1-q} - 1}{1-q} & \text{if } \ q \neq 1, \\ \log u & \text{if } \ q = 1. \end{cases} \tag{2.3.9}$$

The function L_q represents a Box-Cox transformation in statistics and in other contexts it is often called a deformed logarithm. Note that as $q \to 1$, then $L_q(u)$ converges uniformly to $\log(u)$ and the usual definition of Shannon's entropy is recovered. Since the q-logarithm L_q has the following *pseudo-additivity* property:

$$L_q(u_1 u_2) = L_q(u_1) + L_q(u_2) + (1-q)L_q(u_1)L_q(u_2), \tag{2.3.10}$$

the joint information of two random variables X and Y is

$$\mathcal{H}_q(X,Y) = \mathcal{H}_q(X) + \mathcal{H}_q(Y|X) + (1-q)\mathcal{H}_q(X)\mathcal{H}_q(Y|X). \tag{2.3.11}$$

The above display shows that under generalized additivity, the information given

by considering simultaneously X and Y can be larger (or smaller) than the information provided by the variables alone.

Finally, note that there is a simple transformation that relates Tsallis entropy to Rényi entropy. Let $\mathcal{H}^*$ denote Rényi's entropy and $\mathcal{H}_q$ denote the q-entropy. Then,

$$\mathcal{H}_{2-r}(X) = L_r \left\{ \exp \left\{ \mathcal{H}_r^*(X) \right\} \right\}. \tag{2.3.12}$$

Despite this connection, the two information measures differ with respect to their additivity structure, as emphasized by Eq. (2.2.3) and Eq. (2.3.11).

Example 2.3.1 (Bernoulli r.v.)**.** Consider two independent random variables X_1, X_2 on $\Omega = \{0, 1\}$, where $\theta = P(X_k = 1)$, $k = 1, 2$. Then,

$$\mathcal{H}_q(X_1) = -(1-q)^{-1}[\theta^{2-q} + (1-\theta)^{2-q} - 2]. \tag{2.3.13}$$

Moreover, a straightforward calculation shows that the joint information of X_1 and X_2 is

$$\begin{aligned} \mathcal{H}_q(X_1, X_2) = & -2(1-q)^{-1}[\theta^{2-q} + (1-\theta)^{2-q} - 2] \\ & - (1-q)^{-1}[\theta^{2-q} + (1-\theta)^{2-q} - 2]^2. \end{aligned} \tag{2.3.14}$$

Fig. 2.1(a) and Fig. 2.1(b) show that $\mathcal{H}_q(X)$ is a concave function of θ and the shape of the concavity changes with parameter q. Interestingly, when $q > 1$, $\mathcal{H}_q$ flattens, meaning the uncertainty tends to be more homogeneous for values of θ around $1/2$. When θ is close to 0 or 1, $\mathcal{H}_q$ decreases rapidly to 0.

Example 2.3.2. (Multivariate normal distribution). Let $\mathbf{X} = (X_1, \ldots, X_p)$ have

a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.

$$\mathcal{H}_q(\mathbf{X}) = -\int f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left\{L_q f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right\} d\mathbf{x} \tag{2.3.15}$$

$$= -\frac{1}{1-q}\int f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\left[\left(\frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\right)^{1-q} - 1\right] d\mathbf{x} \tag{2.3.16}$$

$$= -\frac{1}{1-q}\left[\left(\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\right)^{2-q}\int e^{-\frac{(2-q)}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}d\mathbf{x} - 1\right] \tag{2.3.17}$$

$$= -\frac{1}{1-q}\left[\frac{(2\pi)^{p/2}(2-q)^{-p/2}|\boldsymbol{\Sigma}|^{1/2}}{\left((2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\right)^{2-q}} - 1\right] \tag{2.3.18}$$

$$= -\frac{\left[(2-q)^{p(1-q)/2}(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\right]^{-(1-q)} - 1}{1-q} \tag{2.3.19}$$

$$= -L_q\left\{(2-q)^{-p/2(1-q)}(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\right\}. \tag{2.3.20}$$

Moreover, one can check that $L_q\left[(2-q)^{-p/2(1-q)}\right] \to -p/2$ and $L_q(\cdot) \to \log(\cdot)$ as, $q \to 1$. Thus, an application of property (2.3.10) to (2.3.20) shows $\mathcal{H}_q(\mathbf{X}) \to \frac{1}{2}\log(2\pi e)^p|\boldsymbol{\Sigma}|$, as $q \to 1$, which is exactly Shannon entropy for a multivariate normal in Eq. (2.1.13).

2.3.1 Relation between differential and discrete entropy

Consider a continuous random variable X with density $p(x)$. Divide the range of X into bins of length δ. By the Mean Value Theorem, there exists a value x_i within each bin such that

$$p(x_i)\delta = \int_{i\delta}^{(i+1)\delta} p(x)dx. \tag{2.3.21}$$

Thus, we define the quantized random variable X_δ as

$$X_\delta := x_i, \quad \text{if} \quad i\delta \le X < (i+1)\delta. \tag{2.3.22}$$

Theorem 2.3.2. *If $p(x)L_q(p(x))$ is integrable in the Riemann sense, then*

$$\delta^{q-1}\mathcal{H}_q(X_\delta) + \delta^{q-1}L_q(\delta) \to -\int p(x)L_q(p(x))dx, \quad as \;\; \delta \to 0. \tag{2.3.23}$$

Proof. Note that

$$P(X_\delta = x_i) = \int_{i\delta}^{(i+1)\delta} p(x)dx = p(x_i)\delta. \tag{2.3.24}$$

Using the pseudo-additivity property (2.3.10), the q-entropy of the quantized random variable can be expressed as

$$\begin{aligned}
\mathcal{H}_q(X_\delta) &= -\sum p_i L_q(p_i) && (2.3.25)\\
&= -\sum p(x_i)\delta L_q(p(x_i)\delta) && (2.3.26)\\
&= -\sum p(x_i)\delta L_q(p(x_i)) - \sum p(x_i)\delta L_q(\delta) && (2.3.27)\\
&\quad - (1-q)L_q(\delta)\sum p(x_i)\delta L_q(p(x_i)).
\end{aligned}$$

Note that the second summand in the last equality is $L_q(\delta)$ since $\sum p(x_i)\delta = \int p(x)dx = 1$. Moreover, if $p(x)L_q(p(x))$ is Riemann integrable, we have

$$\sum p(x_i)\delta L_q(p(x_i)) \to \int p(x)L_q(p(x))dx, \tag{2.3.28}$$

as $\delta \to 0$. This completes the proof. □

Remarks

1. The result in Theorem 2.3.2 highlights the connection between discrete and continuous entropy employing an elementary discretization argument. If $L_q(\cdot)$ is replaced by the usual logarithm, then the statement of the theorem reads:

$$\mathcal{H}_1(X_\delta) + \log\delta \to \mathcal{H}_1(X), \quad \text{as} \;\; \delta \to 0. \tag{2.3.29}$$

   Thus, Shannon's entropy of the quantized random variable is approximately

$\mathcal{H}_1(X) + \log \delta$.

2. Let $q := q_\delta$ be a sequence depending on the bin length δ such that $(1 - q_\delta) \log \delta \to 0$, as $\delta \to 0$. An inspection of the proof shows that

$$\mathcal{H}_q(X_\delta) + L_q(\delta) \to \mathcal{H}_1(X), \quad \text{as} \quad \delta \to 0. \tag{2.3.30}$$

Thus, for a properly chosen sequence q_δ, a δ-quantization of the q-entropy of a continuous random variable X is approximately $\mathcal{H}_1(X) - L_q(\delta)$ and for a properly chosen sequence q_δ it can potentially improve the accuracy given by $\mathcal{H}_1(X_\delta)$. In particular, when δ is close to 0, we can find $q_\delta < 1$ such that $|\log(\delta)| > |L_q(\delta)|$. A graphical representation is given in in Fig. 2.2, where $L_q(\delta)$ is plotted for various values of q.

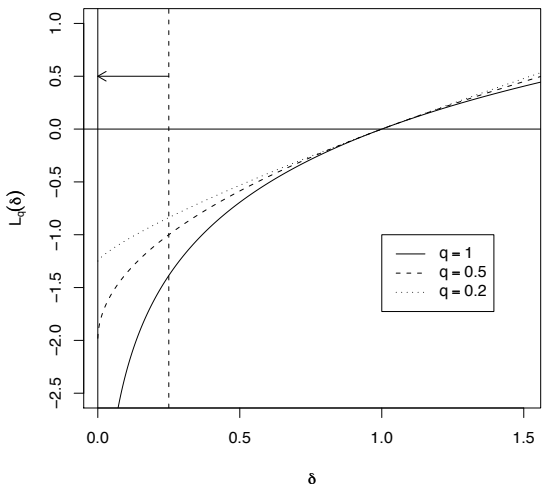


Figure 2.2: Distorted logarithm $L_q(\delta)$ for various values of q. The solid curve $(q = 1)$ represents $\log(\delta)$.

Example 2.3.3. Let X be a uniform random variable on $[0, 1]$ with density $p(x) =$

$I(0 \leq x \leq 1)$ and let $\delta = 1/n$. Moreover, $p(x_i)\delta = \int_{i\delta}^{(i+1)\delta} p(x)dx = 1/n$ and

$$\mathcal{H}_1(X_\delta) = -\sum_{i=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) = -\log(1/n), \tag{2.3.31}$$

$$\mathcal{H}_1(X) = \int p(x) \log p(x) dx = \int_0^1 \log(1) dx = 0. \tag{2.3.32}$$

Following (2.3.29), we have that Shannon's entropy of the discretized random variable X_δ is approximately the entropy of X up to $\log \delta = n$. On the other hand,

$$\mathcal{H}_q(X_\delta) = -\sum_{i=1}^{n} \frac{1}{n} L_q\left(\frac{1}{n}\right) = -L_q(1/n). \tag{2.3.33}$$

Next, note that $-L_q(1/n) < -\log(1/n)$, when $q < 1$. In particular, in order to improve the approximation given by $\mathcal{H}_1(X_\delta)$, it suffices to chose a sequence $q_n < 1$ such that $1 - q_n = o(\log(n)^{-1})$. Figure 2.3 shows the difference $\mathcal{H}_1(X_{1/n}) - \mathcal{H}_{q_n}(X_{1/n})$ versus n for various sequences q_n. The plots exhibit a logarithmically increasing behavior.

2.4 Domains of application

In the last half century, Shannon entropy has found countless applications in many applied and pure sciences (see Cover and Thomas (2006)). As for the case of Shannon entropy, the importance q-entropy is proved by a growing number of applications in several domains. In this section, we summarize a few of them. The interested reader is also referred to Abe (2001) for a detailed treatment of the applications in physics. Gell-Mann (2004) illustrates works in physics as well as in computer science, bio-medical sciences and social sciences. We remark, however, that most of this literature is concerned with the extension of the theory in statistical mechanics introduced by Tsallis (1988).

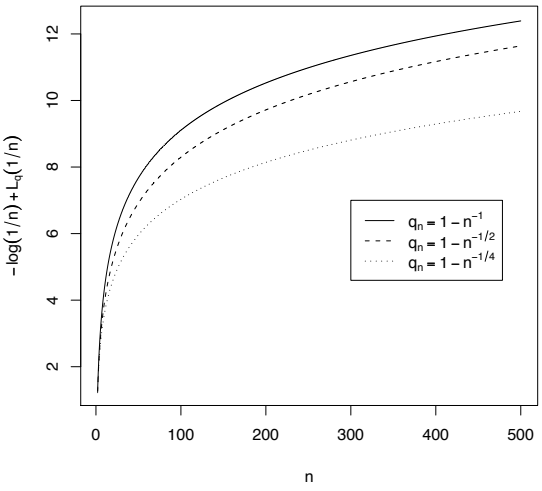


Figure 2.3: Difference between $\mathcal{H}_1(X_\delta)$ and $\mathcal{H}_q(X_\delta)$ for $X \sim U[0,1]$, $\delta = 1/n$ and various choices of the sequence of distortion parameters q_n.

Statistics. In statistics and statistical physics, q-entropy has been used as an alternative to Shannon information when applying Jaynes' *principle of maximum entropy* (MaxEnt) (see Borland et al. (1998); Plastino and Plastino (1999)). The principle of maximum entropy is a method introduced by Jaynes (1957a,b) for analyzing available qualitative information in order to determine a unique probability distribution. It states that the least biased distribution that encodes certain given information is that which maximizes the information entropy. Jaynes introduced maximum entropy thermodynamics by re-interpreting the Gibbs algorithm of statistical mechanics. He suggested that thermodynamics, and in particular thermodynamic entropy, should be seen just as a particular application of a general tool of inference and information theory. The MaxEnt principle shares some features with Bayesian methods as it makes explicit use of prior information. Under Jaynes' inferential paradigm, a considerable amount of work has been done to show that various loss functions can be seen as the convex dual of entropy minimization, subject to constraints. From this standpoint, the classical maximum entropy estimation and maximum likelihood are seen as convex duals of each other (see Altun and Smola (2006)).

Physics. A standard assumption of statistical mechanics is that quantities like energy are "extensive" variables, meaning that the total energy of the system is proportional to the system size; similarly the entropy is also supposed to be extensive. This is justified by appealing to the short-range nature of the interactions which hold matter together, form chemical bonds, etc. Tsallis found that in certain long-range situations energy is not extensive. He provided a generalization of the classic Boltzmann-Gibbs theory using the q-entropy function in combination with Jaynes' formalism. His theory proved useful for a wide class of "anomalous" systems in physics including non-ergotic systems with stationary long-lived states. Other applications concern the physics of turbulence.

Computer sciences. Applications in this domain concern the problem of finding global optima of complex functions with possible multiple local minima (maxima). One of the algorithms proposed for solving this problem is the simulated annealing (SA), which uses entropy at two different steps. Lately, a new version of SA, called generalized simulated annealing (GSA) (Andricioaei and Straub, 1996) involving the q-entropy has been proposed resulting in an increased speed, precision and success rate. Other applications in computer sciences exhibiting common features with statistical mechanics are about the statics and dynamics of internet networks.

Economics. Theory of risk aversion in economics has been developed based on biased averages. The Black-Scholes differential equation for pricing options and its solutions has been generalized with the help of q-entropy, thus obtaining a significant agreement with market data (Anteneodo et al., 2002). Further, the presence of certain "power-law" distributions in economic phenomena can be often explained using ideas related to Tsallis theory (Rajagopal and Abe, 2002; Duarte-Queiros et al., 2005).

Medicine. In this domain entropy, is used especially when processing electroencephalographic (EEG) and electrocardiographic (ECG) signals. In many circumstances it has been found that q-entropy can be used to improve the quality of the signals.

Lately, more insight has been provided on the meaning of the distortion parameter q from the viewpoint of Tsallis' theory. In general, however, the role of q appears to be unclear from a statistical perspective. Although it is common to find applications dealing with direct fitting of q in various contexts, to our knowledge no work has been done to devise the properties of q in relation to statistical inference.

2.5 Properties

The most remarkable feature of q-entropy derives from the nonadditivity of L_q, stated in Eq. (2.3.10). The purpose of the present section is to summarize some known basic properties of the function L_q as well as to investigate extensions of the pseudo-additivity property when considering a set of random variables $\mathcal{V} = (X_1, \ldots, X_n)$.

Difference with respect to log. One useful quantity is the characterization of the difference measured from the usual natural logarithm. Consider the differentiating $L_q(x)$ with respect to q

$$\frac{\partial L_q(x)}{\partial q} = -\frac{x^{1-q}\log x}{1-q} + \frac{x^{1-q}-1}{(1-q)^2} = \frac{L_q(x) - x^{1-q}\log x}{1-q}. \tag{2.5.1}$$

Re-writing the above expression gives,

$$\log(x) - L_q(x) = (1-q)\left[\frac{\partial L_q(x)}{\partial q} + \log(x) L_q(x)\right]. \tag{2.5.2}$$

Upper and lower bounds. For $x > 0$, we have the following inequalities

$$\begin{array}{ll} L_q(x) \geq \dfrac{x-1}{x^q}, & q > 0 \\ L_q(x) \leq \dfrac{x-1}{x^q}, & q < 0 \end{array}. \tag{2.5.3}$$

Taylor-series expansion. If $|x| \leq 1$, computing the Taylor-series expansion about 0 gives

$$L_q(x) = \sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{m!} \lambda(q)_{m-1} x^m, \tag{2.5.4}$$

where $\lambda(q)_{m-1}$ is the shifted factorial $\lambda(q)_k = \Gamma(q+k)/\Gamma(q) = q(q+1)\cdots(q+k-1)$, for $k = 0, 1, 2, \ldots$.

2.5.1 Pseudo-additivity and chain rule

In the following theorem, a formula to compute the q-logarithm of the product of n positive variables is established. The result is extremely useful as it allows us to compute: (i) the mean and the variance of the q-logarithm of the product of random variables and (ii) the q-entropy for any sequence of random variables. An immediate consequence of the Theorem is to establish the chain rule for q-entropy.

Theorem 2.5.1. *Consider a set of positive random variables $\mathcal{V} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ and let $\mathcal{T}_i$ be the set of all combinations of size i of the elements of $\mathcal{V}$. Then,*

$$L_q\left(\prod_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} (1-q)^{i-1} \sum_{T \subseteq \mathcal{T}_i} \prod_{X_k \in T} L_q(X_k). \tag{2.5.5}$$

Proof. The proof of the Theorem is given in Section 2.8. □

Note that for the joint distribution, we have the chain rule

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}, \ldots, x_1) \tag{2.5.6}$$

and a direct application of Theorem 2.5.1 to the right hand side of (2.5.6) gives the following formula for computing the entropy of a collection of random variables.

Corollary 2.5.2 (Chain rule for q-entropy)**.** *Let $X_1, \ldots, X_n$ be drawn according*

to $p(x_1, \ldots, x_n)$. *Then,*

$$\mathcal{H}_q(X_1, \ldots, X_n) = \sum_{i=1}^{n} (1-q)^{i-1} \sum_{T \subseteq \mathcal{T}_i} \prod_{X_k \in T} \mathcal{H}_q(X_k | X_{k-1}, \ldots, X_1). \tag{2.5.7}$$

From (2.5.2), the total information given by n observations from $p(x)$ is

$$\mathcal{H}_q(X_1, \ldots, X_n) = \sum_{i=1}^{n} \mathcal{H}_q(X_i | X_{i-1}, \ldots, X_1) + E\rho_q(X_1, \ldots, X_n), \tag{2.5.8}$$

where $\rho_q(X_1, \ldots, X_n)$ accounts for all interaction terms of order 2 up to n. Thus, the total information of the variable set is interpreted as the sum of the information provided by the individual variables augmented (decreased) by an interaction term depending on q. We refer to the term ρ_q as the *redundancy* or redundant information.

For instance, consider the triplet of random variables (X_1, X_2, X_3). Since $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$, we can write

$$\begin{aligned} E\rho_q(X_1, X_2, X_3) = (1-q)[&\mathcal{H}_q(X_1)\mathcal{H}_q(X_2|X_1) \\ &+ \mathcal{H}_q(X_1)\mathcal{H}_q(X_3|X_1, X_2) + \mathcal{H}_q(X_2|X_1)\mathcal{H}_q(X_3|X_1, X_2)] \\ &- (1-q)^2[\mathcal{H}_q(X_1)\mathcal{H}_q(X_2|X_1)\mathcal{H}_1(X_3|X_2, X_1)]. \end{aligned} \tag{2.5.9}$$

In general, the expected redundancy of the variable set $\mathcal{V} = \{X_1, X_2, \ldots, X_n\}$ can be expressed as

$$\begin{aligned} &E\rho_q(X_1, \ldots, X_n) \\ =&E \sum_{i=2}^{n} (1-q)^{i-1} \sum_{T \subseteq \mathcal{T}_i} (-1)^{\mathrm{card}(\mathcal{V}) - \mathrm{card}(T)} \prod_{X_k \in T} L_q p(X_k | X_{k-1}, \ldots X_1) \\ =&\sum_{i=2}^{n} \sum_{T \subseteq \mathcal{T}_i} (-1)^{\mathrm{card}(\mathcal{V}) - \mathrm{card}(T)} (1-q)^{i-1} \prod_{X_k \in T} \mathcal{H}_q(X_k | X_{k-1}, \ldots X_1), \end{aligned} \tag{2.5.10}$$

where $\mathrm{card}(\cdot)$ denotes the cardinality of a set and $\mathrm{card}(\mathcal{V}) = n$. Clearly, (i) the strength of higher order effects decreases quickly, according to $(1-q)^{i-1}$; (ii) the

further q is from 1, the stronger the contribution of the interaction terms, with lack of redundancy for $q = 1$.

Interestingly, the redundancy of extensive entropy in (2.5.10) has similarities to *interaction information* (McGill, 1954) or *co-information* (Bell, 2003). Interaction information is one of several generalizations of the mutual information, and expresses the synergy bound up in a set of variables, which is present in any subset of those variables. A general expression for interaction information on variable set $\mathcal{V}$ is given by Jakulin and Bratko (2003) as

$$\mathcal{I}(\mathcal{V}) = -\sum_{\mathcal{T}\subseteq\mathcal{V}} (-1)^{\mathrm{card}(\mathcal{V})-\mathrm{card}(T)} \mathcal{H}_1(\mathcal{T}) \tag{2.5.11}$$

$$= -\sum_{i=1}^{n} \sum_{T\subseteq\mathcal{T}_i} (-1)^{\mathrm{card}(\mathcal{V})-\mathrm{card}(T)} \mathcal{H}_1(\mathcal{T}_i) \tag{2.5.12}$$

which is an inclusion-exclusion sum over all subsets $\mathcal{T} \subseteq \mathcal{V}$. Unlike the mutual information, the interaction information can be either positive or negative. This property can be misleading, which has slowed its wider adoption as an information measure in statistics and machine learning. Recently, Jakulin and Bratko (2004) provide a machine learning algorithm which uses the idea of interaction information for describing general correlation patterns in the data.

2.5.2 Entropy rate for i.i.d. random variables

A natural question to ask is: "How does the q-entropy of $(X_1, \ldots, X_n)$ grow with n ?". The rate of growth of information, or *entropy rate* of a sequence of random variables $X_1, \ldots, X_n$ is defined as

$$\mathscr{R}_q(\mathcal{V}) = \lim_{n\to\infty} \frac{1}{n} \mathcal{H}_q(X_1, \ldots, X_n). \tag{2.5.13}$$

If $q = 1$ and the X_is are i.i.d., the derivation of the rate is trivial due to additivity of the logarithm. When $q \neq 1$, it turns out that it is possible to derive a simple formula for $\mathscr{R}_q(\mathcal{V})$. The next theorem gives the mean and the variance of the q-logarithm of the product of positive random variables. The result will be used to

derive entropy rates for an i.i.d. sequence.

Theorem 2.5.3. *Let $X_1, ..., X_n$ be positive i.i.d. random variables. Then,*

(i)

$$EL_q\left(\prod_{i=1}^{n} X_i\right) = \frac{[(1-q)EL_q(X_1)+1]^n - 1}{1-q} \tag{2.5.14}$$

and

$$VarL_q\left(\prod_{i=1}^{n} X_i\right) = \frac{[(1-q)^2 VarL_q(X_1)+1]^n - 1}{(1-q)^2}. \tag{2.5.15}$$

(ii) Moreover, if $\{q_n\}_{n\geq 1}$ is a sequence such that $(1-q_n)n \to 0$ and $E\left[L_{q_n}(X)\right]^k \to E\left[\log(X)\right]^k$ $(k=1,2)$ as $n \to \infty$, then

$$n^{-1}EL_q\left(\prod_{i=1}^{n} X_i\right) \overset{n\to\infty}{\to} E\log(X_1) \tag{2.5.16}$$

and

$$n^{-1}VarL_q\left(\prod_{i=1}^{n} X_i\right) \overset{n\to\infty}{\to} Var\log(X_1). \tag{2.5.17}$$

Proof. The proof of the Theorem is given in Section 2.8. □

From part (i) of the Theorem, one can immediately obtain an expression for the joint entropy of n i.i.d. random variables, by replacing X_1 with $p(X_1)$ in (2.5.14):

$$\mathscr{R}_q(\mathcal{V}) = \lim_{n\to\infty} \frac{1}{n} EL_q\left(\prod_{i=1}^{n} p(X_i)\right) \tag{2.5.18}$$

$$= \lim_{n\to\infty} \frac{1}{n} \frac{[(1-q)\mathcal{H}_q(X_1)+1]^n - 1}{1-q} \tag{2.5.19}$$

$$= \lim_{n\to\infty} \frac{\left[n(1-q)\dfrac{\mathcal{H}_q(X_1)}{n}+1\right]^n - 1}{n(1-q)}. \tag{2.5.20}$$

Using the fact that $(a/n+1)^n \to e^a$, the above expression is equivalent to

$$\lim_{n\to\infty} \frac{e^{n(1-q)\mathcal{H}_q(X_1)}-1}{n(1-q)}. \tag{2.5.21}$$

When q is fixed, the limit expression in (2.5.21) increases (or decreases) quickly in n as the exponential term in the numerator dominates the linear term in the denominator. If $(1-q)\mathcal{H}_q(X_1) > 0$, we have $\mathscr{R}_q = \infty$. Conversely, if $(1-q)\mathcal{H}_q(X_1) < 0$, then $\mathscr{R}_q = 0$. Since $\mathcal{H}_q$ is always positive for discrete random variables, the rate depends on the sign of $(1-q)$. Namely,

$$\mathscr{R}_q(\mathcal{V}) = \begin{cases} 0 & \text{if } \ q > 1 \\ \infty & \text{if } \ q < 1 \end{cases} \tag{2.5.22}$$

For continuous random variables this is not necessarily true and a case by case verification is required.

In general, most interesting cases of the rate are between 0 and ∞. For example, let $q := q_n$ be a sequence depending on n such that $(1-q_n)n \to C$, for some constant C. Then, we obtain constant rate $\mathscr{R}_q(\mathcal{V}) = L_{C+1}e^{\mathcal{H}_1(X_1)}$. In particular, if $(1-q_n)n \to 0$, the usual Shannon rate $\mathscr{R}_1(\mathcal{V})$ is recovered. Finally, note that if $q_n < 1$ then $-L_{q_n}(u) < -\log(u)$, implying that $\mathcal{H}_q(\mathcal{V}) < \mathcal{H}_1(\mathcal{V})$ (with inequality reversed if $q_n > 1$).

2.6 Asymptotic equipartition. Is information additive for small n?

When dealing with information measures, there is an analog of the Law of Large Numbers: the Asymptotic Equipartition Property (AEP). Let $X_1, \ldots, X_n$ be i.i.d. observations from $p(x)$. The AEP states that $-\log p(X_1, \ldots, X_n)/n$ converges to $\mathcal{H}_1(X)$ in probability. In particular,

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) = -E_{F_n} \log p(X), \tag{2.6.1}$$

where $F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$ is the empirical distribution function. The above display emphasizes that, under the logarithmic entropy, the overall information in a sample of size n is the sum of the information given by the individual observations. Thus, if $E_p \log p(X) < \infty$, the weak Law of Large Numbers implies

$$-E_{F_n} \log p(X) \xrightarrow{p} -E_p \log p(X) = \mathcal{H}_1(X), \quad \text{as} \quad n \to \infty. \tag{2.6.2}$$

Consequently, if the sample is large enough, one can anticipate

$$-E_{F_n} \log p(X) \approx -E_p \log p(X) \tag{2.6.3}$$

and, therefore, it seems reasonable to infer about $p(x)$ based on $E_{F_n} \log p^*(X)$, where $p^*(x)$ is a target pdf (or pmf). This approximation plays a key role in a number inferential procedures such as parametric and nonparametric log-likelihood based methods. If the target density $p^*(x)$ is indexed by a parameter vector $\theta \in \Theta$, it is common practice to find the minimizer of $E_{F_n} \log p^*(X)$ and when the solution of the correspondent estimating equation exists, it is the well-known Maximum Likelihood estimator.

Clearly, for larger n, we expect (2.6.3) to be increasingly accurate. However, a reasonable question to ask is whether equipartition holds even for small samples. One way to answer this question is to verify whether the empirical redundancy equals zero:

$$E_{F_n} \rho_q(X_1, \ldots, X_n) = 0, \tag{2.6.4}$$

which from Eq. (2.5.8), this is equivalent to check

$$E_{F_n} L_q(X) = \frac{1}{n} L_q \prod_{i=1} p(X_i). \tag{2.6.5}$$

The above equality is trivially satisfied when $q = 1$ ($L_q = \log$) for any value of the X_is. However, the solution of the above expression is a random variable and if perfect additivity were true for a given sample size n, then we would expect that at least on average the solution of (2.6.5) to be equal to 1. Furthermore, if $q = 1$

and equipartition were satisfied, the right hand side of (2.6.5) is expected to be close to $E_p \log p(X)$.

The above considerations lead us to examine the average behavior of $\widehat{q}_n$, defined as

$$\widehat{q}_n = \{q : E_{F_n} L_q p(X) = E_p \log p(X)\}. \tag{2.6.6}$$

If information were perfectly additive, one can imagine that $E_p \widehat{q}_n = 1$ and the distribution of $\widehat{q}$ given $p(x)$ is symmetric about 1. However, Monte Carlo experiments for several distributions point out that this is not necessarily the case. When n is small or moderate, an asymmetric behavior with $E_p \widehat{q}_n \neq 1$ emerges instead. The following example illustrates instances of such a phenomenon for a continuous and discrete random variable.

Example 2.6.1. In this experiment, we compute Monte Carlo estimates of $\widehat{q}_n$ by solving numerically the equation in (2.6.6). Here, we show results based on 10000 samples from: (i) a Bernoulli with probability of success 0.8 and (ii) an Exponential with rate 1. The considered sample sizes range from 5 to 200. When it was not possible to find the exact zero, we considered the closest value to it in ℓ_1-norm sense.

2.7 Final Remarks

In this introductory chapter, we discussed possible generalizations of Shannon entropy, starting from its axiomatic characterization for discrete random variables. Although the postulates underlying Shannon information are intuitively reasonable and well suited for the type of problems considered by Shannon, more general notions of information have appeared in literature (Rényi, 1961; Havrda and Charvát, 1967; Aczél and Daróczy, 1975). In particular, the additivity property – or additive accumulation – of information is sometimes criticized. The reason is that additivity is the most stringent requirement as it determines completely the logarithmic form of Shannon entropy. We believe that relaxing such a requirement can lead to new and exciting statistical properties of the correspondent information measure.

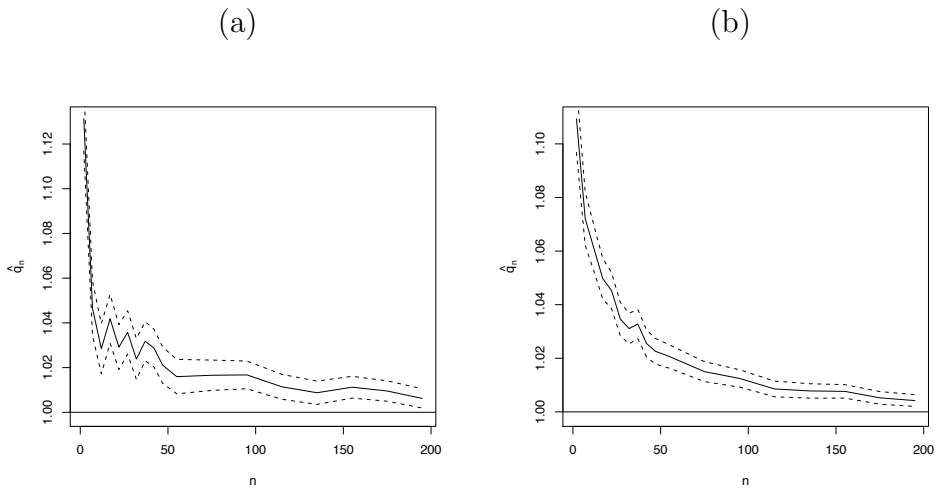


Figure 2.4: Monte Carlo means of $\widehat{q}_n$ with 99% confidence bands based on 10000 replications for a Bernoulli(0.8) (a) and an Exp(1) (b).

Among generalized information measures, we consider the entropy measure introduced by Havrda and Charvát (1967) and re-discovered by Tsallis (1988) in the context of statistical physics. One attractive feature of Havrda-Charvát-Tsallis entropy is the extension of the additivity assumption in a rather natural fashion by means of the quasi-logarithmic function $L_q(\cdot)$. Shannon's entropy is a special case, recovered when $q \to 1$. Some well-know characteristics of the q-entropy functional are presented in comparison to Shannon entropy and new properties are derived as well. In particular, we provided formulas for computing the joint information $\mathcal{H}_q(X_1, \ldots, X_n)$ in the i.i.d. case and the correspondent entropy rate. Interestingly, when q is fixed, the rate either decreases (or increases) exponentially fast to 0 (or ∞). Intermediate situations can be obtained by letting the distortion parameter q depend on the sample size. In particular, we gave conditions on the distortion parameter q for the generalized entropy rate to be equal to the usual rate of Shannon entropy.

An interesting feature of q-entropy is that such a measure explicitly takes into account the interaction information – or redundant information – in a set of variables. In particular, the joint q-entropy of the variable set $(X_1, \ldots, X_n)$ can be

understood as

$$\text{Joint Information} = \sum_i \text{Individual Information of } X_i + \text{Redundancy},$$

where the redundancy term depends on q and the case $q = 1$ corresponds to null redundancy. Since the q-entropy accounts for the additive information components as well as for the contribution given by the interaction of the individual X_is, it seems rather natural to exploit such a measure for describing non-evident dependency structures present of the sequence $X_1, \ldots, X_n$, when only marginal information about the X_is is available.

The results from a preliminary numerical study in Section 2.6 lead us to think that a "hidden" interaction structure of information may be present even for i.i.d. data when the sample size is small. The simulations show that for small or moderate samples, there is an advantage in approximating $-E \log p(x)$ by

$$\sum_{i=1}^{n} \mathcal{H}_q(X_i) = \text{Joint Information} - \text{Redundancy},$$

instead of using the usual log-likelihood functional $\sum_i \log p(X_i)$. The underlying conjecture is that when the sample size is small or moderate, even for statistically independent observations, the information is not perfectly additive and the quantity $\sum_i \mathcal{H}_q(X_i)$ allows to correct for the missing information by choosing an appropriate value of the distortion parameter q. Of course, as the sample size gets larger, we anticipate that the small sample redundancy will disappear.

2.8 Proofs

Let $K_i = \{K_{i1}, K_{i2}, \ldots, K_{iJ_i}\}$ be the set of all combinations of size i from $\{1, 2, \ldots, n\}$, where $J_i = \binom{n}{i}$ denotes the cardinality of K_i.

2.8.1 Proof of Theorem 2.5.1

The formula certainly holds when $n = 1$ and Eq. (2.3.10) holds when $n = 2$. Next, consider

$$
\begin{aligned}
&L_q\left(x_{n+1}\prod_{i=1}^{n}x_i\right)\\
&= L_q(x_{n+1}) + \sum_{i=1}^{n}(1-q)^{i-1}\sum_{j\le J_i}\prod_{k\in K_{ij}}L_q(x_k)\\
&+ (1-q)L_q(x_{n+1})\sum_{i=1}^{n}(1-q)^{i-1}\sum_{j\le J_i}\prod_{k\in K_{ij}}L_q(x_k)\\
&= L_q(x_{n+1}) + \sum_{i=1}^{n}L_q(x_i) + (1-q)\left[\sum_{j\le\binom{n}{2}}\prod_{k\in K_{ij}}L_q(x_k) + \sum_{j\le\binom{n}{1}}\prod_{k\in K_{ij}}L_q(x_k)L_q(x_{n+1})\right]\\
&+ (1-q)^2\left[\sum_{j\le\binom{n}{3}}\prod_{k\in K_{ij}}L_q(x_k) + \sum_{j\le\binom{n}{2}}\prod_{k\in K_{ij}}L_q(x_k)L_q(x_{n+1})\right]\\
&+ \cdots + (1-q)^nL_q(x_1)\cdots L_q(x_{n+1})\\
&= \sum_{i=1}^{n+1}L_q(x_i) + (1-q)\sum_{j\le\binom{n+1}{2}}\prod_{k\in K_{ij}^*}L_q(x_k) + (1-q)^2\sum_{j\le\binom{n+1}{3}}\prod_{k\in K_{ij}^*}L_q(x_k) + \cdots\\
&+ (1-q)^nL_q(x_1)\cdots L_q(x_{n+1}),
\end{aligned}
$$

where $K_{ij}^* = K_{ij}\cup\{n+1\}$. The lemma follows by induction.

2.8.2 Proof of Theorem 2.5.3

By Theorem 2.5.1 we can write

$$EL_q\left(\prod_{i=1}^{n} X_i\right) = E\sum_{i=1}^{n}(1-q)^{i-1}\sum_{j\le J_i}\prod_{k\in K_{ij}} L_q(X_k) \tag{2.8.1}$$

$$= \sum_{i=1}^{n}(1-q)^{i-1}\sum_{j\le J_i} E\prod_{k\in K_{ij}} L_q(X_k) \tag{2.8.2}$$

$$= \sum_{i=1}^{n}(1-q)^{i-1}\binom{n}{i}\left[EL_q(X_1)\right]^i, \tag{2.8.3}$$

where the last line follows by the i.i.d. assumption. By the the Binomial Theorem, we have

$$\sum_{i=0}^{n}\binom{n}{i}(1-q)^i\left[EL_q(X_1)\right]^i = \left[(1-q)EL_q(X_1)+1\right]^n. \tag{2.8.4}$$

This implies

$$EL_q\left(\prod_{i=1}^{n} X_i\right) = \frac{\left[(1-q)EL_q(X_1)+1\right]^n - 1}{1-q}. \tag{2.8.5}$$

For computing the variance, we need to introduce some matrix notation. Let $\mathbf{I}_i$ the $i\times i$ identity matrix and $\mathbf{J}_i$ a column vector of ones of length i. Define $\mathbf{\Delta}_{2^i}^{X_i}$, a diagonal matrix of dimension $2^i\times 2^i$, where d_j $(1\le j\le 2^i)$ represent the elements on the diagonal. In particular,

$$d_j := \begin{cases} 1 & \text{if } j < 2^{i-1}+1 \\ L_q(X_i) & \text{if } j = 2^{i-1}+1 \\ (1-q)L_q(X_i) & \text{if } j > 2^{i-1}+1 \end{cases}. \tag{2.8.6}$$

Then, one can write

$$1 + L_q\left(\prod_{i=1}^{n} X_i\right) \tag{2.8.7}$$

$$=1 + \sum_{i=1}^{n}(1-q)^{i-1}\sum_{j\leq J_i}\prod_{k\in K_{ij}} L_q(X_k) \tag{2.8.8}$$

$$=\mathbf{J}_{2^n}^T\left(\mathbf{I}_{2^{n-1}}\otimes\boldsymbol{\Delta}_{2^1}^{X_1}\right)\left(\mathbf{I}_{2^{n-2}}\otimes\boldsymbol{\Delta}_{2^2}^{X_2}\right)\cdots\left(\mathbf{I}_1\otimes\boldsymbol{\Delta}_{2^n}^{X_n}\right)\mathbf{J}_{2^n} \tag{2.8.9}$$

$$=\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n}\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}, \tag{2.8.10}$$

where the simple product denotes matrix product and $\otimes$ represents Kronecker product. The variance can be computed as

$$Var\left[1 + L_q\left(\prod_{i=1}^{n} X_i\right)\right] \tag{2.8.11}$$

$$=E\left[\left(\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n}\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}\right)\left(\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n}\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}\right)^T\right]$$

$$-E\left[\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n}\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}\right]E\left[\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n}\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}\right]^T. \tag{2.8.12}$$

Since $\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}$ are diagonal matrices, we can write the first summand of the above expression as

$$E\left[\mathbf{J}_{2^n}^T\prod_{i=1}^{n}\left(\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)\left(\mathbf{I}_{2^{i-1}}\otimes\boldsymbol{\Delta}_{2^i}^{X_i}\right)^T\mathbf{J}_{2^n}\right] \tag{2.8.13}$$

Moreover, since in each matrix inside the product is written in terms of independent X_i, from the above expression we have

$$\mathbf{J}_{2^n}^T \prod_{i=1}^{n} E\left[\left(\mathbf{I}_{2^{i-1}} \otimes \boldsymbol{\Delta}_{2^i}^{X_i}\right)\left(\mathbf{I}_{2^{i-1}} \otimes \boldsymbol{\Delta}_{2^i}^{X_i}\right)^T\right] \mathbf{J}_{2^n} \tag{2.8.14}$$

$$=\mathbf{J}_{2^n}^T \prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes E\left[\left(\boldsymbol{\Delta}_{2^i}^{X_i}\right)\left(\boldsymbol{\Delta}_{2^i}^{X_i}\right)^T\right] \mathbf{J}_{2^n}, \tag{2.8.15}$$

where the elements of the diagonal matrix $E\left(\boldsymbol{\Delta}_{2^i}^{X_i}\right)\left(\boldsymbol{\Delta}_{2^i}^{X_i}\right)^T$ can be expressed as Ed_j^2.

Next, consider the second summand in (3.7.2). Since the X_i are independent, we can write

$$E\left[\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes \boldsymbol{\Delta}_{2^i}^{X_i}\right)\mathbf{J}_{2^n}\right] = \mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right]\right)\mathbf{J}_{2^n}. \tag{2.8.16}$$

Therefore the second term in (3.7.2) becomes

$$\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right]\right)\left(\prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right]^T\right)\mathbf{J}_{2^n} \tag{2.8.17}$$

$$=\mathbf{J}_{2^n}^T\left(\prod_{i=1}^{n} \mathbf{I}_{2^{i-1}} \otimes E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right] E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right]^T\right)\mathbf{J}_{2^n}, \tag{2.8.18}$$

where the elements of the diagonal matrix $E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right] E\left[\boldsymbol{\Delta}_{2^i}^{X_i}\right]^T$ are $(Ed_j)^2$. Applying

backwards Eq. (2.8.10) on (2.8.18) and (3.7.6), gives

$$Var\left[1+L_q\left(\prod_{i=1}^{n}X_i\right)\right] \tag{2.8.19}$$

$$=Var\left[L_q\left(\prod_{i=1}^{n}X_i\right)\right] \tag{2.8.20}$$

$$=\sum_{i=1}^{n}(1-q)^{2(i-1)}\sum_{j\leq J_i}\prod_{k\in K_{ij}}EL_q(X_k)^2 \tag{2.8.21}$$

$$-\sum_{i=1}^{n}(1-q)^{2(i-1)}\sum_{j\leq J_i}\prod_{k\in K_{ij}}\left[EL_q(X_k)\right]^2 \tag{2.8.22}$$

$$=\sum_{i=1}^{n}(1-q)^{2(i-1)}\sum_{j\leq J_i}\prod_{k\in K_{ij}}EL_q(X_k)^2-\prod_{k\in K_j}\left[EL_q(X_k)\right]^2 \tag{2.8.23}$$

Since the X_is are identically distributed, the above expressions can be written as

$$\sum_{i=1}^{n}\binom{n}{i}(1-q)^{2(i-1)}\left(EL_q(X_1)^2\right)^i-EL_q(X_1)^{2i} \tag{2.8.24}$$

$$=\sum_{i=1}^{n}\binom{n}{i}(1-q)^{2(i-1)}VarL_q(X_1)^i. \tag{2.8.25}$$

Applying the binomial theorem on (2.8.25), gives

$$Var\left[L_q\left(\prod_{i=1}^{n}X_i\right)\right]=\frac{\left((1-q)^2VarL_q(X_1)+1\right)^n-1}{(1-q)^2}. \tag{2.8.26}$$

This completes part (i) of the Theorem.

Next, consider a sequence q_n, converging to 1. Since $(1+a/n)^n\to e^a$,

$$n^{-1}EL_{q_n}\left(\prod_{i=1}^{n}X_i\right)=n^{-1}\frac{\left[(1-q_n)EL_{q_n}(X_1)+1\right]^n-1}{1-q_n} \tag{2.8.27}$$

is asymptotically equivalent to

$$\frac{e^{EL_{q_n}(X_1)n(1-q_n)}-1}{n(1-q_n)}. \tag{2.8.28}$$

Since $n(1-q) \to 0$ and $EL_{q_n}(X_1) \to E\log(X_1)$ as $n \to \infty$, we have that Eq. (2.8.28) converges to $E\log(X_1)$. A similar calculation shows

$$\frac{((1-q)^2 Var L_{q_n}(X_1)+1)^n - 1}{(1-q_n)^2} \to Var\log(X_1), \tag{2.8.29}$$

as $n \to \infty$.

Chapter 3

Maximum Lq-Likelihood estimation

In this chapter, the Maximum Lq-Likelihood Estimator (MLqE), a new parameter estimator based on nonextensive entropy (Havrda and Charvát, 1967), is introduced. The properties of the MLqE for an exponential family are studied via asymptotic analysis and computer simulations. The behavior of the MLqE is characterized by the degree of distortion q applied to the assumed model. When q is properly chosen for small and moderate sample sizes, the MLqE successfully trades bias for precision, resulting in a substantial reduction of the mean squared error. When the sample size is large and q tends to 1, a necessary and sufficient condition to ensure a proper asymptotic normality and efficiency of MLqE is established.

3.1 Introduction

One of the major contributions to scientific thought of the last century is information theory founded by Claude Shannon in the late 1940s. Its triumph is highlighted by countless applications in various scientific domains including statistics. The fulcrum of information theory is a measure of the "amount of uncertainty" inherent in a probability distribution (usually called Shannon entropy). Provided a distribution $p(x)$, Shannon's entropy is defined as $\mathcal{H}(X) = -E[\log p(X)]$. The quantity $-\log p(x)$ is interpreted as the information content of the outcome x and

$\mathcal{H}(X)$ represents the average uncertainty removed after the actual outcome of X is revealed. The connection between logarithmic (or additive) entropies and inference has been copiously studied; see, e.g., Cover and Thomas (2006). Akaike (1973) introduced a principle of statistical model building based on minimization of expected entropy. In a parametric setting, he pointed out that the usual inferential task of maximizing the log-likelihood function can be equivalently regarded as minimization of the empirical version of Shannon's entropy, $-\sum_{i=1}^{n} \log p(X_i)$. Rissanen proposed the well known minimum description length criterion for model comparison (see e.g., Barron et al. (1998)).

Since the introduction of Shannon's entropy, other and more general measures of information have been developed. Rényi (1961), Aczél and Daróczy (1975) in the mid 60s and 70s proposed generalized notions of information (usually referred as Rényi entropies) by keeping the additivity of independent information but using a more general definition of mean. In a different direction, Havrda and Charvát (1967) proposed *nonextensive* entropies, sometimes referred to as q-order entropies, where the usual definition of mean is maintained while the logarithm is replaced by the more general function $L_q(u) = (u^{1-q} - 1)/(1 - q)$ for $q > 0$. In particular, when $q \to 1$, $L_q(u) \to \log(u)$, recovering the usual Shannon's entropy.

In recent years, q-order entropies have been of considerable interest in different domains of application. Tsallis and colleagues have successfully exploited them in physics (see e.g. Tsallis (1988) and Tsallis et al. (1998)). In thermodynamics, the q-entropy functional is usually minimized subject to some properly chosen constraints, according to the formalism proposed by Jaynes (1957a,b). There is a large literature on analyzing various loss functions as the convex dual of entropy minimization, subject to constraints. From this standpoint, the classical maximum entropy estimation and maximum likelihood are seen as convex duals of each other (see, e.g. Altun and Smola (2006)). Since Tsallis' seminal paper (Tsallis, 1988), q-order entropy has encountered an increasing wave of success and Tsallis' nonextensive thermodynamics, based on such information measure, is nowadays considered the most viable candidate for generalizing the ideas of the famous Boltzmann-Gibbs theory. More recently, a number of applications based on the q-entropy have appeared in other disciplines such as finance, biomedical

sciences, environmental sciences and linguistics (Gell-Mann, 2004).

Despite the broad success, so far little effort has been made to address the inferential implications of using nonextensive entropies from a statistical perspective. In this chapter, we study a new class of parametric estimators based on the q-entropy function, the Maximum Lq-Likelihood Estimator (MLqE). In our approach, the role of the observations is modified by slightly changing the model of reference by means of the distortion parameter q. From this standpoint, Lq-likelihood estimation can be regarded as the minimization of the discrepancy between a general distribution and one that modifies the true distribution to diminish (or emphasize) the role of extreme observations.

In this framework, we provide theoretical insights concerning the statistical usage of the generalized entropy function. In particular, we highlight the role of the distortion parameter q and give the conditions that guarantee asymptotic efficiency of the MLqE. Further, the new methodology is shown to be very useful when estimating high-dimensional parameters and small tail probabilities. This aspect is important in many applications, where we must deal with the fact that the number of observations available is not large in relation to the number of parameters or the probability of occurrence of the event of interest. Standard large sample theory guarantees that the Maximum Likelihood Estimator (MLE) is asymptotically efficient, meaning that when the sample size is large, the MLE is at least as accurate as any other estimator. However, for a small or moderate sample size, it turns out that the MLqE can produce a dramatic improvement in terms of mean squared error at the expense of a slightly increased bias, as will be seen in our numerical results.

For finite sample performance of MLqE, not only the size of $q_n - 1$ but also its sign (i.e., the direction of distortion) are important. It turns out that for different families or different parametric functions of the same family, the beneficial direction of distortion can be different. In addition, for some parameters, MLqE does not produce any improvement. We have found that an asymptotic variance expression of a MLqE is very helpful to decide the direction of distortion for applications.

The chapter is organized as follows. In section 3.2, we examine some information-theoretical quantities and introduce the MLqE; in section 3.3, we present its basic

asymptotic properties. In particular, a necessary and sufficient condition on the choice of q in terms of n to ensure a proper asymptotic normality and efficiency is established. In section 3.4, we consider the plug-in approach for tail probability estimation based on MLqE. The asymptotic properties of the plug-in estimator are derived and its efficiency is compared to the traditional MLE. In section 3.5, we present Monte Carlo simulations and examine the behavior of MLqE in finite sample situations. In section 3.6, concluding remarks are given. Technical proofs of the theorems are deferred to an appendix.

3.2 Generalized entropy and the Maximum Lq-Likelihood Estimator

Consider a measure μ on a measurable space Ω. The Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951; Kullback, 1959) (or relative entropy) between two density functions g and f with respect to μ is

$$\mathcal{D}(f||g) = E_f \log \frac{f(X)}{g(X)} = \int_\Omega f(x) \log \frac{f(x)}{g(x)} d\mu(x). \tag{3.2.1}$$

Note that finding the density g that minimizes $\mathcal{D}(f||g)$ is equivalent to minimizing Shannon's entropy $\mathcal{H}(f, g) = -E_f \log g(X)$.

Definition 3.2.1. *Let f and g be two density functions. The q-entropy of g with respect to f is defined as*

$$\mathcal{H}_q(f, g) = -E_f L_q \{g(X)\}, \quad q > 0, \tag{3.2.2}$$

where $L_q(u) = \log u$ if $q = 1$ and $L_q(u) = (u^{1-q} - 1)/(1 - q)$ otherwise.

The function L_q represents a Box-Cox transformation in statistics and in other contexts it is often called a deformed logarithm. Note that if $q \to 1$, then $L_q(u) \to \log(u)$ and the usual definition of Shannon entropy is recovered.

Let $\mathscr{M} = \{f(x; \theta), \theta \in \Theta\}$ be a family of parametrized density functions and suppose that the true density of observations, denoted by $f(x; \theta_0)$, is a member of

$\mathscr{M}$. Assume further that $\mathscr{M}$ is closed under the transformation

$$f(x;\theta)^{(r)} = \frac{f(x;\theta)^r}{\int_\Omega f(x;\theta)^r d\mu(x)},\ r > 0. \tag{3.2.3}$$

The transformed density $f(x;\theta)^{(r)}$ is often referred to as *zooming* or *escort* distribution (Naudts, 2004; Abe, 2003; Beck and Schlögl, 1993) and the parameter r provides a tool to accentuate different regions of the untransformed true density $f(x;\theta)$. In particular, when $r < 1$ regions with density values close to zero are accentuated, while for $r > 1$ regions with density values further from zero are emphasized.

Consider the KL divergence between $f(x;\theta)$ and $f(x;\theta_0)^{(r)}$:

$$\mathcal{D}_r(\theta_0||\theta) = \int_\Omega f(x;\theta_0)^{(r)} \log \frac{f(x;\theta_0)^{(r)}}{f(x;\theta)} d\mu(x). \tag{3.2.4}$$

Let θ^* be the value such that $f(x;\theta^*) = f(x;\theta_0)^{(r)}$ and assume that differentiation can be passed under the integral sign. Then, clearly θ^* minimizes $\mathcal{D}_r(\theta_0||\theta)$ over θ. Let θ^{**} be the value such that $f(x;\theta^{**}) = f(x;\theta_0)^{(1/q)}$, $q > 0$. Since we have $\nabla_\theta \mathcal{H}_q(\theta_0,\theta)|_{\theta^{**}} = 0$ and $\nabla^2_\theta \mathcal{H}_q(\theta_0,\theta)|_{\theta^{**}}$ is positive definite, $\mathcal{H}_q(\theta_0,\theta)$ has a minimum at θ^{**}.

The derivations above suggest that the task of minimizing the q-entropy can be regarded as equivalent to minimizing the KL divergence between the escort transformation of the true density and the density under exam, when $q = r^{-1}$. Clearly, by considering the divergence with respect to a distorted version of the true density we introduce a certain amount of bias. Nevertheless, the bias can be promptly controlled by an adequate choice of the distortion parameter q, and later we shall discuss the benefits gained from paying such a price for tail probability estimation. The next definition introduces the estimator based on the empirical version of the q-entropy.

Definition 3.2.2. *Let $X_1, ..., X_n$ be an i.i.d. sample from $f(x;\theta_0)$, $\theta_0 \in \Theta$. The*

Maximum L_q-Likelihood Estimator (MLqE) of θ_0 is defined as

$$\widetilde{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} L_q \left[f(X_i;\theta) \right], \quad q > 0, \tag{3.2.5}$$

where L_q is the q-logarithmic function defined in (3.2.2) with $q > 0$.

When $q \to 1$, if the estimator $\widetilde{\theta}_n$ exists, then it approaches the maximum likelihood estimator of the parameters, which maximizes $\sum_i \log f(X_i;\theta)$. In this sense, the ML$q$E extends the classic method, resulting in a general inferential procedure that inherits most of the desirable features of traditional maximum likelihood, and at the same time can improve over the MLE due to variance reduction, as will be seen.

Define

$$U(x;\theta) = \nabla_\theta \log \left\{ f(x;\theta) \right\},$$

$$U^*(X;\theta,q) = U(X;\theta) f(X;\theta)^{1-q}.$$

In general, the estimating equations have the form

$$\sum_{i=1}^{n} U^*(X_i;\theta,q) = 0. \tag{3.2.6}$$

Eq. (3.2.6) offers a natural interpretation of the MLqE as a solution to a *weighted* likelihood. When $q \neq 1$, Eq.(3.2.6) provides a relative-to-the-model downweighting. Observations that disagree with the model receive low weight. In the case $q = 1$, all the observations receive the same weight. The strategy of setting weights that are proportional to the family from which the model is to be chosen appeared in various contexts in literature (e.g., see Windham (1995) and Choi et al. (2000)).

Example 3.2.1. The simple but illuminating case of an exponential distribution will be used as a recurrent example in the course of the chapter. Consider an i.i.d. sample of size n from a distribution with density $\lambda_0 \exp\{-x\lambda_0\}$, $x > 0$ and $\lambda_0 > 0$.

In this case, the L_q-likelihood equation is

$$\sum_{i=1}^{n} e^{-[X_i\lambda-\log\lambda](1-q)}\left(-X_i+\frac{1}{\lambda}\right)=0. \tag{3.2.7}$$

With $q=1$, the usual maximum likelihood estimator is $\widehat{\lambda}=(\sum_i X_i/n)^{-1}=\overline{X}^{-1}$. However, when $q\neq 1$, equation (3.2.7) can be rewritten as

$$\lambda=\left(\frac{\sum_{i=1}^{n} X_i w_i\left(X_i,\lambda,q\right)}{\sum_{i=1}^{n} w_i\left(X_i,\lambda,q\right)}\right)^{-1}. \tag{3.2.8}$$

where $w_i := e^{-[X_i\lambda-\log\lambda](1-q)}$. When $q<1$ the role played by observations corresponding to higher density values are accentuated; when $q>1$ observations corresponding to density values close to zero are accentuated.

3.3 Exponential families and asymptotics of the MLqE

In this section, we discuss the asymptotic properties of the new estimator when the degree of distortion is chosen according to the sample size. In the remainder of the chapter we focus on the exponential families. In particular, we consider density functions of the form:

$$f(x;\theta)=\exp\left\{\theta^{\intercal}b(x)-A(\theta)\right\}, \tag{3.3.1}$$

where $\theta\in\Theta\subseteq\mathbb{R}^p$ is a real valued natural parameter vector, $b(x)$ is the vector of functions with elements $b_j(x)$ $(j=1,\ldots,p)$ and $A(\theta)=\log\int_\Omega e^{\theta^{\intercal}b(x)}d\mu(x)$ is the cumulant generating function (or log normalizer). The true parameter will be denoted by θ_0. Next, we explore consistency, which is a basic requirement for a good estimator.

3.3.1 Consistency

Consider θ_n^*, the value such that

$$E_{\theta_0} U^*(X; \theta_n^*, q_n) = 0. \tag{3.3.2}$$

It can be easily shown that $\theta_n^* = \theta_0/q_n$. Since the actual target of $\widetilde{\theta}_n$ is θ_n^*, to retrieve asymptotic unbiasedness of $\widetilde{\theta}_n$, q_n must converge to 1. We call θ_n^* the *surrogate* parameter of θ_0. We impose the following conditions:

A.1 $q_n > 0$ is a monotone sequence such that $q_n \to 1$ as $n \to \infty$.

A.2 The parameter space Θ is compact and the parameter θ_0 is an interior point in Θ.

In similar contexts, the compactness condition on Θ is used for technical reasons (see e.g., Wang et al. (2004)) as it is the case here.

Theorem 3.3.1. *Under Assumptions A.1 and A.2, with probability going to 1, the L_q-likelihood equation yields a unique solution $\widetilde{\theta}_n$ that is the maximizer of the L_q-likelihood function in Θ. Furthermore, we have $\widetilde{\theta}_n \xrightarrow{P} \theta_0$.*

Remark. When Θ is compact, the MLqE always exists under our conditions, although it is not necessarily unique with probability one.

3.3.2 Asymptotic normality

Theorem 3.3.2. *If Assumptions A.1 and A.2 hold, then we have*

$$\sqrt{n}\; V_n^{-1/2}\left(\widetilde{\theta}_n - \theta_n^*\right) \xrightarrow{\mathscr{D}} N_p\left(0, \mathbf{I}_p\right) \quad as \quad n \to \infty, \tag{3.3.3}$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix, $V_n = J_n^{-1} K_n J_n^{-1}$ and

$$K_n = E_{\theta_0}\left[U^*(X; \theta_n^*, q_n)\right]^T \left[U^*(X; \theta_n^*, q_n)\right] \tag{3.3.4}$$

$$J_n = E_{\theta_0}\left[\nabla_\theta U^*(X; \theta_n^*, q_n)\right]. \tag{3.3.5}$$

A necessary and sufficient condition for asymptotic normality of MLqE around θ_0 *is* $\sqrt{n}(q_n - 1) \to 0$.

Let $m(\theta) := \nabla_\theta A(\theta)$ and $D(\theta) := \nabla^2_\theta A(\theta)$. Note that K_n and J_n can be expressed as

$$K_n = c_{2,n}\left(D(\theta_{2,n}) + [m(\theta_{2,n}) - m(\theta^*_n)][m(\theta_{2,n}) - m(\theta^*_n)]^{\mathtt{T}}\right), \tag{3.3.6}$$

and

$$J_n = c_{1,n}(1 - q_n)D(\theta_{1,n}) - c_{1,n}D(\theta^*_n) \tag{3.3.7}$$

$$+ c_{1,n}(1 - q_n)[m(\theta_{1,n}) - m(\theta^*_n)][m(\theta_{1,n}) - m(\theta^*_n)]^{\mathtt{T}}, \tag{3.3.8}$$

where $c_{k,n} = \exp\{A(\theta_{n,k}) - A(\theta_0)\}$ and $\theta_{k,n} = k\theta_0(1/q_n - 1) + \theta_0$. When $q_n \to 1$, it is seen that $V_n \to -D(\theta_0)$, the asymptotic variance of the MLE. When $\Theta \subseteq \mathbb{R}^1$ we use the notation σ^2_n for the asymptotic variance in place of V_n. Note that the existence of moments are ensured by the functional form of the exponential families (e.g., see Lehmann and Casella (1998)).

Remarks. (i) When q is fixed, the MLqE is a regular M-estimator (Huber, 1981a), which converges in probability to $\theta^* = \theta_0/q$. (ii) With the explicit expression of θ^*_n, one may consider correcting the bias of MLqE by using the estimator $q_n\widetilde{\theta}_n$. The numerical results are not promising in this direction.

Example 3.3.1 (Exponential distribution)**.** The surrogate parameter is $\theta^*_n = \lambda_0/q_n$ and a lengthy but straightforward calculation shows that the asymptotic variance of the MLqE of λ_0 is

$$\sigma^2_n = \left(\frac{\lambda_0}{q_n}\right)^2 \left[\frac{q_n^2 - 2q_n + 2}{q_n^3\,(2 - q_n)^3}\right] \to \lambda_0^2 \tag{3.3.9}$$

as $n \to \infty$. By Theorem 3.3.2, we conclude that $n^{1/2}\sigma_n^{-1}(\widetilde{\lambda}_n - \lambda_0/q_n)$ converges weakly to a standard normal distribution as $n \to \infty$. Clearly, the asymptotic calculation does not produce any advantage of MLqE in terms of reducing the limiting variance. However, for an interval of q_n, we have $\sigma^2_n < \lambda^2_0$ (see Section 3.4.3) and, based on our simulations, an improvement of the accuracy is achieved

in finite sample sizes as long as $0 < q_n - 1 = o(n^{-1/2})$, which ensures a proper asymptotic normality of $\widetilde{\lambda}_n$. For the re-scaled estimator $q_n\widehat{\lambda}_n$, the expression $q_n^2\sigma_n^2$ is larger than 1 unless $q = 1$, which suggests that $q_n\widehat{\lambda}_n$ may be at best no better than $\widehat{\lambda}_n$.

Example 3.3.2 (Multivariate normal distribution)**.** Consider a multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Two convenient matrix operators in this setting are the vec$(\cdot)$ (vector) and vech$(\cdot)$ (vector-half). Namely, vec : $\mathbb{R}^{r\times p} \mapsto \mathbb{R}^{rp}$ stacks the columns of the argument matrix. For symmetric matrices, vech : $\mathbb{S}^{p\times p} \mapsto \mathbb{R}^{p(p+1)/2}$ stacks only the unique part of each column that lies on or below the diagonal (McCulloch, 1982). Further, for a symmetric matrix M, define the extension matrix G as vec$M = G$ vech$M$. Thus, $\theta_0 = (\boldsymbol{\mu}^T, \text{vech}^\intercal\boldsymbol{\Sigma})^\intercal$ and under such a parametrization, its easy to show the surrogate parameter solving (3.3.2) is $\theta_n^* = (\boldsymbol{\mu}^\intercal, \sqrt{q_n}\text{vech}^\intercal\boldsymbol{\Sigma})^\intercal$, where interestingly the mean component does not depend on q_n. In fact, for symmetric distributions about the mean, it can be shown that the distortion imposed to the model affects the spread of the distribution but leaves the mean unchanged. Consequently, the MLqE is expected to improve the estimation of $\boldsymbol{\Sigma}$ without much effect on $\boldsymbol{\mu}$. This will be clearly seen in our simulation results (see Section 3.5.4). The calculation in Appendix B shows that the asymptotic variance of the MLqE of θ_0 is the block-diagonal matrix

$$V_n = \begin{pmatrix} \frac{(2-q)^{2+p}}{(3-2q)^{1+\frac{p}{2}}}\boldsymbol{\Sigma} & 0 \\ 0 & \frac{4q^2[(3-2q)^2+1](2-q)^{4+p}}{[(2-q)^2+1]^2(3-2q)^{2+p/2}}[G^\intercal(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})G]^{-1} \end{pmatrix}. \tag{3.3.10}$$

3.4 Estimation of the tail probability

In this section, we address the problem of tail probability estimation, using the popular plug-in procedure, where the point estimate of the unknown parameter is substituted into the parametric function of interest. We focus on one-dimensional case, i.e. $p = 1$, and derive the asymptotic distribution of the plug-in estimator for the tail probability based on the MLq method.

Let $\alpha(x;\theta) = P_\theta(X \le x)$ or $\alpha(x;\theta) = 1 - P_\theta(X \le x)$, depending on whether we are considering the lower tail or the upper tail of the distribution. Without

loss of generality, we focus on the latter from now on, and assume $\alpha(x;\theta) > 0$ for all x (of course $\alpha(x;\theta) \to 0$ as $x \to \infty$). When x is fixed, under some conditions, the familiar delta method shows that $\sqrt{n}[\sigma_n^{1/2}\alpha'(x;\theta_n^*)]^{-1}[\alpha(x;\widetilde{\theta}_n) - \alpha(x;\theta_n^*)]$ converges weakly to a standard Normal distribution as $n \to \infty$. However, in most applications a large sample size is usually demanded in order to obtain accurate estimates of a small tail probability in terms of relative error. As a consequence, the classical problem setup with x fixed may be inadequate, as it ignores possible difficulty due to smallness of the tail probability in relation to the sample size n. We consider instead a framework that better reflects the nature of the problem.

3.4.1 Asymptotic normality of the plug-in MLq estimator

We are interested in estimating $\alpha(x_n;\theta_0)$, where $x_n \to \infty$ as $n \to \infty$. For $\theta^* \in \Theta$ and $\delta > 0$, define

$$\beta(x;\theta^*;\delta) = \sup_{\theta \in \Theta \cap [\theta^* - \frac{\delta}{\sqrt{n}}, \theta^* + \frac{\delta}{\sqrt{n}}]} \left| \frac{\alpha''(x;\theta)}{\alpha''(x;\theta^*)} \right|, \tag{3.4.1}$$

and $\gamma(x;\theta) = \alpha''(x;\theta)/\alpha'(x;\theta)$.

Theorem 3.4.1. *Let θ_n^* be as in the previous section. Under Assumptions A.1 and A.2, if $n^{-1/2}\,|\gamma(x_n;\theta_n^*)|\,\beta(x_n;\theta_n^*;\delta) \to 0$ for each $\delta > 0$, then*

$$\sqrt{n}\frac{\left(\alpha(x_n;\widetilde{\theta}_n) - \alpha(x_n;\theta_n^*)\right)}{\sigma_n \alpha'(x_n;\theta_n^*)} \xrightarrow{\mathscr{D}} N(0,1)\,,$$

where $\sigma_n = E_{\theta_0}[U^*(X;\theta_n^*)^2]/E_{\theta_0}[\partial U^*(X;\theta,q_n)/\partial\theta|_{\theta_n^*}]$.

Remarks. (i) For the main requirement of the theorem on the order of the sequence x_n, it is easiest to be verified on a case by case basis. For instance, in the case of the exponential distribution in Eq.(3.7.3), for $x_n > 0$,

$$\beta(x_n;\lambda_n^*;\delta) = \sup_{\lambda \in \lambda_n^* \pm \frac{\delta}{\sqrt{n}}} \frac{e^{-x_n\lambda}x_n^2}{e^{-x_n\lambda_n^*}x_n^2} \le \sup_{\lambda \in \lambda_n^* \pm \frac{\delta}{\sqrt{n}}} e^{x_n|\lambda - \lambda_n^*|} = e^{\frac{\delta x_n}{\sqrt{n}}}.$$

Moreover, $\gamma(x_n;\lambda_n^*) = -x_n$. So, the condition reads $n^{-1/2}x_n e^{\frac{\delta x_n}{\sqrt{n}}} \to 0$, i.e., $n^{-1/2}x_n \to 0$. (ii) The plug-in estimator based on $q_n\widetilde{\theta}_n$ has been examined as well. With $q_n \to 1$, we did not find any significant advantage.

In the new setting we relate explicitly the amount of information available in the sample to both the "rarity" of the event under examination and the distortion parameter. In the next section, we will use this new framework to compare the MLqE of the tail probability, $\alpha(x_n;\widetilde{\theta}_n)$, with the one based on the traditional MLE, $\alpha(x_n;\widehat{\theta}_n)$.

In many applications, the quantity of interest is quantile instead of the tail probability. In our setting, the quantile function is defined as $\rho(s;\theta) = \alpha^{-1}(s;\theta)$, $0 < s < 1$ and $\theta \in \Theta$. Next, we present the analogue of Theorem 3.3.2 for the plug-in estimator of the quantile. Define

$$\beta_1(s;\theta^*;\delta) = \sup_{\theta\in\Theta\cap[\theta^*-\frac{\delta}{\sqrt{n}},\theta^*+\frac{\delta}{\sqrt{n}}]} \left|\frac{\rho''(s;\theta)}{\rho'(s;\theta^*)}\right|, \ \delta > 0 \tag{3.4.2}$$

and $\gamma_1(s;\theta) = \rho''(s;\theta)/\rho'(s;\theta)$.

Theorem 3.4.2. *Let $0 < s_n < 1$ be a nonincreasing sequence such that $s_n \searrow 0$ as $n \to \infty$ and let θ_n^* and q_n be as in Theorem 3.4.1. Under Assumptions A.1 through A.3, for every sequence s_n such that $n^{-1/2} \mid \gamma_1(x_n;\theta_n^*)\mid \ \beta_1(x;\theta_n^*;\delta) \to 0$ for each $\delta > 0$,*

$$\sqrt{n}\frac{\Big(\rho(s_n;\widetilde{\theta}_n) - \rho(s_n;\theta_n^*)\Big)}{\sigma_n\rho'(s_n;\theta_n^*)} \xrightarrow{\mathscr{D}} N(0,1).$$

3.4.2 Relative efficiency between MLE and MLqE

In Section 3.3, we we showed that when $(q_n - 1)\sqrt{n} \to 0$, the MLqE is asymptotically as efficient as the MLE. For tail probability estimation, with $x_n \to \infty$, it is unclear if MLqE performs efficiently.

Consider w_n and v_n, two estimators of a parametric function $g_n(\theta)$ such that both $\sqrt{n}\,(w_n - a_n)/\sigma_n$ and $\sqrt{n}\,(v_n - b_n)/\tau_n$ converge weakly to a standard Normal distribution as $n \to \infty$ for some deterministic sequences a_n, b_n, $\sigma_n > 0$ and $\tau_n > 0$.

Definition 3.4.1. *Define*

$$\Lambda(w_n, v_n) := \frac{(b_n - g_n(\theta))^2 + \tau_n^2/n}{(a_n - g_n(\theta))^2 + \sigma_n^2/n}. \tag{3.4.3}$$

The bias adjusted asymptotic relative efficiency of w_n with respect to v_n is $\lim_{n\to\infty} \Lambda(w_n, v_n)$, provided that the limit exists.

It can be easily verified that the definition does not depend on the specific choice of a_n, b_n, σ_n and τ_n among equivalent expressions.

Corollary 3.4.3. *Under the conditions of Theorem 3.4.1, when q_n is chosen such that*

$$n^{1/2}\alpha(x_n; \theta_n^*)\alpha(x_n; \theta_0)^{-1} \to 1 \text{ and } \alpha'(x_n; \theta_n^*)\alpha'(x_n; \theta_0)^{-1} \to 1, \tag{3.4.4}$$

then $\Lambda(\alpha(x_n; \widehat{\theta}_n), \alpha(x_n; \widetilde{\theta}_n)) = 1$.

The result says that when q_n is chosen sufficiently close to 1, asymptotically speaking, the MLqE is as efficient as the MLE.

Example 3.4.1 (Continued)**.** In this case, we have $\alpha(x_n; \lambda) = e^{-\lambda x_n}$ and $\alpha'(x_n; \lambda) = -x_n e^{-\lambda x_n}$. For sequences x_n and q_n such that $x_n/\sqrt{n} \to 0$ and $(q_n - 1)\sqrt{n} \to 0$, we have that

$$\sqrt{n}\frac{\left(e^{-\widetilde{\lambda}_n x_n} - e^{-\frac{\lambda_0}{q_n} x_n}\right)}{\lambda_0 x_n e^{-\frac{\lambda_0}{q_n} x_n}} \xrightarrow{\mathscr{D}} N(0, 1). \tag{3.4.5}$$

When $q_n = 1$ for all n, we recover the usual plug-in estimator based on MLE. With the asymptotic expressions given above,

$$\Lambda(\alpha(x; \widehat{\lambda}_n), \alpha(x; \widetilde{\lambda}_n)) = \frac{n}{\lambda_0^2 x_{L,n}^2}(e^{-x_n(\lambda_0/q_n - \lambda_0)} - 1)^2 + e^{-2x_n(\lambda_0/q_n - \lambda_0)}, \tag{3.4.6}$$

which is greater than 1 when $q_n > 1$. Thus, no advantage in terms of MSE is expected by considering $q_n > 1$ (which introduces bias and enlarges the variance at the same time).

Although in limits MLqE is not more efficient than MLE, MLqE can be much better than MLE due to variance reduction as will be clearly seen in Section 3.5.

The following calculation provides a heuristic understanding. Let $r_n = 1 - 1/q_n$. Add and subtract 1 in (3.4.6), obtaining

$$\frac{nr_n^2 L_{q_n}\left(e^{-x_n\lambda_0}\right)^2}{\lambda_0^2 x_{L,n}^2} + r_n L_{q_n}\left(e^{2x_n\lambda_0}\right) + 1 < nr_n^2 + r_n 2x_n\lambda_0 + 1, \tag{3.4.7}$$

where the last inequality holds as $L_{1/q_n}(u) < \log(u)$ for any $u > 0$ and $q < 1$. Next, we impose (3.4.7) to be smaller than 1 and solve for q_n, obtaining

$$T_n := \left(1 + \frac{2\lambda_0 x_n}{n}\right)^{-1} < q_n < 1. \tag{3.4.8}$$

This provides some insights on the choice of the sequence q_n in accordance to the size of the probability to be estimated. If q_n approaches 1 too quickly from below, the gain obtained in terms of variance vanishes rapidly as n becomes larger. On the other hand, if q_n converges to 1 too slowly, the bias part dominates the variance and the MLE outperforms the MLqE. This understanding is confirmed in our simulation study.

3.4.3 Discussion

1. For estimating θ_0, when $q_n \to 1$, the asymptotic variance of MLqE is equivalent to that of MLE. When q_n is fixed, a non-vanishing bias is introduced, but an advantage may be obtained in terms of variance reduction. For instance, in the variance expression (3.3.9) one can easily check that $(q^2 - 2q + 2)/[q^5(2-q)^3] < 1$ for $1 < q < 1.40069$; thus, choosing the distortion parameter in such a range gives $\sigma_n^2 < \lambda_0^2$.

2. For estimating the tail probability, when $q_n \to 1$, the asymptotic variance of MLqE can be of a smaller order than that of MLE, although there is a bias that approaches 0. In particular:

   (i) MLqE cannot be asymptotically more efficient than MLE.

   (ii) MLqE is asymptotically as efficient as MLE when q_n is chosen to be close enough to 1. In the case of tail probability for the exponential

distribution we need to choose q_n such that $(q_n - 1)x_n \to 0$.

(iii) For such an order of q_n, although the MLqE and MLE are asymptotically equivalent in efficiency, we conjecture that there is a higher order effect favoring the MLqE due to variance reduction. This is supported by our simulations results.

(iv) The best choice of the distortion parameter to minimize the mean square error depends upon the form of the density function of interest, which is involved in the calculation of the variance of the plug-in estimator of the tail probability. For distributions belonging to the exponential family, it can be shown that $0 < q_n < 1$. When $q_n \searrow 1$ slowly enough compared to $|b^{-1}(x_n)|$, the variance of the plug-in estimator becomes increasingly larger than that of MLE.

3.4.4 Choice of the distortion parameter q

When applying the MLqE method, an important issue is the selection of the distortion parameter q, in terms of the sample size. One possible approach to handling this problem is to minimize the estimated asymptotic mean squared error of the estimator when it is mathematically treatable. In the case of the exponential distribution, by Theorem 3.4.1 we have the following expression for the asymptotic mean squared error:

$$MSE(q, \lambda_0) = \left(e^{-\frac{\lambda_0}{q}x_n} - e^{-\lambda_0 x_n}\right)^2 + \left(\frac{\lambda_0}{q}\right)^2 n^{-1} \left(\frac{q^2 - 2q + 2}{q^3(2-q)^3}\right) x_n^2 e^{-2\frac{\lambda_0}{q}x_n}. \tag{3.4.9}$$

However, since λ_0 is unknown, we consider

$$q^* = \underset{q \in (0,1)}{\arg\min} \left\{ MSE(q, \widehat{\lambda}) \right\}, \tag{3.4.10}$$

where $\widehat{\lambda}$ is the MLE. This will be used in our simulation study.

3.5 Monte Carlo results

In this section, the performance of the MLqE in finite samples is explored via simulations. Our study includes: (i) assessment of accuracy for tail probability estimation and reliability of confidence intervals (ii) assessment of the performance of MLqE for estimating multidimensional parameters, including regression settings with generalized linear models. The standard MLE is used as a benchmark throughout the study.

3.5.1 On the choice of q

Clearly, the choice of q_n influences the finite sample performance of MLqE. In the literature on applications of non-extensive entropy, although some discussions on choosing q have been made often from physical considerations, it is unclear how to do it from statistics perspectives. In particular, the direction of distortion (i.e., $q > 1$ or $q < 1$) needs to be decided. We have the following observations/thoughts.

1. For estimating the parameters in an exponential family, although $|q_n - 1|n^{-1/2}$ guarantees the right asymptotic normality, one direction of distortion typically reduces the variance of estimation and consequently improves the MSE. In the exponential distribution case, q_n needs to be slightly greater than 1, but for estimating the covariance matrix for multivariate normal observations, q_n needs to be slightly smaller than 1. In section 3, an asymptotic covariance matrix for the MLqE is given. For a given family, the expression can be used to find the beneficial direction of distortion. Our numerical investigations confirm this understanding.

2. For tail probability estimation for the exponential distribution case, q_n needs to be larger than 1 for better performance, which is the opposite for estimating the parameter λ itself. Thus the optimal choice of q_n (in particular to emphasize or de-emphasize the larger density values) is not a characteristic of the family but also depends on the parametric function to be estimated.

3. For some parametric functions, the MLqE makes little change. For the multivariate normal family, the surrogate value of the mean parameter stays

exactly the same while the variance parameters are altered.

4. We have found that given the right distortion direction, choices of q_n with $|1-q_n|$ between $1/n$ and $1/\sqrt{n}$ usually improves – to different extents – over the MLE.

In this section, we present both deterministic and data driven approaches on choosing q_n. First, deterministic choices are used to explore the possible advantage of the MLqE for tail probability estimation with q_n approaching to 1 fast when x is fixed and q_n approaching to 1 slowly when x increases with n. Then, the data driven choice in Section 3.4.4 is applied. For multivariate normal and GLM families, where estimation of the MSE or prediction error becomes analytically cumbersome, we choose $q_n = 1 - 1/n$, which satisfies $1 - q_n = o(n^{-1/2})$ that is needed for asymptotic normality around θ_0. In all considered cases, numerical solution of Eq. (3.2.6) is found using variable metric algorithm (e.g., Goldfarb (1970)), where the ML solution is chosen as the starting value.

3.5.2 Mean squared error: role of the distortion parameter q

In the first group of simulations, we compare the estimators of the true tail probability $\alpha = \alpha(x; \lambda_0)$, obtained via the ML$q$ method and the traditional maximum likelihood approach. Particularly, we are interested in assessing the relative performance of the two estimators for different choices of the sample size by taking the ratio between the two mean squared errors, $MSE\left(\hat{\alpha}_n\right)/MSE\left(\widetilde{\alpha}_n\right)$. The simulations are structured as follows: (i) For any given sample size $n \geq 2$, a number $B = 10000$ of Monte Carlo samples $X_1, ..., X_n$ is generated from an exponential distribution with parameter $\lambda_0 = 1$. (ii) For each sample, the MLq and ML estimates of α, respectively $\widetilde{\alpha}_{n,k} = \alpha(x; \widetilde{\lambda}_{n,k})$ and $\hat{\alpha}_{n,k} = \alpha(x; \widehat{\lambda}_{n,k})$, $k = 1, ..., B$, are obtained. (iii) For each sample size n, the relative performance between the two estimators is evaluated by the ratio $\widehat{R}_n = MSE_{MC}\left(\hat{\alpha}_n\right)/MSE_{MC}\left(\widetilde{\alpha}_n\right)$, where MSE_{MC} denotes the Monte Carlo estimates of the mean squared error. In addition, let $\overline{y}_1 = B^{-1}\sum_{k=1}^{B}(\widehat{\alpha}_{n,k} - \alpha)^2$ and $\overline{y}_2 = B^{-1}\sum_{k=1}^{B}(\widetilde{\alpha}_{n,k} - \alpha)^2$. By the central

limit theorem, for large values of B, $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2)'$ approximately has a bi-variate normal distribution with mean $(MSE(\hat{\alpha}_n), MSE(\widetilde{\alpha}_n))'$ and certain covariance matrix Γ. Thus, the standard error for $\widehat{R}_n$ can be computed by the Delta Method (Ferguson, 1996) as

$$se\left(\widehat{R}_n\right) = B^{1/2}\left(\frac{\widehat{\gamma}_{11}}{\overline{y}_2} - 2\widehat{\gamma}_{12}\frac{\overline{y}_1}{\overline{y}_2^3} + \widehat{\gamma}_{22}\frac{\overline{y}_1^2}{\overline{y}_2^4}\right)^{1/2}$$

where $\widehat{\gamma}_{11}$, $\widehat{\gamma}_{22}$ and $\widehat{\gamma}_{12}$ denote respectively the Monte Carlo estimates for the components of the covariance matrix Γ.

Case 1: fixed α and q. Fig. 3.1 illustrates the behavior of $\widehat{R}_n$ for several choices of the sample size. In general, we observe that for relatively small sample sizes, $\widehat{R}_n > 1$ and the MLqE clearly outperforms the traditional MLE. Such a behavior is much more accentuated for smaller values of the tail probability to be estimated. In contrast, when the sample size is larger, the bias component plays an increasingly relevant role and eventually we observe that $\widehat{R}_n < 1$. This case is presented in Fig. 3.1/(a) for values of the true tail probability $\alpha = .01, .005, .003$ and a fixed distortion parameter $q = 0.5$. Moreover, the results presented in Fig. 3.1/(b) show that smaller values of the distortion parameter q accentuate the benefits attainable in a small sample situation.

Case 2: fixed α and $q_n \nearrow 1$. In the second experimental setting, illustrated in Fig. 3.2/(a), the tail probability α is fixed, while we let q_n be a sequence such that $q_n \nearrow 1$ and $0 < q_n < 1$. For illustrative purposes we choose the sequence $q_n = \left[1/2 + e^{0.3(n-20)}\right] / \left[1 + e^{0.3(n-20)}\right]$, $n \geq 2$ and study R_n for different choices of the true tail probability to be estimated. For small values of the sample size, the chosen sequence q_n converges relatively slowly to 1 and the distortion parameter produces benefits in terms of variance. In contrast, when the sample size becomes larger, q_n adjusts quickly to one. As a consequence, for large samples the MLqE exhibits the same behavior shown by the traditional MLE.

Case 3: $\alpha_n \searrow 0$ and $q_n \nearrow 1$. The last experimental setting of this subsection examines the case where both the true tail probability and the distortion parameter change depending on the sample size. We consider sequences of distortion parameters converging slowly relative to the sequence of quantiles x_n. In particu-

lar we set $q_n = 1 - [10\log(n+10)]^{-1}$ and $x_n = n^{\frac{1}{2+\delta}}$. In the simulation described in Fig. 3.2/(b) we illustrate the behavior of the estimator for $\delta = 0.5, 1.0$ and 1.5, confirming the theoretical findings discussed in section 3.4.

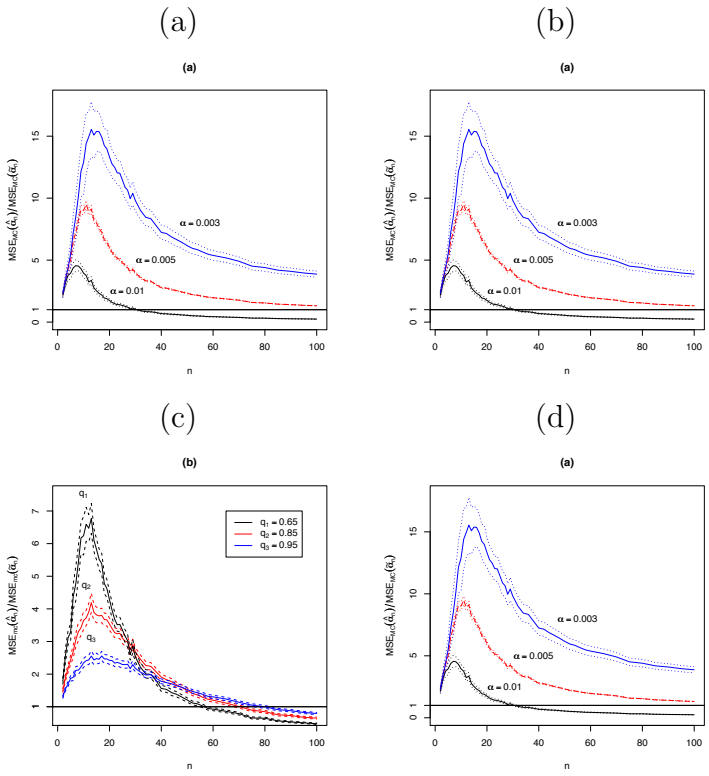
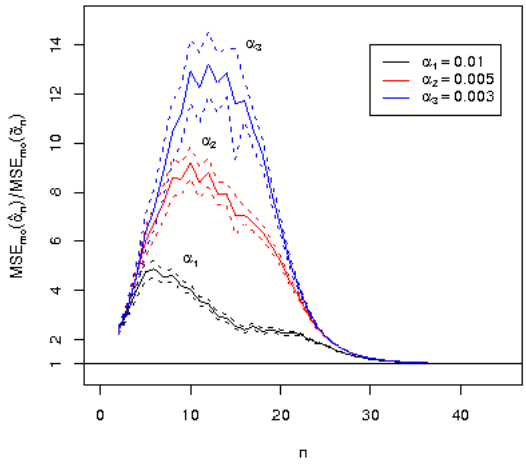


Figure 3.1: Monte Carlo Mean Squared Error ratio computed from $B = 10000$ samples of size n. In (a) we use a fixed distortion parameter $q = 0.5$ and true tail probability $\alpha = 0.01, 0.005, 0.003$. The dashed lines represent 99% confidence bands. In (b) we set $\alpha = 0.003$ and the distortion parameters $q = 0.65, 0.85, 0.95$. The dashed lines represent 90% confidence bands.

3.5.3 Asymptotic and bootstrap confidence intervals

The main objective of the simulations presented in this sub-section is twofold: (a) to study the reliability of MLqE based confidence intervals constructed using three commonly used methods: asymptotic normality, parametric and nonparametric bootstraps; (b) to compare the results with those obtained using MLE. The struc-

(a)



(b)

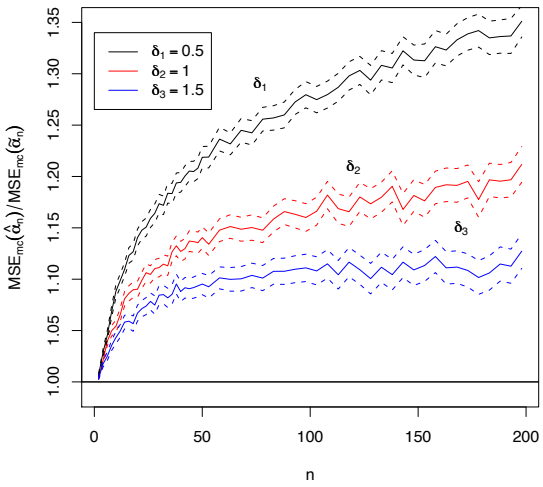


Figure 3.2: (a)Monte Carlo Mean Squared Error ratio computed from $B = 10000$ samples of size n, for different values of the true probability ($\alpha = .01, .005, .003$). The distortion parameter is computed as $q_n = \left[1/2 + e^{0.3(n-20)}\right] / \left[1 + e^{0.3(n-20)}\right]$. (b) Monte Carlo Mean Squared Error ratio computed from $B = 10000$ samples of size n. We use sequences $q_n = 1 - [10\log(n+10)]^{-1}$ and $x_n = n^{\frac{1}{2+\delta}}$ ($\delta = 0.5, 1.0$ and 1.5). The dashed lines represent 99% confidence bands.

ture of simulations is similar to that of section 3.5.2, but a data-driven choice of q_n is used: (ii) For each sample, first we compute $\widehat{\lambda}_n$, the MLE of λ_0; we substitute $\widehat{\lambda}_n$ in Eq. (3.4.9) and solve it numerically in order to obtain q^* as described in section 3.4.4. (iii) For each sample, the MLq and ML estimates of the tail probability α are obtained. The standard errors of the estimates are computed using three different methods: the asymptotic formula derived in (3.4.5), nonparametric and parametric bootstrap. The number of replicates employed in bootstrap re-sampling is 500. We then construct 95% confidence intervals accordingly and check the coverage of the true value α.

In Table 3.1 we show the Monte Carlo means of $\widehat{\alpha}_n$ and $\widetilde{\alpha}_n$, their standard deviations and the standard errors computed with the three methods described above. In addition, we report the Monte Carlo average of the estimates of optimal distortion parameter q^*. When $q^* = 1$, the results refer to the MLE case. Not surprisingly, q^* approaches 1 as the sample size increases. When the sample size is small, the MLqE has a smaller standard deviation and better performance. When n is larger, the advantage of MLqE diminishes. As far as the standard errors are concerned, the asymptotic method and the parametric bootstrap seem to provide values somewhat closer to the Monte Carlo standard deviation for the considered sample sizes.

In Table 3.2 we compare the accuracy of 95% confidence intervals and report the relative length of the intervals for MLqE over those for MLE. Although the coverage probability for MLqE is slightly smaller than that of MLE (in the order of 1%), we observe a substantial reduction in the interval length for all the considered cases. The most evident benefits occur when the sample size is small. Furthermore, note that, in general, the intervals computed via parametric bootstrap outperform the other two methods in terms of coverage and length.

3.5.4 Multivariate normal distribution

In this subsection, we evaluate the MLq methodology for estimating the mean and covariance of a multivariate normal distribution. We generate $B = 10000$ samples from a multivariate normal $N_p(\mu, \Sigma)$, where μ is the p-dimensional unknown mean

n	q^*	Estimate	St.Dev.	se_{asy}	se_{boot}	se_{pboot}
15	.939	.009489	.010975	.010472	.011923	.010241
	1.000	.013464	.014830	.013313	.013672	.015090
25	.959	.009693	.008417	.008470	.009134	.008298
	1.000	.012108	.010517	.009919	.010227	.010950
50	.977	.010108	.006261	.006326	.006575	.006249
	1.000	.011385	.007354	.006894	.007083	.007318
100	.988	.010158	.004480	.004568	.004680	.004549
	1.000	.010789	.004908	.004778	.004880	.004943
500	.998	.010006	.002014	.002052	.002061	.002050
	1.000	.010122	.002055	.002070	.002073	.002087

Table 3.1: MC means and standard deviations of estimators of α, along with the MC mean of the standard error computed using: (i) asymptotic normality, (ii) bootstrap and (iii) parametric bootstrap. The true tail probability is $\alpha = .01$ and $q = 1$ corresponds to the MLE.

vector and Σ is the unknown $(p \times p)$ covariance matrix. In our simulation, the true mean is $\mu = 0$ and the ij-th element of Σ is $\rho^{|i-j|}$, where $-1 < \rho < 1$. To gauge performance for the mean we employed the usual L_2-norm. For the covariance matrix, we considered the loss function

$$\Delta(\Sigma, \widehat{\Sigma}_q) = \mathrm{tr}(\Sigma^{-1}\widehat{\Sigma}_q - \mathbf{I})^2, \tag{3.5.1}$$

where $\widehat{\Sigma}_q$ represents the MLq estimate of Σ with $q = 1 - 1/n$. Note that the loss is 0 when $\Sigma = \widehat{\Sigma}_q$ and is positive otherwise. Moreover, the loss is invariant to the transformations $A\Sigma A^{\intercal}$ and $A\widehat{\Sigma}_q A^{\intercal}$ for a nonsingular matrix A. The use of such a loss function is common in literature (e.g., Huang et al. (2006)).

In Table 3.3, we show simulation results for moderate or small sample sizes ranging from 10 to 100 for various dimensions of the covariance matrix Σ. The entries

		Asympt.		Boot.		Par.Boot.	
n	q^*	Coverage(%)	RL	Coverage(%)	RL	Coverage(%)	RL
15	.939	79.2	0.787	89.1	0.865	92.9	0.657
	1.000	80.9		88.4		92.5	
25	.958	83.4	0.854	91.8	0.890	93.6	0.733
	1.000	84.3		90.8		94.2	
50	.977	87.1	0.918	92.3	0.928	93.9	0.824
	1.000	88.4		91.6		93.4	
100	.988	91.1	0.956	93.3	0.960	94.7	0.889
	1.000	92.2		92.9		94.3	
500	.998	94.5	0.991	95.0	0.995	95.2	0.962
	1.000	94.7		94.6		94.8	

Table 3.2: MC coverage rate of 95% confidence intervals for α, computed using (i) asymptotic normality, (ii) boostrap and (iii) parametric bootstrap. RL is the length of the intervals of MLqE over that of MLE. The true tail probability is $\alpha = .01$ and $q = 1$ corresponds to the MLE.

in the table represent the Monte Carlo estimates of the ratio $\Delta(\Sigma, \widehat{\Sigma}_1)/\Delta(\Sigma, \widehat{\Sigma}_q)$, where $\widehat{\Sigma}_1$ is the usual ML estimate multiplied by the correction factor $n/(n-1)$. The standard error of the ratio is computed via the Delta method. Clearly, the MLqE performs well for smaller sample sizes. Interestingly, the squared error for the MLqE reduces dramatically compared to that of the MLE as the dimension increases. Remarkably, when $p = 8$ the gain in accuracy persists even for larger sample sizes, ranging from about 22% to 84%. We tried various structures of Σ and obtained performances comparable to the ones presented. For μ we found that MLqE performs nearly identically to MLE for all choices of p and n, which is not surprising given the findings in Section 3.2.3. For brevity we omit the results on μ.

n	$p=$	1	2	4	8
10		1.225(0.018)	1.298(0.019)	1.740(0.029)	1.804(0.022)
15		1.147(0.014)	1.249(0.017)	1.506(0.021)	1.840(0.026)
25		1.083(0.011)	1.153(0.012)	1.313(0.016)	1.562(0.020)
50		1.041(0.007)	1.052(0.007)	1.199(0.011)	1.377(0.015)
100		1.018(0.005)	1.033(0.005)	1.051(0.006)	1.222(0.011)

Table 3.3: Monte Carlo mean of $\Delta(\Sigma, \widehat{\Sigma}_1)$ over that of $\Delta(\Sigma, \widehat{\Sigma}_q)$ with standard error in parenthesis.

3.5.5 Generalized linear models

Our methodology can be promptly extended to the popular framework of the generalized linear models. Consider the regression setting where each outcome of the dependent variables, Y, is drawn from a distribution in the exponential family. The mean η of the distribution is assumed to depend on the independent variables, X, through $E(Y|X) = \eta = g^{-1}(X^{\mathsf{T}}\beta)$, where X is the design matrix, β is a p-dimensional vector of unknown parameters and g is the link function. In our simulations, we consider two notable instances: (i) Y from an exponential distribution with link function $\exp(-x^{\mathsf{T}}\beta) = \eta$; (ii) Y from a Bernoulli distribution with link function $x^{\mathsf{T}}\beta = \log(\eta/(1-\eta))$. The first case is useful as it represents the exponential regression model, which is a basic setup for time-to-event analysis. The latter is the popular logistic regression model.

We initialize the simulations by generating design points randomly drawn from the unit hypercube $[-1, 1]^p$. The entries of the true vector of coefficients β are assigned by sampling p points at random in the interval $[-1, 1]$, obtaining values $\beta = (-.57, 0.94, 0.16, -.72, 0.68, .92, .80, .04, .64, .34, .38, .47)$. The values of X and β are kept fixed during the simulations. Then, 1000 Monte Carlo samples of $Y|X$ are generated according to the two models described above and for each sample MLq and ML estimates are computed. The prediction error based on independent out-of-sample observations is:

$$PE_q = \frac{1}{10^3} \sum_{j=1}^{10^3} \left(Y_j^{\text{test}} - \eta(X_j^{\text{test}} \widehat{\beta}_{q,n}) \right)^2, \tag{3.5.2}$$

where $\widehat{\beta}_{q,n}$ is the MLqE of β. In Table 3.4 we present the prediction error for various choices of n and p. For both models, the MLqE outperforms the classic MLE for all considered cases. The benefits from MLqE can be remarkable also when the dimension of the parameter space is larger. This is particularly evident in the case of the exponential regression, where the prediction error of MLE at least twice that of MLqE. In one case, when $n = 25$ and $p = 12$, the MLqE is about nine times more accurate. This is mainly due to MLqE's stabilization of the variance component, which for the MLE tends to be large quickly when n is very small compared to p. Although for the logistic regression we observe a similar behavior, the gain in high dimension becomes more evident for larger n.

p	$n =$	25	50	100	250
			Exp. Regression		
2		2.549(.003)	2.410(.002)	2.500(.003)	2.534(.003)
4		2.469(.002)	2.392(.002)	2.543(.002)	2.493(.002)
8		4.262(.012)	2.941(.004)	3.547(.006)	3.582(.006)
12		9.295(.120)	3.644(.008)	3.322(.005)	5.259(.027)
			Logistic. Regression		
2		1.156(.006)	1.329(.006)	1.205(.003)	1.385(.003)
4		1.484(.022)	1.141(.003)	1.502(.007)	1.353(.003)
8		1.178(.008)	1.132(.003)	1.290(.004)	1.300(.002)
12		1.086(.005)	1.141(.003)	1.227(.003)	1.329(.002)

Table 3.4: Monte Carlo mean of PE_1 over that of PE_q for exponential and logistic regression with standard error in parenthesis.

3.6 Concluding remarks

In this work we have introduced the MLqE, a new parametric estimator inspired by a class of generalized information measures that have been successfully used in several scientific disciplines. The MLqE may also be viewed as a natural extension of the classical MLE: it can preserve the large sample properties of the MLE, while – by means of a distortion parameter q – allowing modification of the trade-off between bias and variance in small or moderate sample situations.

We emphasized both from theoretical and empirical standpoints the meaning of the link between q and the sample size. The Monte Carlo simulations support that when the sample size is small or moderate relative to the tail probability to be estimated, the MLqE successfully trades bias for variance, obtaining an overall reduction of the mean squared error, sometimes very dramatically. Simulations also show that for estimating covariance matrix of multivariate normal distribution and GLM models, the MLq method can substantially improve accuracy. More research on the practical choices of q and their theoretical properties, as well as high order performance comparison between MLqE and MLE, will be valuable.

Proofs of results from Chapter 3

In all of the following proofs we denote $\psi_n(\theta) := n^{-1} \sum_{i=1}^n \nabla_\theta L_{q_n}\left(f(X_i;\theta)\right)$. Since $f(x;\theta) = e^{\theta b(x) - A(\theta)}$, we have

$$\psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n e^{(1-q_n)(\theta b(X_i) - A(\theta))} \left(b(X_i) - D(\theta)\right). \tag{3.6.1}$$

The MLq equation sets $\psi_n(\theta) = 0$ and solve for θ. Define $\varphi(x,\theta) := \theta^{\mathrm{T}} b(x) - A(\theta)$, and thus $f(x;\theta) = e^{\varphi(x,\theta)}$. When clear from the context, $\varphi(x,\theta)$ is denoted by φ.

3.7 Proof of Theorem 3.3.1

Define $\psi(\theta) := E_{\theta_0} \nabla_\theta \log\left(f(X;\theta)\right)$. Since f has the form in (3.3.1), we can write $\psi(\theta) = E_{\theta_0}\left[b(X) - D(\theta)\right]$. We want to show uniform convergence of $\psi_n(\theta)$ to $\psi(\theta)$ for all $\theta \in \Theta$ in probability, i.e.

$$\sup_{\theta\in\Theta} \left|\left| \frac{1}{n} \sum_{i=1}^n \left(e^{(1-q_n)(\theta^{\mathrm{T}} b(X_i) - A(\theta))} - 1 \right) \left(b(X_i) - D(\theta)\right) \right|\right|_1 \xrightarrow{p} 0, \tag{3.7.1}$$

where $||\cdot||_1$ denotes the ℓ_1-norm. Let $s(X_i;\theta) := e^{(1-q_n)(\theta^{\mathrm{T}} b(X_i) - A(\theta))} - 1$ and $t(X_i;\theta) := b(X_i) - D(\theta)$. By the Cauchy-Schwartz inequality, the left hand side of Eq. (3.7.1) can be bounded as follows:

$$|LHS| \quad \leq \sup_{\theta\in\Theta} \left\{ \sum_{j=1}^p \sqrt{\frac{1}{n}\sum_{i=1}^n s(X_i;\theta)^2} \sqrt{\frac{1}{n}\sum_{i=1}^n t_j(X_i;\theta)^2}, \right\}. \tag{3.7.2}$$

where t_j denotes the jth element of the vector $t(X_i;\theta)$. It follows that for (3.7.1), it sufficies to show $n^{-1}\sum_i \sup_\theta s(X_i;\theta)^2 \xrightarrow{p} 0$ and $n^{-1}\sum_i \sup_\theta t_j(X_i;\theta)^2$ is bounded in probability. Since Θ is compact, $\sup_\theta |D(\theta)| \leq (c_1^{(1)}, c_2^{(1)}, \ldots, c_p^{(1)})$ for some positive constants $c_j^{(1)} < \infty$, $j = 1, \ldots, p$ and we have

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in\Theta} t_j(X_i;\theta)^2 \leq \frac{2}{n}\sum_{i=1}^{n} b_j(X_i)^2 + 2\left(c_j^{(1)}\right)^2, \tag{3.7.3}$$

where the last inequality from the basic fact that $(a-b)^2 \leq 2a^2 + 2b^2$ $(a, b \in \mathbb{R})$. The last expression in (3.7.3) is bounded in probability by some constant and $E_{\theta_0} b_j(X)^2 < \infty$ for all $j = 1, \ldots, p$. Next, note that

$$\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n} s(X_i;\theta)^2 \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in\Theta} e^{2(1-q_n)(\theta^{\mathrm{T}}b(X_i)-A(\theta))} - \frac{2}{n}\sum_{i=1}^{n} \inf_{\theta\in\Theta} e^{(1-q_n)(\theta^{\mathrm{T}}b(X_i)-A(\theta))} + 1.$$

Thus, to show $n^{-1}\sum_i \sup_\theta s(X_i;\theta)^2 \xrightarrow{p} 0$, it sufficies to obtain $n^{-1}\sum_i \sup_\theta e^{2(1-q_n)\varphi(\theta)} - 1 \xrightarrow{p} 0$ and $n^{-1}\sum_i \inf_\theta e^{(1-q_n)\varphi(\theta)} - 1 \xrightarrow{p} 0$. Actually, since Θ is compact and $\sup_\theta e^{-A(\theta)} < c^{(2)}$ for some $c^{(2)} < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{\theta\in\Theta} e^{2(1-q_n)(\theta^{\mathrm{T}}b(X_i)-A(\theta))} \leq \frac{1}{n}\sum_{i=1}^{n} e^{2|1-q_n|(|\log c^{(2)}| + {\theta^{(*)}}^{\mathrm{T}}|b(X_i)|)}, \tag{3.7.4}$$

where $\theta_j^{(*)} = \max\left\{|\theta_{j,0}^{(*)}|, |\theta_{j,1}^{(*)}|\right\}$, $j = 1, \ldots, p$, and $(\theta_{j,0}^{(*)}, \theta_{j,1}^{(*)})$ represent elementwise boundary points of θ_j. For $r = 1, 2$

$$E_{\theta_0}\left[e^{2|1-q_n|(|\log c^{(2)}| + {\theta^{(*)}}^{\mathrm{T}}|b(X)|)}\right]^r \tag{3.7.5}$$

$$= e^{2r|1-q_n||\log c^{(2)}| - A(\theta_0)} \int e^{[2r|1-q_n|\mathrm{sign}\{b(x)\}\theta^{(*)} + \theta_0]^{\mathrm{T}} b(x)} d\mu(x). \tag{3.7.6}$$

We decompose Ω into 2^p subsets in terms of the sign of the elements of $b(x)$. That is, $\Omega = \bigcup_{k=1}^{2^p} B_k$, where

$$\begin{aligned} B_1 &= \{x \in \Omega : b_1(x) \geq 0, b_2(x) \geq 0, \ldots, b_{p-1}(x) \geq 0, b_p(x) \geq 0\} \\ B_2 &= \{x \in \Omega : b_1(x) \geq 0, b_2(x) \geq 0, \ldots, b_{p-1}(x) \geq 0, b_p(x) < 0\} \\ B_3 &= \{x \in \Omega : b_1(x) \geq 0, b_2(x) \geq 0, \ldots, b_{p-1}(x) < 0, b_p(x) \geq 0\} \end{aligned} \tag{3.7.7}$$

and so on. Note that $\text{sign}\{b(x)\}$ stays the same for each B_i, $i = 1, \ldots, 2^p$. Also because θ_0 is an interior point, when $|1 - q_n|$ is small enough, the integral in (3.7.6) on B_i is finite and by Dominated Convergence Theorem,

$$\int_{B_k} e^{[2r|1-q_n|\text{sign}\{b(x)\}\theta^{(*)}+\theta_0]^{\mathrm{T}} b(x)} d\mu(x) \overset{n\to\infty}{\to} \int_{B_k} e^{\theta_0^{\mathrm{T}} b(x)} d\mu(x). \tag{3.7.8}$$

Consequently,

$$\int e^{[2r|1-q_n|\text{sign}\{b(x)\}\theta^{(*)}+\theta_0]^{\mathrm{T}} b(x)} d\mu(x) \overset{n\to\infty}{\to} \int e^{\theta_0^{\mathrm{T}} b(x) - A(\theta_0)} d\mu(x) = 1. \tag{3.7.9}$$

It follows that the mean and the variance of $\sup_\theta e^{2(1-q_n)[\theta^{\mathrm{T}} b(X) - A(\theta)]}$ converge to 1 and 0, respectively, as $n \to \infty$. Therefore, a straightforward application of Chebychev's inequality gives

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in \Theta} e^{2(1-q_n)[\theta^{\mathrm{T}} b(X_i) - A(\theta)]} \xrightarrow{p} 1, \quad n \to \infty. \tag{3.7.10}$$

An analogous argument shows that

$$\frac{1}{n} \sum_{i=1}^{n} \inf_{\theta \in \Theta} e^{(1-q_n)[\theta^{\mathrm{T}} b(X_i) - A(\theta)]} \xrightarrow{p} 1, \quad n \to \infty. \tag{3.7.11}$$

Therefore, we have established $n^{-1} \sum_i \sup_\theta s(X_i; \theta)^2 \xrightarrow{p} 0$. Hence, (3.7.1) holds. By applying lemma 5.9, p.46 Van der Vaart (1998), we know that with probability converging to 1, the solution of the MLq equations is unique and it maximizes the

MLqE.

3.8 Proof of Theorem 3.3.2

By Taylor's theorem, there exist a random point $\widetilde{\theta}$, in the line segment between θ_n^* and $\widetilde{\theta}_n$, such that with probability converging to one we have

$$\begin{aligned} 0 &= \psi_n\left(\underline{\mathrm{X}};\widetilde{\theta}_n\right) \\ &= \psi_n\left(\underline{\mathrm{X}};\theta_n^*\right) + \dot{\psi}_n\left(\underline{\mathrm{X}};\theta_n^*\right)\left(\widetilde{\theta}_n - \theta_n^*\right) + \frac{1}{2}(\widetilde{\theta}_n - \theta_n^*)^{\mathtt{T}}\ddot{\psi}_n(\underline{\mathrm{X}};\widetilde{\theta})(\widetilde{\theta}_n - \theta_n^*), \end{aligned} \tag{3.8.1}$$

where $\dot{\psi}_n$ is a $p \times p$ matrix of first-order derivatives and, similarly to van der Vaart Van der Vaart (1998) p.68, $\ddot{\psi}_n$ denotes a p-vector of $(p \times p)$ matrices of second-order derivatives, respectively; $\underline{\mathrm{X}}$ denotes the data vector. We can rewrite the above expression as

$$-\sqrt{n}\ \dot{\psi}(\theta_n^*)^{-1}\psi_n\left(\underline{\mathrm{X}};\theta_n^*\right) = \dot{\psi}(\theta_n^*)^{-1}\dot{\psi}_n\left(\underline{\mathrm{X}};\theta_n^*\right)\sqrt{n}\left(\widetilde{\theta}_n - \theta_n^*\right) \tag{3.8.2}$$

$$+\ \dot{\psi}(\theta_n^*)^{-1}\frac{\sqrt{n}}{2}(\widetilde{\theta}_n - \theta_n^*)\ddot{\psi}_n(\underline{\mathrm{X}};\widetilde{\theta})(\widetilde{\theta}_n - \theta_n^*), \tag{3.8.3}$$

where $\dot{\psi}(\theta) = E_{\theta_0}\nabla_\theta^2 L_{q_n} f(X;\theta)$. Note that,

$$\dot{\psi}(\theta) = E_{\theta_0} e^{(1-q_n)\varphi(\theta)}\left[(1-q_n)\nabla_\theta\varphi(\theta)^{\mathtt{T}}\nabla_\theta\varphi(\theta) - \nabla_{\theta\theta}^2\varphi(\theta)\right] \tag{3.8.4}$$

$$= K_{1,n}E_{\mu_{1,n}}\left[(1-q_n)\nabla_\theta\varphi(\theta)^{\mathtt{T}}\nabla_\theta\varphi(\theta) - \nabla_{\theta\theta}^2\varphi(\theta)\right] \tag{3.8.5}$$

where $\mu_{k,n} = k(1-q_n)\theta + \theta_0$ and $K_{k,n} = e^{A(\mu_{n,k})-A(\theta_0)}$. For $k, l \in \{1, \ldots, p\}$, we have

$$\begin{aligned}
\left\{E_{\mu_{n,1}} \nabla_\theta \varphi(\theta)^{\mathtt{T}} \nabla_\theta \varphi(\theta)\right\}_{kl} =& E_{\mu_{n,1}}\left[\left(b_k(X) - m_k(\theta)\right)\left(b_l(X) - m_l(\theta)\right)\right] && (3.8.6)\\
=& E_{\mu_{n,1}}\left[\left(b_k(X) - m_k(\mu_{n,1}) + m_k(\mu_{n,1}) - m_k(\theta)\right)\right. && (3.8.7)\\
& \left.\times \left(b_l(X) - m_l(\mu_{n,1}) + m_l(\mu_{n,1}) - m_l(\theta)\right)\right] && (3.8.8)\\
=& E_{\mu_{n,1}}\left[\left(b_k(X) - m_k(\mu_{n,1})\right)\left(b_l(X) - m_l(\mu_{n,1})\right)\right] && (3.8.9)\\
& + \left[\left(m_k(\mu_{n,1}) - m_k(\theta)\right)\left(m_l(\mu_{n,1}) - m_l(\theta)\right)\right], && (3.8.10)
\end{aligned}$$

where the first summand is the kl-th element of the covariance matrix $-D(\theta)$ evaluated at $\mu_{n,1}$. Since Θ is compact, $\left\{\dot{\psi}(\theta)\right\}_{kl} \leq C^*_{kl} < \infty$, for some constants C^*_{kl}, $k, l \in \{1, \ldots, p\}$. We take the following steps to derive asymptotic normality.

Step 1. We first show that the left hand side of (3.8.3) converges in distribution. Define the vector $Z_{n,i} := \nabla_\theta L_{q_n} f(X_i, \theta^*_n) - E_{\theta_0} \nabla_\theta L_{q_n} f(X_i, \theta^*_n)$ in $\mathbb{R}^p$. Consider an arbitrary vector $a \in \mathbb{R}^p$ and let $W_{n,i} := a^{\mathtt{T}} Z_{n,i}$ and $\overline{W}_n = n^{-1}\sum_i W_{n,i}$. Since $W_{n,i}$ $(1 \leq i \leq n)$ form a triangular array where $W_{n,i}$ are rowwise i.i.d, we check the Lyapunov condition. In our case, the condition reads

$$n^{-1/3}\left(EW^2_{n,1}\right)^{-1}\left(E\left[W^3_{n,1}\right]\right)^{2/3} \to 0, \quad \text{as} \quad n \to \infty. \tag{3.8.11}$$

Next, denote $\mu_{n,k} = \theta_0 + k(1-q_n)\theta^*_n$. One can see that

$$\left(E\left[W^3_{n,1}\right]\right)^{1/3} = K_n \left(E_{\mu_{n,3}}\left[\sum_{j=1}^{p} a_j \left(b_j(X) - m_j(\theta^*_n)\right)\right]^3 \right)^{2/3}$$

where $K_n = \exp\left\{-\frac{2}{3}A(\theta_0) - 2(1-q_n)A(\theta^*_n) + \frac{2}{3}A(\mu_{n,3})\right\}$ and $K_n \to 1$ as $n \to \infty$. Since θ_0 is an interior point in Θ (compact) the above quantity is uniformly upper bounded in n by some finite constant. Next, consider

$$E\left[W^2_{n,1}\right] = E\left[a^{\mathtt{T}} Z_{n,1} Z^{\mathtt{T}}_{n,1}\right] = a^{\mathtt{T}} E\left[Z_{n,1} Z^{\mathtt{T}}_{n,1}\right] a$$

A calculation similar to that in (3.8.10) for the matrix $Z_{n,1}Z^{\mathtt{T}}_{n,1}$ shows that the

above quantity satisfies

$$a^{\intercal}[-D(\mu_{n,2}) + M_n]a \to -a^{\intercal}D(\theta_0)a > 0,\ n \to \infty, \tag{3.8.12}$$

where the kl-th element of M_n is

$$\{M_n\}_{kl} = \left(m_k(\mu_{n,2}) - m_k(\theta_n^*)\right)\left(m_l(\mu_{n,2}) - m_l(\theta_n^*)\right) \tag{3.8.13}$$

and $\mu_{n,2} \to \theta_0$ and $\theta_n^* \to \theta_0$, as $n \to \infty$. This shows that condition (3.8.11) holds and $\sqrt{n}\ (E\left[W_{n,1}^2\right])^{-1/2} a^{\intercal}\overline{W}_n \xrightarrow{\mathscr{D}} N_1(0,1)$. Hence, by the Cramer-Wold device (e.g. see Van der Vaart (1998)), we have

$$\sqrt{n}\ \left[EZ_{n,1}Z_{n,1}^{\intercal}\right]^{-1/2}\overline{W}_n \xrightarrow{\mathscr{D}} N_p(0, \mathbf{I}_p). \tag{3.8.14}$$

Step 2. Next, we want to convergence in probability of $\dot{\psi}\left(\theta_n^*\right)^{-1}\dot{\psi}_n\left(\underline{\mathrm{X}}, \theta_n^*\right)$ to $\mathbf{I}_p$. For $k, l \in \{1, \ldots, p\}$, Given $\varepsilon > 0$, we have

$$P_{\theta_0}\left(\left|\left\{\dot{\psi}_n\left(\underline{\mathrm{X}}, \theta_n^*\right)\right\}_{kl} - \left\{\dot{\psi}\left(\theta_n^*\right)\right\}_{kl}\right| > \varepsilon\right) \leq n^{-1}\varepsilon^{-2}E_{\theta_0}\left[\frac{\partial^2}{\partial\theta_k\theta_l}L_{q_n}\left(f\left(X;\theta\right)\right)\Big|_{\theta_n^*}\right]^2 \tag{3.8.15}$$

by the i.i.d. assumption and Chebychev's inequality. When $|1 - q_n| \leq 1$, the expectation in (3.8.15) is

$$\begin{aligned}
&E_{\theta_0}\left[e^{2(1-q_n)\varphi(\theta_n^*)}[(1-q_n)(b_k(X) - m_k(\theta_n^*))(b_l(X) - m_l(\theta_n^*)) + D(\theta_n^*)^2]\right]^2 \\
&\leq 2E_{\mu_{n,2}}\left[\left((b_k(X) - m_k(\theta_n^*)(b_l(X) - m(\theta_n^*)\right)^2 + D(\theta_n^*)^4\right] \\
&\times \exp\left\{-A(\theta_0) - 2(1-q_n)A(\theta_n^*) + A(\mu_{n,2})\right\}
\end{aligned}$$

where the inequality passage follows from the triangle inequality and the fact that $q_n \leq 1$. Since Θ is compact and the existence of fourth moments is ensured for exponential families, the above quantity is upper bounded by some finite constant. Therefore, the right hand side of (3.8.15) is upper bounded by a constant that converges to zero as $n \to \infty$. Since convergence in probability holds for each

$k, l \in \{1, \ldots, p\}$ and $p < \infty$, we have that the matrix difference $|\dot{\psi}_n(\underline{\mathrm{X}}, \theta_n^*) - \dot{\psi}(\theta_n^*)|$ converges in probability to the zero matrix. From the calculation carried out in (3.8.4), one can see that $\dot{\psi}(\theta_n^*)$ is a deterministic sequence such that $\dot{\psi}(\theta_n^*) \to \dot{\psi}(\theta_0) = -\nabla_\theta^2 A(\theta_0)$, as $n \to \infty$. Thus, we have

$$|\dot{\psi}_n(\underline{X}; \theta_n^*) - \dot{\psi}(\theta_0)| \leq |\dot{\psi}_n(\underline{X}; \theta_n^*) - \dot{\psi}(\theta_n^*)| + |\dot{\psi}(\theta_n^*) - \dot{\psi}(\theta_0^*)| \xrightarrow{p} 0 \tag{3.8.16}$$

as $n \to \infty$. Therefore, $\dot{\psi}(\theta_n^*)^{-1} \dot{\psi}_n(\underline{\mathrm{X}}, \theta_n^*) \xrightarrow{p} \mathbf{I}_p$.

Step 3. Here, we show that the second on the right hand side of (3.8.3) is neglegible. Let $g(\underline{\mathrm{X}}; \theta)$ be an element of the array $\ddot{\psi}_n(\underline{\mathrm{X}}, \theta)$ of dimension $p \times p \times p$. For some fixed $\overline{\overline{\theta}}$ in the line segment between $\widetilde{\theta}$ and θ_n^*, we have that

$$|g(\underline{\mathrm{X}}; \widetilde{\theta}) - g(\underline{\mathrm{X}}; \theta_n^*)| = |\nabla_\theta g(\underline{\mathrm{X}}, \overline{\overline{\theta}})^{\mathrm{T}}| \, |\widetilde{\theta} - \theta_n^*| \leq \sup_{\theta \in \Theta} |\nabla_\theta g(\underline{\mathrm{X}}, \theta)| \, |\widetilde{\theta} - \theta_n^*|. \tag{3.8.17}$$

A calculation shows that the h-th element of the gradient vector in the expression above is

$$\begin{aligned} \{\nabla_\theta g(\underline{\mathrm{X}}, \theta)\}_h = n^{-1} \sum_{i=1}^{n} e^{(1-q_n)\varphi(\theta)} \big[&(1-q_n)^3 \varphi(\theta)^{(1)} + (1-q_n)^2 \varphi(\theta)^{(2)} \\ &+ (1-q_n)\varphi(\theta)^{(3)} + \varphi(\theta)^{(4)} \big], \end{aligned} \tag{3.8.18}$$

for $h \in \{1, \ldots, p\}$, where $\varphi^{(k)}$ denotes the product of the partial derivatives of order k taken with respect to the parameter θ. As shown before in the proof of Theorem 3.3.1, $\sup_\theta e^{(1-q_n)\varphi(X_i, \theta)}$ has finite expectation when $|1 - q_n|$ is small enough. Thus, by Markov's inequality, $\sup_\theta \, |g'(\underline{\mathrm{X}}, \theta)|$is bounded in probability. In addition, recall that that the deteriministic sequence $\dot{\psi}(\theta_n^*)$ converges to a constant. Hence, $\dot{\psi}(\theta_n^*)^{-1} \ddot{\psi}_n\left(\underline{\mathrm{X}}; \widetilde{\theta}_0\right)$ is bounded in probability.

Since the third term in the expansion (3.8.3) is of higher order than the second term, by combining steps 1, 2 and 3 and applying Slutsky's Lemma we obtain the desired asymptotic normality result. Note that when q_n converges to 1 slowly enough so that $\theta_n^* - \theta_0$ is of a larger order than $n^{-1/2}$, then clearly $\sqrt{n}(\widetilde{\theta}_n - \theta_0)$ does not converge in distribution. Therefore, a necessary and sufficient condition for asymptotic normality of MLqE around θ_0 is $\sqrt{n}(q_n - 1) \to 0$.

3.9 Proof of Theorem 3.4.1

From the second order Taylor expansion of $\alpha(x_n;\widetilde{\theta}_n)$ about θ_n^* one can obtain

$$\begin{aligned}
&\sqrt{n}\frac{\Big(\alpha(x_n;\widetilde{\theta}_n)-\alpha(x_n;\theta_n^*)\Big)}{\sigma_n\alpha'(x_n;\theta_n^*)}\\
&=\sqrt{n}\frac{(\widetilde{\theta}_n-\theta_n^*)}{\sigma_n}+\frac{1}{2\sigma_n}\frac{\alpha''(x_n;\widetilde{\theta})}{\alpha'(x_n;\theta_n^*)}\sqrt{n}(\widetilde{\theta}_n-\theta_n^*)^2\\
&=\sqrt{n}\frac{(\widetilde{\theta}_n-\theta_n^*)}{\sigma_n}+\frac{1}{2\sigma_n}\frac{\alpha''(x_n;\theta_n^*)}{\alpha'(x_n;\theta_n^*)}\frac{\alpha''(x_n;\widetilde{\theta})}{\alpha''(x_n;\theta_n^*)}\sqrt{n}(\widetilde{\theta}_n-\theta_n^*)^2, \qquad (3.9.1)
\end{aligned}$$

where $\widetilde{\theta}$ is a value between $\widetilde{\theta}_n$ and θ_n^*. We need to show that the second term in (3.9.1) converges to zero in probability, i.e.,

$$\frac{\alpha''(x_n;\theta_n^*)}{\alpha'(x_n;\theta_n^*)}\frac{\alpha''(x_n;\widetilde{\theta})}{\alpha''(x_n;\theta_n^*)}\frac{\sigma_n}{\sqrt{n}}\frac{n(\widetilde{\theta}_n-\theta_n^*)^2}{\sigma_n^2}\xrightarrow{p}0. \qquad (3.9.2)$$

Since $\sqrt{n}(\widetilde{\theta}_n-\theta_n^*)/\sigma_n\xrightarrow{\mathscr{D}}N(0,1)$ and σ_n is upper bounded, we need

$$\frac{\alpha''(x_n;\theta_n^*)}{\alpha'(x_n;\theta_n^*)\sqrt{n}}\frac{\alpha''(x_n;\widetilde{\theta})}{\alpha''(x_n;\theta_n^*)}\xrightarrow{p}0. \qquad (3.9.3)$$

This holds under the assumptions of the theorem. This completes the proof of the theorem.

3.10 Proof of Theorem 3.4.2

The rationale presented here is analogous to that of Theorem 3.4.1. From the second order Taylor expansion of $\rho(\widetilde{\theta}_n,s)$ about θ_n^* one can obtain

$$\sqrt{n}\frac{\rho(s_n;\widetilde{\theta}_n)-\rho(s_n;\theta_n^*)}{\sigma_n\rho'(s_n;\theta_n^*)}=\sqrt{n}\frac{(\widetilde{\theta}_n-\theta_n^*)}{\sigma_n}+\frac{1}{2\sigma_n}\frac{\rho''(s_n;\bar{\bar{\theta}})}{\rho'(s_n;\theta_n^*)}\sqrt{n}(\widetilde{\theta}_n-\theta_n^*)^2. \qquad (3.10.1)$$

where $\bar{\bar{\theta}}$ is a value between $\widetilde{\theta}_n$ and θ_n^*. The assumptions combined with Theorem 3.3.2 imply that the second term in Eq.(3.10.1) converges to 0 in probability.

Hence, the central limit theorem follows from Slutsky's Lemma.

3.11 Asymptotic distribution of the MLqE: exponential distribution

First, consider the equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda} \int_0^\infty L_q \left(f(x;\lambda) \right) f(x;\lambda_0) dx \\ &= \lambda^{-q} \int_0^\infty (1-\lambda x) e^{-[\lambda(1-q)+\lambda_0]x} dx \\ &= \lambda^{-q} \lambda_0 \left[\frac{1}{\lambda(1-q)+\lambda_0} - \frac{\lambda}{(\lambda(1-q)+\lambda_0)^2} \right] \end{aligned}$$

and the solution of the the above equation can be easily computed as $\lambda^* = \lambda_0/q$. Next consider,

$$E_{\lambda_0} \left[\frac{\partial}{\partial \lambda} L_q f(x;\lambda) \right]^2 = \lambda_0 \lambda^{-2q} \int_0^\infty (1 - 2\lambda x + \lambda^2 x^2) e^{-[2\lambda(1-q)+\lambda_0]x} dx$$

Computing the integrals and setting $\lambda = \lambda^*$ gives

$$\begin{aligned} E_{\lambda_0} \left[\frac{f'(x;\lambda^*)}{f(x;\lambda^*)^q} \right]^2 &= \lambda_0 \left(\frac{\lambda_0}{q} \right)^{-2q} \left[\frac{1}{\lambda_0(\frac{2}{q}-1)} - \frac{2(\lambda_0/q)}{\lambda_0^2(\frac{2}{q}-1)^2} + \frac{2(\lambda_0/q)^2}{\lambda_0^3(\frac{2}{q}-1)^3} \right] \\ &= q \left(\frac{\lambda_0}{q} \right)^{-2q} \left[\frac{q^2-2q+2}{(2-q)^3} \right]. \end{aligned}$$

Next compute

$$\begin{aligned}
E_{\lambda_0}\left[\frac{\partial}{\partial\lambda}\frac{f'(x;\lambda)}{f(x;\lambda)^q}\right] &= \lambda_0\lambda^{-q}\frac{\partial}{\partial\lambda}\left[\int_0^\infty (1-\lambda x)e^{-[\lambda(1-q)+\lambda_0]x}dx\right] \\
&= \frac{\partial}{\partial\lambda}\left[\frac{\lambda_0\lambda^{-q}}{\lambda(1-q)+\lambda_0} - \frac{\lambda^{1-q}\lambda_0}{[\lambda(1-q)+\lambda_0]^2}\right] \\
&= 2\lambda_0\left(1-q\right)\frac{\lambda^{1-q}}{\left(\lambda_0+\lambda\left(1-q\right)\right)^3} - 2\frac{\lambda_0}{\lambda^q}\frac{1-q}{\left(\lambda_0+\lambda\left(1-q\right)\right)^2} \\
&- q\frac{\lambda_0}{\lambda^{q+1}\left(\lambda_0+\lambda\left(1-q\right)\right)}.
\end{aligned}$$

Substituting for $\lambda = \lambda^*$ in the above expression and reorganizing gives:

$$E_{\lambda_0}\left[\frac{\partial}{\partial\lambda}\frac{f'(x;\lambda)}{f(x;\lambda)^q}\bigg|_{\lambda=\lambda^*}\right] = -\frac{q^3}{\lambda_0\left(\lambda_0/q\right)^q}.$$

Finally, an asymptotic variance is obtained as

$$\sigma^2(\lambda^*) = \frac{E_{\lambda_0}\left[\frac{f'(x;\lambda^*)}{f(x;\lambda^*)^q}\right]^2}{\left(E_{\lambda_0}\,\frac{\partial}{\partial\lambda}\frac{f'(x;\lambda)}{f(x;\lambda)^q}\Big|_{\lambda=\lambda^*}\right)^2} = \left(\frac{\lambda_0}{q_n}\right)^2\left[\frac{q^2-2q+2}{q^3\left(2-q\right)^3}\right].$$

3.12 Asymptotic distribution of the MLqE: multivariate normal

Two matrix operators used in this calculation are the vec$(\cdot)$ (vector) and vech$(\cdot)$ (vector-half). Namely, vec : $\mathbb{R}^{r\times p} \mapsto \mathbb{R}^{rp}$ stacks the columns of the argument matrix. For symmetric matrices, vech : $\mathbb{S}^{p\times p} \mapsto \mathbb{R}^{p(p+1)/2}$ extracts the unique elements (it stacks only the unique part of each column that lies on or below the diagonal) (e.g., see McCulloch (1982)). Furthermore let G be the unique matrix such that vec$D = G$vech$D$, where $D \in \mathbb{S}^{p\times p}$ is a symmetric matrix. Usually, G is called the "extension matrix". The log-likelihood function of a multivariate normal

is

$$\ell(\theta) = \log f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{p}{2}(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathtt{T}}\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu}). \tag{3.12.1}$$

Recall that the surrogate parameter is $\theta^* = (\boldsymbol{\mu}^{\mathtt{T}}, q\ \mathrm{vech}^{\mathtt{T}}\boldsymbol{\Sigma})^{\mathtt{T}}$. The asymptotic variance is computed as $V = J^{-1}(\theta^*)K(\theta^*)J^{-1}(\theta^*)$, where

$$K(\theta^*) = E_{\theta_0}[f(\mathbf{x};\theta^*)^{2(1-q)}U(\mathbf{x};\theta^*)^{\mathtt{T}}U(\mathbf{x};\theta^*)] \tag{3.12.2}$$

$$= c_2 E^{(2)}[U(\mathbf{x};\theta^*)^{\mathtt{T}}U(\mathbf{x};\theta^*)], \tag{3.12.3}$$

and

$$J(\theta^*) = -qE_{\theta_0}[f(\mathbf{x};\theta^*)^{1-q}U(\mathbf{x};\theta^*)^{\mathtt{T}}U(\mathbf{x};\theta^*)] \tag{3.12.4}$$

$$= -qc_1 E^{(1)}[U(\mathbf{x};\theta^*)^{\mathtt{T}}U(\mathbf{x};\theta^*)], \tag{3.12.5}$$

where $E^{(r)}$ denotes expectation taken with respect to a Normal with mean $\boldsymbol{\mu}$ and covariance matrix $[r(1-q)+1]^{-1}\boldsymbol{\Sigma}$, $r = 1, 2$ and the normalizing constant c_r is:

$$c_r := E_{\theta_0}[f(\mathbf{x};\theta^*)^{r(1-q)}] = \frac{\int e^{-\frac{r(1-q)+1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathtt{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}d\mathbf{x}}{(2\pi)^{rp(1-q)/2}|q\ \boldsymbol{\Sigma}|^{r(1-q)/2}(2\pi)^{1/2}|\boldsymbol{\Sigma}|^{1/2}} \tag{3.12.6}$$

$$= \frac{(r(1-q)+1)^{-p/2}}{(2\pi q^p|\boldsymbol{\Sigma}|)^{r(1-q)/2}}. \tag{3.12.7}$$

Note that K and J can be partitioned into block form

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \quad J = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}, \tag{3.12.8}$$

where K_{11} and J_{11} depend on second-order derivatives of U with respect to $\boldsymbol{\mu}$, K_{22} and J_{22} depend on second-order derivatives with respect to vech$\boldsymbol{\Sigma}$. The off-diagonal matrices K_{12}, K_{21} depend on mixed derivatives of U with respect $\boldsymbol{\mu}$ and $\mathrm{vech}^{\mathtt{T}}\boldsymbol{\Sigma}$. Since the mixed moments of order three are zero, one can check that $K_{21} = K_{12}^{\mathtt{T}} = 0$. Consequently, only the calculation of K_{11}, K_{22}, J_{11} and J_{22} is

required and the expression of the asymptotic variance is given by

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix} := \begin{bmatrix} J_{11}^{-1} K_{11} J_{11}^{-1} & 0 \\ 0 & J_{22}^{-1} K_{22} J_{22}^{-1} \end{bmatrix}. \tag{3.12.9}$$

Next, we compute the entries of K and J using the approach employed by McCulloch McCulloch (1982) for the usual log-likelihood function. First, we use standard matrix differentiation to compute K_{11} and J_{11},

$$K_{11} = c_2 E^{(2)} \left[(q\boldsymbol{\Sigma})^{-1} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} (\mathbf{x} - \boldsymbol{\mu}) (q\boldsymbol{\Sigma})^{-1} \right] \tag{3.12.10}$$

$$= c_2 q^{-2} \left[2(1-q) + 1 \right]^{-1} \boldsymbol{\Sigma}^{-1} \tag{3.12.11}$$

and similarly one can obtain $J_{11} = -c_1 q^{-1} \left[(1-q) + 1 \right]^{-1} \boldsymbol{\Sigma}^{-1}$. Some straightforward algebra gives

$$V_{11} = J_{11}^{-1} K_{11} J_{11}^{-1} = \frac{(2-q)^{2+p}}{(3-2q)^{1+\frac{p}{2}}} \boldsymbol{\Sigma}. \tag{3.12.12}$$

Next, we compute V_{22}. Let $\mathbf{z} := \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ using the following relationship derived by McCulloch McCulloch (1982), p.682:

$$\begin{aligned} E[\nabla_{\mathrm{vech}\boldsymbol{\Sigma}} \ell(\theta)]^{\mathrm{T}} [\nabla_{\mathrm{vech}\boldsymbol{\Sigma}} \ell(\theta)] = 1/4 \; G^{\mathrm{T}} (\boldsymbol{\Sigma}^{-1/2} \otimes \boldsymbol{\Sigma}^{-1/2}) (E[(\mathbf{z} \otimes \mathbf{z})(\mathbf{z}^{\mathrm{T}} \otimes \mathbf{z}^{\mathrm{T}})] \\ - \mathrm{vec}\mathbf{I}_p \mathrm{vec}^{\mathrm{T}} \mathbf{I}_p) (\boldsymbol{\Sigma}^{-1/2} \otimes \boldsymbol{\Sigma}^{-1/2}) G. \end{aligned} \tag{3.12.13}$$

Moreover, a result by Magnus and Neudecker Magnus and Neudecker (1979), p 388, shows

$$E[(\mathbf{z} \otimes \mathbf{z})(\mathbf{z}^{\mathrm{T}} \otimes \mathbf{z}^{\mathrm{T}})] = \mathbf{I}_p + K_{p,p} + \mathrm{vec}\mathbf{I}_p \mathrm{vec}^{\mathrm{T}} \mathbf{I}_p \tag{3.12.14}$$

where $K_{p,p}$ denotes the commutation matrix (see Magnus and Neudecker Magnus and Neudecker (1979)). To compute K_{22} and J_{22}, we need to evaluate (3.12.13) at $\theta^* = (\boldsymbol{\mu}^{\mathrm{T}}, q \; \mathrm{vech}^{\mathrm{T}} \boldsymbol{\Sigma})^{\mathrm{T}}$, replacing the expectation operator with $c_r E^{(r)}[\cdot]$. In

particular,

$$\left\{E^{(r)}[(\mathbf{z}\otimes\mathbf{z})(\mathbf{z}^{\mathsf{T}}\otimes\mathbf{z}^{\mathsf{T}})]_{\theta^*} - \mathrm{vec}\mathbf{I}_p\mathrm{vec}^{\mathsf{T}}\mathbf{I}_p\right\} G \tag{3.12.15}$$

$$= (r(1-q)+1)^{-2}\left\{\mathbf{I}_p + K_{p,p}\right\} G \tag{3.12.16}$$

$$= 2(r(1-q)+1)^{-2}G, \tag{3.12.17}$$

where the last equality follows from the fact that $K_{p,p}G = G$. Therefore,

$$K_{22} = 1/(4q^2)c_2\ G^{\mathsf{T}}(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})(E[(\mathbf{z}\otimes\mathbf{z})(\mathbf{z}^{\mathsf{T}}\otimes\mathbf{z}^{\mathsf{T}})] \tag{3.12.18}$$

$$- \mathrm{vec}\mathbf{I}_p\mathrm{vec}^{\mathsf{T}}\mathbf{I}_p)(\boldsymbol{\Sigma}^{-1/2}\otimes\boldsymbol{\Sigma}^{-1/2})G \tag{3.12.19}$$

$$= 1/(4q^2)c_2[(r(1-q)+1)^{-2}+1]G^{\mathsf{T}}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})G \tag{3.12.20}$$

$$= 1/(4q^2)\frac{[(2(1-q)+1)^{-2}+1]\,(3-2q)^{-p/2}}{4\,(2\pi q^p|\boldsymbol{\Sigma}|)^{2-q}}G^{\mathsf{T}}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})G. \tag{3.12.21}$$

A similar calculation gives

$$J_{22} = 1/(4q^2)\frac{[(2-q)^{-2}+1]\,(2-q)^{-p/2}}{(2\pi q^p|\boldsymbol{\Sigma}|)^{(2-q)/2}}G^{\mathsf{T}}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})G. \tag{3.12.22}$$

Finally, we assemble (3.12.21) and (3.12.22) obtaining

$$V_{22} = J_{22}^{-1}K_{22}J_{22}^{-1} = \frac{4q^2[(3-2q)^2+1](2-q)^{4+p}}{[(2-q)^2+1]^2(3-2q)^{2+p/2}}[G^{\mathsf{T}}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})G]^{-1}. \tag{3.12.23}$$

Chapter 4

Information, divergences and likelihood

In parametric and nonparametric density estimation, one popular approach is to minimize some appropriate data-based divergence between an assumed model and the true model underlying the data. The resulting estimators are generally called *minimum divergence estimators*. In this chapter, a survey of some notable representatives of this class of methods is presented. The goal is to overview both the motivating rationale as well as some of their common features and strength. In Section 4.1, we consider the maximum likelihood estimator (MLE) and the weighted likelihood estimator (WLE), emphasizing their relationship with the Kullback-Leibler (KL) divergence. In Section 4.2, minimum disparity procedures are discussed, which include minimization of KL divergence, Hellinger distance and Chi-square divergence as special cases. In Section 4.3, estimators based on minimization of ℓ_2-norm and the minimum density power divergence estimator (MDPDE) are described. The latter includes the KL divergence and the ℓ_2-norm as special instances. A discussion and final remarks are provided in Section 4.4.

4.1 KL divergence and likelihood principle

4.1.1 Akaike's observation

In 1973, Akaike (1973) proposed a new approach to measure the goodness of fit of an estimated statistical model under the name of "an information criterion". His ideas, grounded in the information-theoretical concept of relative entropy, or Kullback-Leibler divergence (Kullback and Leibler, 1951; Kullback, 1959), had heavy consequences on statistical inference and influenced the way of thinking about model selection for the years to come. His paper *New Look at Statistical-Model Identification* (Akaike, 1974) is one of the most cited works among statisticians and practitioners and currently counts more than 5400 citations in the Science Citation Index.

Provided a random sample $X = (X_1, \ldots, X_n)$ from a parametric distribution with density $f(x; \theta_0)$, $\theta_0 \in \Theta \subseteq \mathbb{R}^p$, Akaike pointed out the relationship between the usual inferential task of maximizing the log-likelihood function $\sum_i \log \{f(X_i; \theta)\}$ and the minimization of relative entropy or KL divergence. In many circumstances, the statistician wishes to compare a general target density $f(x; \theta)$ with the "true" density $f(x; \theta_0)$. The comparison can be performed without loss of efficiency using the likelihood ratio $\lambda(x) = f(x; \theta_0)/f(x; \theta)$. Thus, any decision procedure aimed to discriminate between the two distributions, provided that a sample point x is observed, could be equivalently carried out employing $\Phi(\lambda(x))$ for some well-behaved function $\Phi(\cdot)$.

The mean amount of deviation by choosing a particular θ, measured with respect to θ_0 can be conveniently gauged by averaging out the randomness of X for $\Phi(\lambda(x))$, resulting in the following divergence measure:

$$\mathcal{D}(\theta; \theta_0, \Phi) := E_{\theta_0}\left[\Phi\left(\frac{f(x; \theta_0)}{f(x; \theta)}\right)\right]. \tag{4.1.1}$$

When $\Phi(\cdot)$ is convex, such measures belong to the class of f-divergences, previously examined by Csiszár (1967).

Under conditions of interchangeability of integration and differentiation, we

have the following second order Taylor expansion:

$$\mathcal{D}(\theta;\theta_0,\Phi) = \mathcal{D}(\theta_0;\theta_0,\Phi) + \mathcal{D}'(\theta_0;\theta_0,\Phi)(\theta-\theta_0) + \frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\mathcal{D}''(\theta_0;\theta_0,\Phi)(\theta-\theta_0) + o(||\theta-\theta_0||^2). \tag{4.1.2}$$

One can easily check that $\mathcal{D}(\theta_0;\theta_0,\Phi) = \Phi(1)$, $\mathcal{D}'(\theta_0;\theta_0,\Phi) = 0$ and the third term involves $\mathcal{D}''(\theta_0;\theta_0,\Phi) = \Phi''(1)I(\theta)$, where $I(\theta_0)$ is Fisher's information matrix. Moreover, it is possible to restrict further the form of Φ by requiring $\Phi(1) = 0$ and $\Phi''(1) \neq 0$, in order to make $\mathcal{D}(\theta;\theta_0,\Phi)$ sensitive to small variations around θ_0. Under these conditions, $\mathcal{D}(\theta;\theta_0,\Phi)$ behaves similarly to a distance.

Further, consider the effect of taking n independent observations. The divergence is then:

$$\mathcal{D}_n(\theta;\theta_0,\Phi) := \int\cdots\int \Phi\left(\frac{\prod_i f(x_i;\theta_0)}{\prod_i f(x_i;\theta)}\right) dF(x_1\theta_0)\times\cdots dF(x_n\theta_0), \tag{4.1.3}$$

where $F(x_i;\theta_0)$ denotes the cdf of X_i. Akaike concluded that a reasonable choice is $\Phi(\cdot) = \log(\cdot)$ because

$$\mathcal{D}_n(\theta_0;\theta_0,\Phi) = 0 \tag{4.1.4}$$

and

$$\mathcal{D}'_n(\theta_0;\theta_0,\Phi) = 0, \tag{4.1.5}$$

implying that the divergence behaves similarly to a distance also in the case when n observations are considered. Most importantly, the additivity property of the logarithm implies

$$\mathcal{D}''_n(\theta_0;\theta_0,\Phi) = n\mathcal{D}''_1(\theta_0;\theta_0,\Phi). \tag{4.1.6}$$

The last equation is critical as it determines completely the functional form of $\Phi(\cdot)$ as logarithm. It states that the information provided by an independent sample is proportional to the sample size n. Equivalently, the amount of information brought

into play by an additional observation is directly proportional to the total amount of observations available.

These heuristics justify the KL divergence as a "good" measure of discrepancy between the true and the candidate density. In this setting, the KL divergence is

$$\mathcal{D}(\theta;\theta_0,\log) := E_{\theta_0}\left[\log\left(\frac{f(x;\theta_0)}{f(x;\theta)}\right)\right]. \tag{4.1.7}$$

Next, consider an estimate $\widehat{\theta}$ of θ_0. Based on the above considerations, a reasonable risk function is

$$\begin{aligned} R(\widehat{\theta},\theta_0) &= E_{\theta_0}\left[\log\left(\frac{f(X;\theta_0)}{f(X;\widehat{\theta})}\right)\right] \\ &= n^{-1}\sum_{i}^{n}\log\left(\frac{f(X_i;\theta_0)}{f(X_i;\widehat{\theta})}\right) + o_p(\sqrt{n}), \end{aligned} \tag{4.1.8}$$

where the last equality follows by the law of large numbers. Thus, under the model conditions and for large sample sizes, we expect that the minimizer of (4.1.8) is close to the minimum of

$$\sum_{i=1}^{n}\log\left(\frac{f(X_i;\theta_0)}{f(X_i;\widehat{\theta})}\right). \tag{4.1.9}$$

Luckily enough, the minimization of the above expression can be done without knowing the true value of θ_0, due to the properties of the logarithm. This is exactly the same as maximizing the usual log-likelihood function and explains the relationship between KL divergence and maximum likelihood principle.

However, we remark that these considerations are fully valid only for identically distributed X_is. What if the random variables are not identical? The following section describes one way to approach such a situation within the information-theoretical framework provided by Akaike.

4.1.2 Almost identical populations: The weighted likelihood

A famous argument by Stein (1956) made clear that bias can be successfully traded for variance, resulting in an overall gain in terms of precision. Stein pointed out that precision could be "borrowed" from data drawn independently from populations other than the one about which inferences were to be made. Specifically, when the goal is to estimate simultaneously normal population means from independent samples, then the sample averages can be outperformed in terms of expected combined squared-errors of estimation. Surprisingly, each of the improved mean estimators benefits from the data coming from the other populations. This naturally poses the question whether the likelihood principle and the information-theoretical paradigm put forward by Akaike can be extended to describe Stein's phenomenon. Hu and Zidek (2002) tackled this issue by seeking predictive distributions that minimize the relative entropy subject to certain data-dependent constraints. These constraints have the role to capture similarities of the inferential population of interest and others from which independent samples are drawn. This reasoning leads to the weighted likelihood.

Hu and Zidek assume the existence of m distinct populations with densities $f(x_1;\theta_1),\ldots,f(x_m;\theta_m)$. Such distributions are thought to resemble one another in varying degrees and, in some respects, they contain mutual information. Further, it is assumed that observations from the m populations are sampled independently. The inferential interest is on the true joint distribution $f(x;\theta_0) = f(x_1;\theta_{0,1}) \times \cdots \times f(x_m;\theta_{0,m})$. The degree of "resemblance" among the m distributions can be characterized using the KL divergence, or relative entropy and following the general paradigm proposed by Akaike (1973) the goal is then to minimize

$$\int \log\left\{\frac{f(x|\theta_0)}{f(x|\theta)}\right\} dF(x;\theta_0), \tag{4.1.10}$$

with respect to $\theta = (\theta_1,\ldots,\theta_m)$. Since the m populations are independent this is

equivalent to minimize

$$-\int \log\{f(x|\theta)\}dF(x;\theta_0) = -\sum_{j=1}^{m}\int \log\{f(x_j|\theta_j)\}dF(x_j;\theta_{0,j}). \tag{4.1.11}$$

However, it is known that the distributions resemble each other, meaning that they are not too far apart in a KL sense:

$$-\int \log\{f(x_j|\theta_j)\}dF(x_j;\theta_{0,j}) \leq \int \log\{f(x_i|\theta_i)\}dF(x;\theta_{0,j}) \leq c_{ij}, \tag{4.1.12}$$

for all $i \neq j$. In particular $c_{i,j}$ is a constant representing the degree of reciprocal resemblance between density i and density j. Using the Lagrange method one can express the minimization of (4.1.10) as a constrained optimization problem and obtain the new objective function

$$\sum_{j=1}^{m}\lambda_{ij}\int \log\{f(x_i|\theta_i)\}dF(x_j;\theta_{0,j}), \tag{4.1.13}$$

where the constants $\lambda_{i,j}$ depend on a proper choice of $c_{i,j}$. In practice, since the true densities are unknown, we estimate them using the empirical distributions of the data $\widehat{F}_j$. Thus, setting $F(x_j;\theta_{0,j}) = \widehat{F}_j$ gives

$$\sum_{j=1}^{m}\frac{\lambda_{ij}}{n_j}\sum_{i=1}^{n_j}\log\{f(X_{i,j}|\theta_i)\}, \tag{4.1.14}$$

where n_j is number of observations from the j-th population. The above expression is the weighted log-likelihood function and the corresponding minimizer is the weighted likelihood estimator (WLE). The estimator achieves good performance upon a proper choice of the weights. Wang et al. (2004) show both consistency and asymptotic normality of MLE under appropriate assumptions. In these proofs the weights are adaptive as they are allowed to depend on the data.

An important example is given when data are sampled from one main distribution $f(x_1,\theta_1)$, with the exception of a few outlying observations, which are drawn from other populations. In this case the weights $\lambda_{1,j}$ express the degree for which

the associated data are thought to be outliers and the WLE is given by optimizing

$$\sum_{j=1}^{n} \lambda_{1j} \log \{f(X_j|\theta_1)\} = \sum_{j=1}^{n} \log \left\{ f(X_j|\theta_1)^{\lambda_{1j}} \right\}. \tag{4.1.15}$$

The idea of optimizing functions similar to (4.1.14) and (4.1.15) is found in literature in various contexts. For example, the weighted likelihood approach extends the local likelihood method of Hastie and Tibshirani (1987) and 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. The work by Hu and Zidek has the merit of bringing all these approaches to a common ground under the information-theoretical paradigm proposed by Akaike. In practice, however, the advantages of weighted likelihood methods rely heavily on a sensible selection of the weights, which is performed using "ex-post" data-driven procedures such as cross-validation (Wang and Zidek, 2005b).

4.2 Disparities: efficiency or robustness?

Although the traditional MLE is asymptotically efficient, in practice the large sample property is satisfied when two important conditions hold: the assumed model mimics well the distribution of the data and the sample size is sufficiently large. Such conditions are not met in many real-world applications. This motivated the creation of a large body of literature aimed to produce estimators that are not unduly affected by small departures from model assumptions, compared to the MLE. We saw that the WLE is one of such estimators under a proper choice of the weights. In this section, we give an overview of other estimators derived using particular classes of Csiszár's or f-divergence measures: the minimum power disparity estimators (MPDE) and the generalized negative exponential disparity estimators (GNEDE).

4.2.1 Power density disparities

In traditional statistics, there are two important – but somewhat competing – approaches to parametric estimation. Efficiency is the dominant paradigm when the model minimics well the data at hand. Robustness is preferred when (usually few) deviations from the assumed model are present. Lindsay (1994) considered classes of robust methods that have much in common with the Hellinger distance estimators proposed by Beran (1977).

Beran's work pioneered the idea that robustness and efficiency have much in common as they are both intertwined in their relationship with the sample size, under the model choice. Remarkably, his analyses led to a different paradigm for robustness where the degree an observation is actually an outlier depends on both: (i) the sample size and (ii) its probability of occurrence under the specified model. To have an intuition of this, consider three observations $X_1 = X_2 = X_3 = -3$ from a normal distribution with zero mean and unit standard deviation. Although this might be somewhat unusual if the sample size is small, say 10, it is not surprising at all for a sample of size of 1000.

Consider a sample space $\Omega = \{0, 1, \ldots, m\}$, with m possibly infinite and a family of probability densities on Ω and let $f(x;\theta)$, $\theta \in \Theta$ be a parametric target density. Assume that a random sample $X_1, \ldots, X_n$ is drawn from $t(x)$ and let $d(x)$ denote the proportion of the n observations which had value x. Lindsay considered estimating equations for the true density $t(x)$ – not necessarily belonging to the assumed class of parametric models – of the form

$$\sum_x A(\delta(x))\nabla_\theta f(x;\theta) = 0, \tag{4.2.1}$$

where $\delta(x)$ represents the Pearson's residual defined as

$$\delta(x) = \frac{d(x) - f(x;\theta)}{f(x;\theta)} \tag{4.2.2}$$

and $\delta(x) \in [-1, \infty]$. Further, it is required that $A(\cdot)$, called the *residual adjustment function* (RAF), is strictly increasing and twice differentiable on $[-1, \infty]$,

with $A(0) = 0$ and $A'(0) = 1$. The function $A(\delta(x))$ carries relevant information about the trade-off between efficiency and robustness by means of its curvature and determines how strongly the estimator downweights larger outliers. Lindsay demonstrates the superiority of RAFs as opposed to *influence functions* (Huber, 1981a) to describe the robustness behavior of the estimator. For example, this is clear for the case of Hellinger distance estimators, for which the influence function can be unbounded.

One interesting consideration is that solving estimation equations like (4.2.1) corresponds to the minimization of a measure of "distance" between the data and the model $f(x;\theta)$, usually referred to as *disparity*. More precisely, for any pair of densities $f(x;\theta)$ and $d(x)$ a disparity is defined as

$$\rho(d(x), f(x;\theta)) = \sum_x f(x;\theta) G(\delta(x)) \tag{4.2.3}$$

where $G(\cdot)$ is a real-valued function thrice-differentiable on $[-1, \infty)$ with $G(0) = 0$. An important class of such measures is given by the *power divergence disparities*:

$$\rho(d(x), f(x;\theta)) = \sum_x f(x;\theta) \frac{[1+\delta(x)]^{\lambda+1} - 1}{\lambda(\lambda+1)} \tag{4.2.4}$$

considered first by Cressie and Read (1984) in the context of goodness-of-fit testing procedures. Remarkably, three important divergence measures can be recovered as special cases for specific choices of λ: the KL divergence, the Hellinger distance and the Chi-squared divergence.

For the power divergence disparities, the RAFs can be expressed as:

$$A(\delta) = (1+\delta)G'(\delta) - G(\delta) = \lambda[(\delta+1)^{1/\lambda} - 1], \tag{4.2.5}$$

which allows the investigation of the robustness behavior of the estimator to outliers. In Fig. 4.1, we display a plot of $A(\delta)$ for power divergence disparities. The plotted cases correspond to maximum likelihood ($\lambda = 0$), Hellinger distance ($\lambda = -1/2$) and Pearson's Chi-Square divergence ($\lambda = -1$). The maximum likelihood – represented by the straight line – assigns the same weight to all the ob-

servations regardless of the corresponding value of the Pearson's residual. Instead, Hellinger distance and Pearson's Chi-Square depart significantly from linearity, putting lower weight to stronger outliers. This illustrates how the curvature of $A(\delta)$ contains information on the robustness of the corresponding estimator.

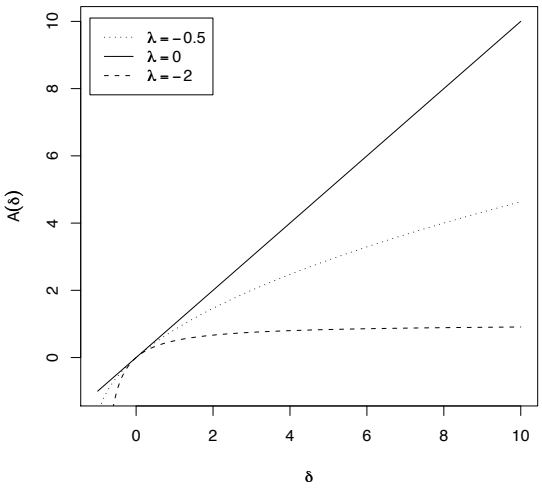


Figure 4.1: Residual adjustment function (RAF) as a function of Pearson's residual δ for power divergence disparities. Cases corresponding to maximum likelihood ($\lambda = 0$), Hellinger distance ($\lambda = -1/2$) and Pearson's Chi-Square divergence ($\lambda = -1$).

4.2.2 Generalized negative exponential disparities

A closely related class of disparities is represented by the family of generalized negative exponential disparities (GNED), defined by taking

$$G(\delta) = \begin{cases} (\exp(-\lambda\delta) - 1 + \lambda\delta)/\lambda^2 & \text{if } \lambda > 0 \\ \delta^2/2 & \text{if } \lambda = 0 \end{cases}. \tag{4.2.6}$$

For the first time, Bhandari et al. (2006) used the GNED for robust estimation, illustrating its competitive performance compared to the Hellinger distance. They

consider the use of such disparities for both parameter estimation and hypothesis testing. In the latter case it is shown that the proposed estimator has better power breakdown than the Hellinger deviance test. The RAF for GNED estimators is given by

$$A(\delta) = \begin{cases} \lambda^{-2}[(\lambda+1) - ((\lambda+1) + \lambda\delta)e^{-\lambda\delta}] & \text{if } \lambda > 0 \\ \delta + \delta^2/2 & \text{if } \lambda = 0 \end{cases}. \tag{4.2.7}$$

In Fig. 4.2, we plotted the RAFs for GNED corresponding to $\lambda = \{0, 0.5, 1, 2, 4\}$. The robustness trait of this type of divergence is somewhat similar to that of power density disparities. However, for properly chosen λ, the RAF is nearly linear in a neighborhood of 0, showing sensitivity to outliers close to that of the MLE. In our plot this can be seen when $\lambda = 0$. At the same time, note that $A(\delta)$ becomes increasingly flat when the Pearson residual is larger, resulting in increased robustness to unusual observations.

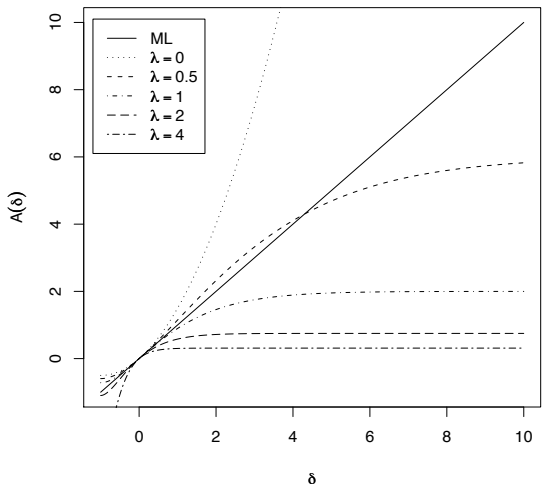


Figure 4.2: Residual adjustment function (RAF) as a function of Pearson's residual δ for generalized negative exponential divergences for $\lambda = \{0, 0.5, 1, 2, 4\}$ in comparison with the maximum likelihood (ML).

4.2.3 Minimum disparity estimation

Consider a random sample $X_1, \ldots, X_n$ from a distribution $t(x)$, possibly continuous in x. When using power divergence disparities or GNED, $d(x)$ is usually estimated using an appropriate nonparametric technique. A common strategy is to consider

$$d_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{4.2.8}$$

where K is a smooth kernel function and h_n is the bandwidth corresponding to the sample size n. For discrete models, $d_n(x)$ is just the proportion of sample values that equal to x for any x in the sample space. In practice, estimation is done by replacing $d(x)$ in the Pearson's residual in Eq.(4.2.1) by $d_n(x)$, giving estimating equations of the form:

$$\sum_{i=1}^{n} A\left(\frac{d_n(X_i) - f(X_i;\theta)}{f(X_i;\theta)}\right) \nabla_\theta f(x;\theta) = 0. \tag{4.2.9}$$

Clearly, any standard nonparametric approach presents a strong drawback concerning all the complications related to the choice of the bandwith h_n. In the next section and in Chapter 5, we will describe estimators with characteristics similar to those discussed in the present section, but with the advantage of avoiding nonparametric analysis.

4.3 Minimum integrated square error

The integrated square error – also referred as to ℓ_2-norm – has enjoyed a long tradition as the goodness-of-fit criterion of choice in nonparametric density estimation. Scott (2001) investigated the use of integrated square error as a theoretical and practical estimation tool for various classes of parametric statistical models. He argues in favor of the robustness virtues of the resulting estimator, especially in high-dimensional problems. He also shows that the inefficiency of this measure compared to the MLE is comparable to that of the median versus the mean. In a similar direction, Basu et al. (1998) propose a generalization that smoothly

bridges between the KL divergence – hence, maximum likelihood estimation – and ℓ_2-norm. In what follows, we will review these two approaches.

4.3.1 Minimum integrated square error estimation

The motivation for parametric ℓ_2-estimation (L2E), originates in the derivation of the nonparametric least squares cross-validation algorithm for choosing the bandwidth h of a histogram (see Bowman (1984)). In a parametric setting, that problem can be formulated as

$$\arg\min_{\theta\in\Theta} \int \left[f(x;\theta) - f(x;\theta_0)\right]^2 dx, \tag{4.3.1}$$

where θ_0 is the unknown true parameter. Given a random sample $X_1, \ldots, X_n$ from $f(x;\theta_0)$, the estimator proposed by Scott is obtained by expanding the square in (4.3.1):

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \int f(x;\theta)^2 dx + \int f(x;\theta_0)^2 dx + 2\int f(X_i;\theta) f(X_i;\theta_0) dx \tag{4.3.2}$$

$$= \arg\min_{\theta\in\Theta} \int f(x;\theta)^2 dx - \frac{2}{n}\sum_{i=1}^{n} f(X_i;\theta), \tag{4.3.3}$$

where in the last line the true density $f(x;\theta_0)$ is estimated by dF_n, the density corresponding to the empirical distribution of the data. This type of approach is extremely interesting as it highlights how an often-used nonparametric criterion can be applied to parametric problems related to robustness. However, the parameters are shown to be relatively inefficient compared to maximum likelihood estimates at the correct model. Actually, for large samples, testing always rejects the models built using this criterion (Scott, 2001).

Finally, we remark that when dealing with high dimensional problems and massive data sets, robustness to deviations from the assumed model can be a critical aspect and the performance of MLE can be really poor. Since the MLE gives the same weight for all the observations, it blurs the distinction between "good" and "bad" data increasingly with the number of dimensions. Instead, the

L2E can successfully handle a substantial fraction of bad data, giving reasonable approximations even in high dimensions.

4.3.2 Minimum Density Power Divergence estimation

Basu et al. (1998) developed a minimum divergence method that includes L2E estimation as a special case. Their approach is appealing as – unlike other robust methods such as minimum Hellinger distance estimation – it avoids the use of nonparametric density estimation and the complications related to the bandwidth selection. They consider the so-called *density power divergences*. In spite of the similarity of the names, these must not be confounded with the previously discussed power divergences by Cressie and Read (1984).

The density power divergence between two densities g and f is defined as

$$\mathcal{D}_\alpha(g, f) = \int \left\{ g^{1+\alpha}(x) - \left(1 + \frac{1}{\alpha}\right) f(x) g(x)^\alpha + \frac{1}{\alpha} f^{1+\alpha}(z) dx \right\}, \tag{4.3.4}$$

where the integral is undefined for $\alpha = 0$. However, if $\alpha \to 0$, $\mathcal{D}_\alpha(g, f)$ is actually the KL divergence. In their paper, Basu and colleagues propose a procedure based on the minimization of an estimate of expression (4.3.4). Note that for $\alpha = 1$, $\mathcal{D}_1(g, f) = \int (g(x) - f(x))^2 dx$ is actually the ℓ_2-divergence and the estimator minimizes the ℓ_2-distance between the two densities f and g. Similarly to Scott's approach, their version of the ℓ_2-distance estimator does not require any smoothing techniques for problems involving continuous densities. Consequently, their estimator creates a smooth bridge between maximum likelihood and robust ℓ_2-distance, allowing for adjusting the trade-off between efficiency and robustness.

Consider random observations $X_1, \ldots, X_n$ drawn from f. Replacing f by the empirical distribution of the data and setting $g := f(x; \theta)$, $\theta \in \Theta$, gives the following estimation equation:

$$\sum_{i=1}^{n} U(X_i; \theta) f(X_i; \theta)^\alpha - \int U(x; \theta) f(x; \theta)^{1+\alpha} dx = 0, \tag{4.3.5}$$

where $U(x; \theta)$ represents the maximum likelihood score function. Note that the

equation is satisfied when $f(x;\theta)$ equals the true density.

Further insights on the behavior of the estimator can be gained by looking at the case where $f(x;\theta)$ is a location model. In that case, the estimating equations can be written as

$$\sum_{i=1}^{n} U(X_i;\theta) f(X_i;\theta)^{\alpha} = 0. \tag{4.3.6}$$

From the expression above it is clear that the MDPDE downweights observations that are inconsistent with the model when $\alpha > 0$, which explains the robust character of this method. However, the amount of robustness can be promptly exchanged for efficiency by means of the parameter α, when needed.

Among the appealing features of the MDPDE, there is also a major drawback: there is no universal way to choose the parameter α. Given a particular model, ne way to proceed could be to fix the efficiency loss at a certain low level, like five or ten percent. A related idea is to fix the maximum level of the resulting influence curve at some acceptable threshold. However, all we know is that α controls the amount of efficiency that is traded for robustness, but no other meaning is associated with it. Hence, one is forced to go back to the initial problem and introduce ex-post data-driven techniques to guarantee a sensible choice of α.

4.4 Discussion

In this chapter, we surveyed various parametric estimation methods. All these methods share an important characteristic: they are derived by minimizing some reasonably chosen divergence between the assumed model and the true distribution underlying the data. The most popular archetype among them is the MLE, understood as the minimization of the empirical version of the KL divergence. Despite the fact that the MLE guarantees desirable asymptotic properties at the model specifications – including efficiency – the needed requirements hardly ever take place in real-world applications. This has motivated alternative methods in the past few decades.

The weighted likelihood methodology for example, which includes notable spe-

cial cases such as the local likelihood by Hastie and Tibshirani (1987), is simply obtained by KL minimization based on data-dependent constraints. When the weights are properly chosen, the WLE trades successfully bias for variance resulting in an overall gain in terms of mean squared error. Under particular choices of the weights, the WLE can be used for robust estimation, downweighting the observations that are inconsistent with the assumed model. However, when the weights cannot be specified before observing the data, an additional and nontrivial layer of complexity adds to the estimation problem. Currently, the choice of weights is handled by using data-driven procedures such as cross-validation.

Another common feature of the discussed methodologies is that the estimating equations always boil down to a weighted version of the log-likelihood score function. In the WLE, the weights are constants depending on the data (Eq. (4.1.14)). When minimizing power disparities, the weights depend on the data through the function $A(\delta)$, where δ is Pearson's residual. In minimum density power divergence estimation, for location models, the weights are written as power transformations of the model under exam (Eq. (4.3.6)). Not surprisingly, remarkable robustness is obtained in these cases as the estimation strategy provides a natural relative-to-model downweighting.

The disparity divergence methods advocated by Lindsay (1994) are shown to provide robust generalizations of the KL divergence and in practice these ideas have the potential to work well, as witnessed by numerous contributions on this literature in recent years. However, the disparity minimization approach suffers from two ponderous drawbacks. The first is some level of nonparametrics required to choose the bandwidth of the kernel smoother. The other concerns the choice of the distortion parameter λ that controls the trade-off between efficiency and robustness. Besides, these two issues are strictly intertwined as both robustness and kernel bandwidth selection are unavoidably bound to the sample size. As a consequence, the meaning of λ as a "controller" of the efficiency is blurred and depends heavily on the bandwidth choice.

The methods proposed by Basu et al. (1998) and Scott (2001) partially overcome these issues because they do not involve kernel smoothing. The MDPDE of Basu et al. creates a smooth bridge between ℓ_2-norm and KL divergence by

means of a parameter α, which does not necessarily depend on the sample size. Are the L2E and MDPDE preferable to the power disparity divergences? In more extreme terms, should we prefer Hellinger distance over ℓ_2-norm? The answer is not univocally determined so far and depends on the problem setup. In general, however, it is known that minimum Hellinger distance estimators are second order efficient and at the same time have robustness qualities similar to ℓ_2-norm.

Chapter 5

Efficient and robust parametric density estimation via q-entropy minimization

In this chapter, we consider parametric density estimation based on minimizing an empirical version of the Havrda-Charvát-Tsallis nonextensive entropy (Havrda and Charvát, 1967; Tsallis, 1988). The resulting estimator, the Maximum Lq-Likelihood Estimator (MLqE), is indexed by a single distortion parameter q. 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 bandwidth selection. The MLqE is studied using asymptotic analysis, simulations and real-world data, showing that it reconciles 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 re-weighting strategy is also described.

5.1 Introduction

In parametric estimation, one approach is to compute the parameters of interest by minimizing some divergence between an appropriate density and a data-based approximation of true model density underlying the data. The successful tradition in this area dates back to Rao (1994) and Kullback and Leibler (1951). Undoubtedly, 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 (1973). Despite the fact that the MLE is asymptotically efficient, in practice the large sample property is satisfactory when two important conditions hold: (i) the assumed class of models can mimic well the actual 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 (1977) first introduced a density minimum divergence estimator based on Hellinger distance. Similar research lines were embraced by Lindsay (1994), Park (2003) and Bhandari et al. (2006). Basu et al. (1998) 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 bandwidth choice, which can be hard to handle in high-dimensional problems.

In a different direction, Hu and Zidek (2002) 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 (1987) and it shares its underlying purpose with other methods such as weighted least squares and kernel smoothers in 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 (Wang

and Zidek, 2005a).

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 (Havrda and Charvát, 1967; Tsallis, 1988). Tsallis and colleagues have successfully employed such measures in the context of statistical mechanics (e.g., see Tsallis (1988)). More recently, applications have appeared in finance, as well as social, biomedical and environmental sciences (Gell-Mann, 2004). 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, the Maximum Lq-Likelihood Estimator (MLqE) has been studied in Chapter 3 in the context of exponential families. When q is judiciously chosen and the sample size is small or moderate, 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 reconcile 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" depends 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 for 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 approach has the perk of not involving any nonparametric analysis,

yet maintaining reasonable performances.

In addition to the appealing properties, our methodology answers three important needs of the practitioner: 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 computing 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 5.2, we introduce a class of quasi-logarithmic divergences and point out their connection with nonextensive entropies. In section 5.3, we introduce the MLqE for parametric families. In section 5.4, convergence in probability and asymptotic normality results of MLqE are provided in light of existing M-estimation theory. In addition, we discuss the trade-off between bias and variance for particular families of distributions. In section 5.5, 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 5.6, we apply the method to real-world examples and assess the finite-sample performance of the MLqE via Monte Carlo simulations.

5.2 Power divergences and nonextensive entropy

Consider a family of models $\mathcal{F}$ having densities (or probability mass functions) $\{g\}$ with respect to a σ-finite measure μ on Ω. Given a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, the large class of Csiszár divergences (Csiszár, 1967) between two densities f and g is given by

$$\int_{\Omega} f(x)\varphi\left\{\frac{f(x)}{g(x)}\right\} d\mu(x). \tag{5.2.1}$$

Commonly, optimization is performed with respect to g and different choices of φ lead to different divergence measures. Perhaps, the most common choice is $\varphi(\cdot) = \log(\cdot)$, which yields the popular KL divergence, or *relative entropy* (Kullback and Leibler, 1951). Consider instead the following family of divergences.

Definition 5.2.1. *Define the power divergence between $g(x)$ and $f(x)$ as*

$$\mathcal{D}_q(g||f) = -\frac{1}{q}\int_\Omega f(x) L_q\left\{\frac{g(x)}{f(x)}\right\} d\mu(x), \tag{5.2.2}$$

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

$$NCS(g||f) = \int_\Omega \frac{[f(x) - g(x)]^2}{2f(x)} d\mu(x); \tag{5.2.3}$$

Hellinger distance:

$$HD(g||f) = \int_\Omega \left[\sqrt{g(x)} - \sqrt{f(x)}\right]^2 d\mu(x); \tag{5.2.4}$$

KL divergence:

$$KL(g||f) = \int_\Omega f(x) \log\left\{\frac{f(x)}{g(x)}\right\} d\mu(x); \tag{5.2.5}$$

Pearson's Chi-squared:

$$PCS(g||f) = \int_\Omega \frac{[f(x) - g(x)]^2}{2g(x)} d\mu(x); \tag{5.2.6}$$

by letting $q = -1, 1/2, 1, 2$, respectively. The same type of divergence has been considered by Cressie and Read (1984) in relation to goodness-of-fit tests and by Lindsay (1994) in the context of robust estimation. Although in general the power

divergence is not a distance as it lacks symmetry, it enjoys the following important discrimination property.

Theorem 5.2.1. *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

$$-\frac{1}{q}\int_\Omega f(x) L_q\left\{\frac{g(x)}{f(x)}\right\} d\mu(x) \geq -\frac{1}{q} L_q \int_\Omega f(x)\frac{g(x)}{f(x)} d\mu(x) = 0. \tag{5.2.7}$$

□

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

$$\mathcal{H}_1(g||f) = -\int_\Omega f(x) \log\{g(x)\}\, d\mu(x). \tag{5.2.8}$$

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:

$$L_q(u_1 u_2) = L_q(u_1) + L_q(u_2) + (1-q) L_q(u_1) L_q(u_2),\ u_1, u_2 > 0,\ q > 0 \tag{5.2.9}$$

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

$$-q\mathcal{D}_q(g||f) = \int L_q(g) + L_q(f^{-1}) + (1-q) L_q(g) L_q(f^{-1}) f d\mu \tag{5.2.10}$$

$$= \int \frac{f^{q-1} - 1}{1-q} + \frac{g^{1-q} f^{q-1} - f^{q-1}}{1-q} f d\mu \tag{5.2.11}$$

$$= \int L_q(g) f^q d\mu - \int L_q(f) f^q d\mu. \tag{5.2.12}$$

Since in (5.2.12) we integrate with respect to 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 transformation $f^{(\alpha)}(x) := f(x)^\alpha / \int f(x)^\alpha d\mu(x)$, $\alpha > 0$. Replacing the true density f in Eq.(5.2.12) with the transformation $f^{(1/q)}$ gives

$$\mathcal{D}_q(g||f^{(1/q)}) = Z^{-1}(q) \left(\int L_q(f^{(1/q)}) f d\mu - \int L_q(g) f d\mu \right), \tag{5.2.13}$$

where $Z(q) = q \int f^{1/q} d\mu$. Therefore, minimizing (5.2.13) is equivalent to minimizing the *q-entropy*, or nonextensive entropy, defined as

$$\mathcal{H}_q(g||f) = -\int_\Omega f(x) L_q \left\{ g(x) \right\} d\mu(x). \tag{5.2.14}$$

Given observations $X_1, \ldots, X_n$ from f, our program is to minimize $\mathcal{H}_q(g||f)$. Since f is unknown, we replace the above expectation with one taken with respect to 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 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.

5.3 The Maximum Lq-Likelihood method

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

$$\mathcal{H}_q(\theta||\theta_0) = -\int_\Omega f(x;\theta_0) L_q \left\{ f(x;\theta) \right\} d\mu(x). \tag{5.3.1}$$

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 5.3.1. *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*

$$\widehat{\theta}_{q,n} = \arg\max_{\theta\in\Theta} \ell_q(\theta) := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} L_q\left[f(X_i;\theta)\right], \quad q > 0, \tag{5.3.2}$$

where L_q is the q-logarithmic function defined in (5.2.2).

When $q \to 1$, if the estimator $\widehat{\theta}_{n,1}$ exists, it approaches 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

$$\Psi_n(\theta) := n^{-1}\sum_{i=1}^{n} U(X_i;\theta) f(X_i;\theta)^{1-q} = 0, \tag{5.3.3}$$

where $U(x;\theta) = \nabla_\theta \log\{f(x;\theta)\}$ is the maximum likelihood score function. When $q \neq 1$, Eq.(5.3.3) 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 (1995) and Choi et al. (2000) propose similar strategies to robustify estimators. 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 (5.3.3) equals to equation (2.4) in Basu et al. (1998), when $q = 1 - \alpha$. Our perspective shares features with these approaches and ultimately highlights the role played by nonextensive entropy measures in downweighting with respect to the model rather than the data.

5.4 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 (1981b) and Hampel et al. (2005)). 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:

$$K_q(\theta) := \int_\Omega f(x;\theta)^{2(1-q)} U(x;\theta) U(x;\theta)^\intercal f(x;\theta_0) d\mu(x) \tag{5.4.1}$$

and

$$J_q(\theta) := \int_\Omega f(x;\theta)^{(1-q)} [(1-q) U(x;\theta) U(x;\theta)^\intercal - I(x;\theta)] f(x;\theta_0) . d\mu(x) \tag{5.4.2}$$

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 $J_q(\theta^*)^{-1} K_q(\theta^*) J_q(\theta^*)^{-1}$.

5.4.1 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

$$\Psi(\theta) := \int_\Omega L_q\{f(x;\theta)\} f(x;\theta_0) d\mu(x) < \infty \tag{5.4.3}$$

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 5.4.1. *Assume the following conditions:*
(C.1) For $\theta \in \Theta$, $\psi(x,\theta)$ is continuous almost everywhere;

(C.2) For all sufficiently small balls B, $\sup_{\theta \in B} \{\psi(x,\theta)\}$ *is measurable and*

$$E_{\theta_0} \sup_{\theta \in B} \{\psi(x,\theta)\} < \infty. \tag{5.4.4}$$

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* $K \subset \Theta$*,*

$$P\left(||\widehat{\theta}_{q,n} - \theta^*|| > \varepsilon \wedge \widehat{\theta}_{q,n} \in K\right) \to 0. \tag{5.4.5}$$

Proof. The proof is given in Van der Vaart (1998), Theorem 5.14. □

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

Lemma 5.4.2. *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*

$$|\psi(x;\theta_1) - \psi(x;\theta_2)| \leq \gamma(x)||\theta_1 - \theta_2||, \quad \textit{and} \quad E||\gamma(x)||^2 < \infty. \tag{5.4.6}$$

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 5.4.3. *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

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

$$\nabla_\theta[f(x;\theta)^{(1-q)}U^{\intercal}(x;\theta)] = (1-q)\frac{\nabla_\theta[f(x;\theta)]}{f(x;\theta)^q}U^{\intercal}(x;\theta) + f(x;\theta)^{(1-q)}I(x;\theta) \quad (5.4.7)$$

$$= f(x;\theta)^{(1-q)}[(1-q)U(x;\theta)U^{\intercal}(x;\theta) + I(x;\theta)] \quad (5.4.8)$$

The result follows from the given condition. □

The next theorem states the asymptotic normality of the MLqE.

Theorem 5.4.4. *Let θ^* be an interior point of Θ, and suppose the conditions of Lemma 5.4.3 and Lemma 5.4.3 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*

$$\sqrt{n}(\widehat{\theta}_{q,n} - \theta^*) \xrightarrow{\mathcal{D}} N\left(0, J_q(\theta^*)^{-1}K_q(\theta^*)J_q(\theta^*)^{-1}\right) \quad (5.4.9)$$

where K_q and J_q are given in Eq.(5.4.1) and Eq.(5.4.2).

Proof. By Lemma 5.4.3, we can write the following Taylor expansion of $\psi(\theta)$:

$$\Psi(\theta) = \Psi(\theta^*) + \frac{1}{2}(\theta-\theta^*)^T\nabla_\theta^2\Psi(\theta^*)(\theta-\theta^*) + o(||\theta-\theta^*||^2). \quad (5.4.10)$$

By the Lipschitz condition (5.4.6), the desired result follows immediately from applying Theorem 5.21 in Van der Vaart (1998). □

Note that assumption in Lemma 5.4.3 the interchangeability of integration and differentiation, implicitly implies the conditions for the existence of J_q. Such requirements are

$$E_{\theta_0}\left[i_{kj}(x;\theta)f(x;\theta)^{1-q}\right] < \infty, \quad \text{and} \quad E_{\theta_0}\left[u_{kj}(x;\theta)f(x;\theta)^{1-q}\right] < \infty, \quad (5.4.11)$$

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.

5.4.2 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 (Huber (1981b),Hampel et al. (2005)). 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 sandwich estimator (e.g., see Huber Huber (1981b)). Let $k(x) = f(x;\widehat{\theta}_{q,n})^{1-q}U(x;\widehat{\theta}_{q,n})$. The variance estimator is

$$\widehat{\mathrm{Var}}(\widehat{\theta}_{q,n}) = (n-1)^{-1}\widehat{J}_q^{-1}\sum_{i=1}^{n} k(X_i)k(X_i)^{\intercal}\widehat{J}_q^{-1}, \tag{5.4.12}$$

where $\widehat{J}_q$ is obtained by replacing $\widehat{\theta}_{q,n}$ in the expression of the influence function and taking expectation with respect to 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.

5.4.3 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).

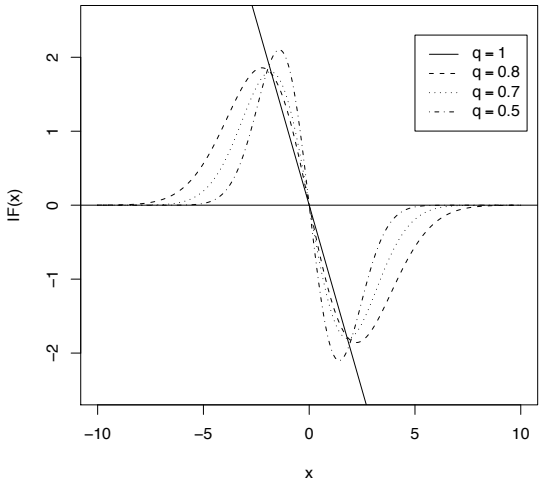
Lemma 5.4.5. *Let $\Psi(\theta)$ be as in Eq.(5.4.3). If $f(x;\theta)$ is an exponential family and the conditions given in Lemma 5.4.3 are satisfied, then $\theta^* = \theta_0/q$ maximizes $\Psi(\theta)$.*

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

$$\nabla_\theta\Psi(\theta) = \int_\Omega \frac{\nabla_\theta[f(x;\theta)]}{f(x;\theta)^q} f(x;\theta_0)d\mu(x). \tag{5.4.13}$$

For exponential families we have that $f(x;\theta_0/q) \propto f(x;\theta_0)^{1/q}$. Thus, by the given

(a)



(b)

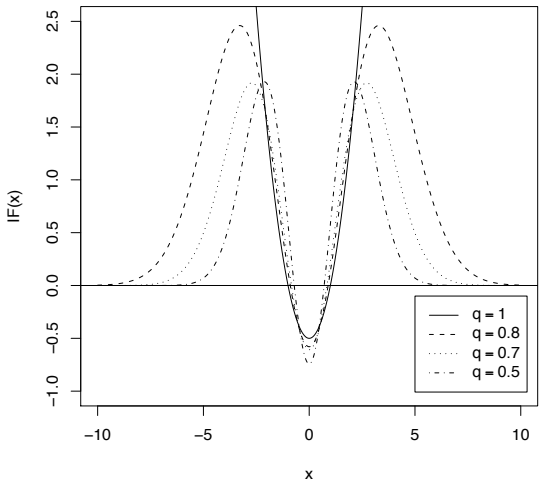


Figure 5.1: Influence functions for estimating the mean (a) and the standard deviation (b) of a standard normal distribution for various choices of q.

conditions, we have $\nabla_\theta \Psi(\theta^*) = \nabla_{\theta^*} \int_\Omega f(x;\theta^*) d\mu(x) = 0$. The second derivative is

$$\nabla^2_\theta \Psi(\theta) = \int_\Omega \frac{\nabla^2_\theta f(x;\theta)}{f(x;\theta)^q} f(x;\theta_0) d\mu(x) - q \int_\Omega \frac{[\nabla_\theta f(x;\theta)]^\intercal [\nabla_\theta f(x;\theta)]}{f(x;\theta)^{q+1}} f(x;\theta_0) d\mu(x). \tag{5.4.14}$$

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. (5.2.13) 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 bandwidth choice. However, simulations for various 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.

5.4.4 Trade-off between bias and variance

5.4.4.1 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 5.4.4 becomes the inverse of the Fisher information. Thus, the ratio of the asymptotic mean squared errors is

computed as

$$\Lambda(q,n;\theta_0) := \frac{aMSE(1,n;\theta_0)}{aMSE(q,n;\theta_0)} = \frac{\mathrm{Tr}\left(J_1(\theta_0)^{-1}K_1(\theta_0)J_1(\theta_0)^{-1}\right)}{n||\theta^* - \theta_0||^2 + \mathrm{Tr}\left(J_q(\theta^*)^{-1}K_q(\theta^*)J_q(\theta^*)^{-1}\right)}, \tag{5.4.15}$$

where $\mathrm{Tr}(\cdot)$ 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 5.4.1. Consider the exponential distribution with density $\theta_0 \exp\{-x\theta_0\}$, $x > 0$, $\theta_0 > 0$. In Chapter 3, we computed J_q and K_q, obtaining

$$\theta^* = \theta_0/q, \quad J_q(\theta^*) = -\frac{q^3}{\theta_0(\theta_0/q)^q}, \quad K_q(\theta^*) = q\left[\frac{q^2-2q+2}{(2-q)^3}\right]\left(\frac{\theta_0}{q}\right)^{-2q}. \tag{5.4.16}$$

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

$$J_q(\theta^*)^{-2}K_q(\theta^*) = \theta_0^2\frac{q^2-2q+2}{q^5(2-q)^3}. \tag{5.4.17}$$

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

$$\Lambda(q,n) = \left[n\left(\frac{1-q}{q}\right)^2 + \frac{q^2-2q+2}{q^5(2-q)^3}\right]^{-1}, \tag{5.4.18}$$

which turns out to be independent from θ_0.

Example 5.4.2. 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 $\theta_0^2(1-\sqrt{q})^2$. A calculation shows that the asymptotic variance is

$$J_q(\theta^*)^{-2}K_q(\theta^*) = \theta_0^2\frac{(3-2q+q^2)}{4(2-q)^{5/2}q^{3/2}}. \tag{5.4.19}$$

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

relative efficiency is

$$\Lambda(q,n) = \left(2n(1-\sqrt{q})^2 + \frac{(3-2q+q^2)}{2(2-q)^{5/2}q^{3/2}}\right)^{-1}, \tag{5.4.20}$$

which, as for the case of the exponential distribution, does not depend on the true value of the parameter.

In Fig.5.2, 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-adjusted efficiency larger than 1.

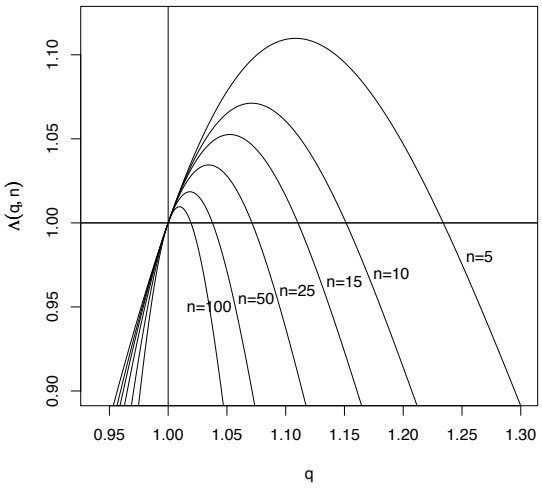
Finite sample efficiency

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. 5.3, where the ratio of the Monte Carlo mean squared errors of the MLE over that of the MLqE is $\widehat{R}(q,n) = \sum_{b=1}^{B}(\widehat{\theta}_{1,n} - \theta_0)^2 / \sum_{b=1}^{B}(\widehat{\theta}_{q,n} - \theta_0)^2$ 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.

5.5 Re-weighting algorithm

One of the main perks of the MLqE is that a simple and fast algorithm is automatically available. Fixed q, Eq. (5.3.3) 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.

(a)



(b)

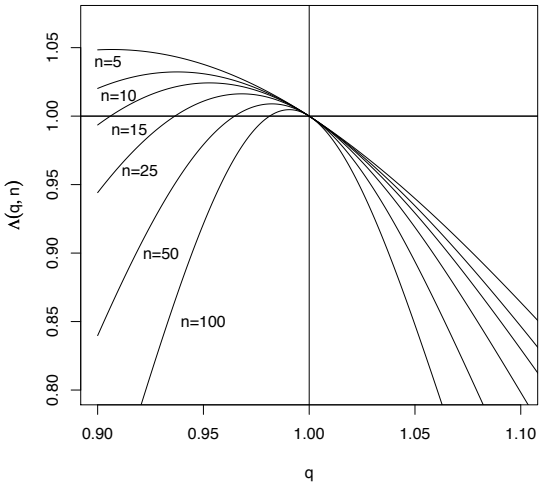
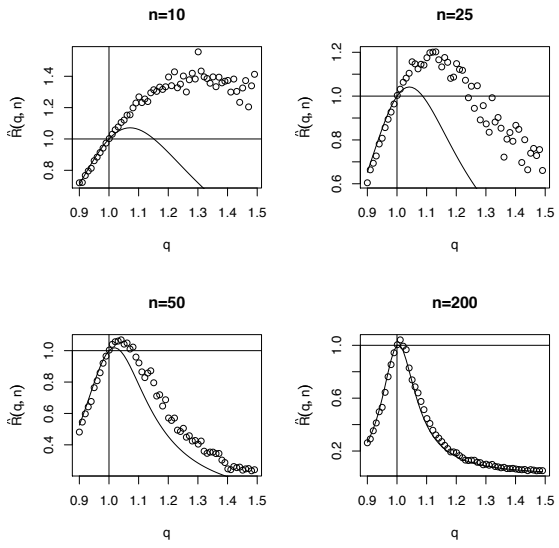


Figure 5.2: Bias-adjusted relative efficiency betwen MLE and MLqE for different sample sizes as in Eq.(5.4.18) and in Eq.(5.4.20), for an exponential (a) and a scale normal (b).

$\mathrm{Exp}(\theta_0)$



$\mathrm{N}(0, \theta_0)$

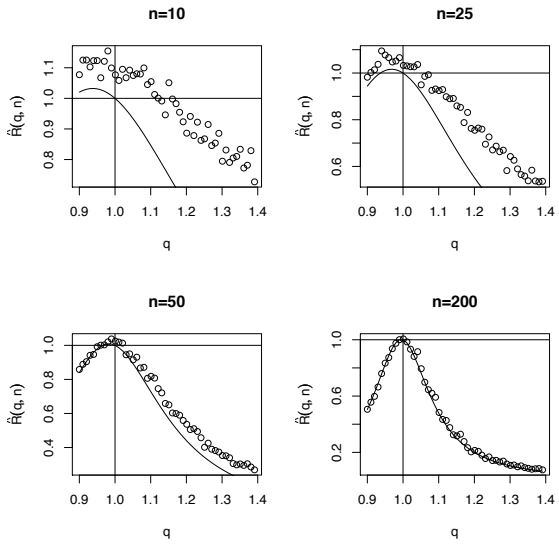


Figure 5.3: Monte Carlo relative efficiency between MLqE and MLE against q for an exponential and a normal for various sample sizes. The solid line is the bias-adjusted relative efficiency as in Eq.(5.4.18) and in Eq. (5.4.20). The Monte Carlo sample size is 1500.

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 $s > 0$,

$$\theta^{(s+1)} = \left\{ \theta : \sum_{i=1}^{n} w^* \left(X_i; \theta^{(s)} \right) U(X_i; \theta) = 0 \right\}, \tag{5.5.1}$$

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 straightforward variable transformation, as illustrated in the following two examples.

Example 5.5.1 (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

$$w_i = \left[\sum_{j=1}^{n} \exp \left\{ -(X_j - X_i) \widehat{\theta}^{(s-1)} (1 - q) \right\} \right]^{-1}. \tag{5.5.2}$$

Example 5.5.2 (Multivariate normal distribution)**.** For computing the unknown mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, proceed as follows:

1. Initialize, by setting $\widehat{\boldsymbol{\mu}}^{(0)} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}}^{(0)} = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^{(0)})^{\mathsf{T}} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^{(0)})$, 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 $\boldsymbol{\Sigma}^{(s)} = \sum_{i=1}^{n} w_i^{(s-1)} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^{\mathsf{T}} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$,

where

$$w_i^{(s)} = \frac{f(\mathbf{x}_i; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})^{1-q}}{\sum_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})^{1-q}} \tag{5.5.3}$$

Finally, if asymptotically unbiased estimates are desired, one can set $\widehat{\boldsymbol{\mu}}^{(s^*)} = \widehat{\boldsymbol{\mu}}^{(s^*-1)}$ and $\widehat{\Sigma}^{(s^*)} = q \widehat{\Sigma}^{(s^*-1)}$, where s^* is the step at which convergence is reached.

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 (1995). First, note that the re-weighting 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

$$\begin{aligned}\nabla_\tau h(\tau) = (q-1)\left\{E_{F_n}\left[f(X;\tau)^{1-q}I(X;\tau)\right]\right\}^{-1} \\ \times E_{F_n}\left[f(X;\tau)^{1-q}U(X;\tau)^{\intercal}U(X;h(\tau))\right].\end{aligned} \tag{5.5.4}$$

The above derivative can be restated as

$$\nabla_\tau h(\tau) = (1-q)W(\tau)\left[\mathbf{I} + (1-q)W(\tau)\right]^{-1}, \tag{5.5.5}$$

where

$$\begin{aligned}W(t) = -\left\{E_{F_n}\nabla_\tau\left[f(X;\tau)^{1-q}U(X;h(\tau))\right]\right\}^{-1} \\ \times E_{F_n}\left[f(X;\tau)^{1-q}U(X;\tau)^{\intercal}U(X;h(\tau))\right].\end{aligned} \tag{5.5.6}$$

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 $W(\theta_0) = q^{-1}\mathbf{I}$. By substituting in Eq.(5.5.5)m we obtain a diagonal matrix with diagonal elements equal to $(1-q)$. The local convergence rate is related to the largest eigenvalue of (5.5.5) at the solution (Johnson and Riess, 1982). 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 the distortion parameter is to 1, the faster the algorithm.

Figure 5.5 illustrates the convergence rates of the re-weighting algorithm for q ranging from 0.5 to 1.5. The dotted lines correspond to 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 $|1-q|$.

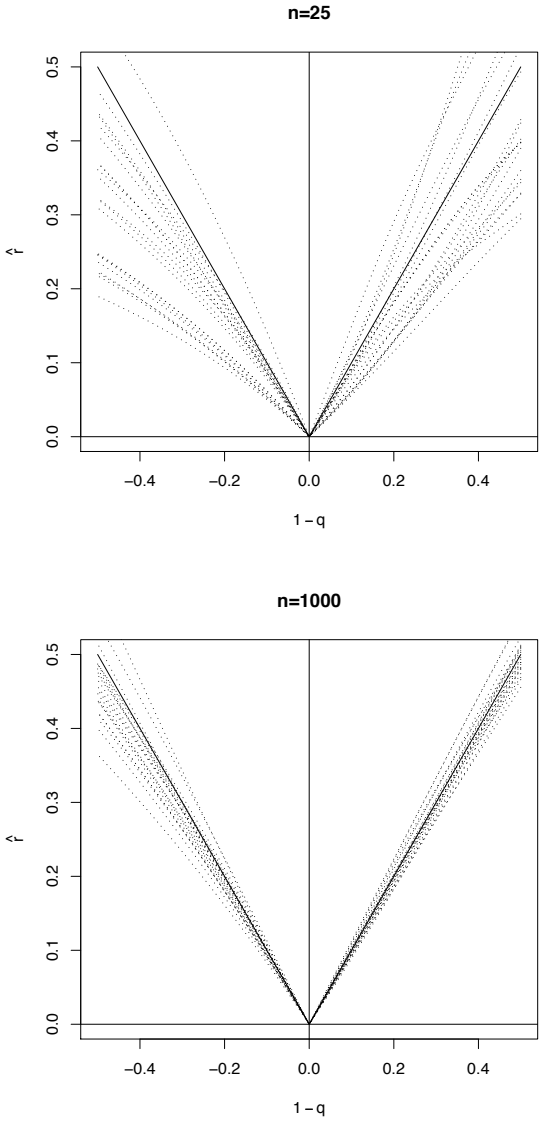


Figure 5.4: The dotted lines are the estimated convergence rates of the algorithm for 20 samples of size 25 (left panel) and 1000 (right panel) from an Exp(1). The solid line corresponds to $\widehat{r} = |1 - q|$.

5.5.1 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 alter 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 when data are sampled from a correctly specified model. In this situation the MLqE can improve upon the MLE by reducing the variance at expenses of a slightly increased bias, when the sample size is moderate or small. Under these circumstances, a reasonable 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: Automatic choice of q via Windham's criterion. Unlike other optimization methods the convergence rate of the re-weighting algorithm described in section 5.5 yields a statistical interpretation. Evaluate expression (5.5.6) at $\widehat{\theta}_{q,n}$, obtaining

$$W(\widehat{\theta}_{q,n})^2 = \widehat{J}_q(\widehat{\theta}_{q,n})^{-1} E_{F_n}\left[f(X;\widehat{\theta}_{q,n})^{1-q} U(X;\tau)^{\intercal} U(X;\widehat{\theta}_{q,n})\right]^2 \widehat{J}_q(\widehat{\theta}_{q,n})^{-1}, \quad (5.5.7)$$

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

$$\left(E_{F_n}\left[f(X;\widehat{\theta}_{q,n})^{1-q} U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right)^2 \quad (5.5.8)$$

$$\leq \left(E_{F_n}\left[f(X;\widehat{\theta}_{q,n}))^{2(1-q)} U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right) \quad (5.5.9)$$

$$\times \left(E_{F_n}\left[U(X;\widehat{\theta}_{q,n})^{\intercal} U(X;\widehat{\theta}_{q,n})\right]\right), \quad (5.5.10)$$

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

$$W(\widehat{\theta}_{q,n})^{-2} \geq \widehat{I}_q(\widehat{\theta}_{q,n}) \left[\widehat{J}_q(\widehat{\theta}_{q,n}) \widehat{K}_q(\widehat{\theta}_{q,n}) \widehat{J}_q(\widehat{\theta}_{q,n}) \right] \tag{5.5.11}$$

and $W(\widehat{\theta}_{q,n})^{-2}$ is an empirical upper bound for efficiency. The above calculation emphasizes that the convergence rate of the algorithm contains information about the efficiency of the estimates through equation (5.5.5). In a similar situation, Windham (1995) considered equating diagonal elements of (5.5.5), 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 distortion parameters in a grid $Q_k = \{q_1, \ldots, q_k\}$, corresponding convergence rates can be computed as

$$\widehat{r}_{q_j} = \frac{||\widehat{\theta}_{q_j}^{(S)} - \widehat{\theta}_{q_j}^{(S-1)}||}{||\widehat{\theta}_{q_j}^{(S-1)} - \widehat{\theta}_{q_j}^{(S-2)}||}, \; 1 \leq j \leq k, \tag{5.5.12}$$

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

$$\widehat{q} = \underset{q \in Q_k}{\arg\max} \left\{ (1-q)^2 \left(\widehat{r}_q^{-1} - 1 \right)^2 \right\}. \tag{5.5.13}$$

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.

5.6 Numerical studies

5.6.1 Examples

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

Example 5.6.1. In this example, we consider $n = 799$ observations of time intervals (in seconds) between successive pulses along a nerve fiber in Hand et al. (1994) (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 drew $B = 250$ subsamples of size $n^* < n$ from the original sample and computed the quadratic error $\mathcal{E}(q, n^*) := B^{-1} \sum_{b=1}^{B} (\widehat{\theta}_{q,n^*} - \widehat{\theta}_{1,n})^2$.

$n^* =$	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 5.1: 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.

The results in Table 5.1 illustrate that setting q slightly larger than one improves the accuracy. The gain is sensible when the sample size is small and persists even for larger samples.

Example 5.6.2. In this example, we apply our method to Newcomb's dataset, representing 66 measurements of the passage time of light. Among others, Brown

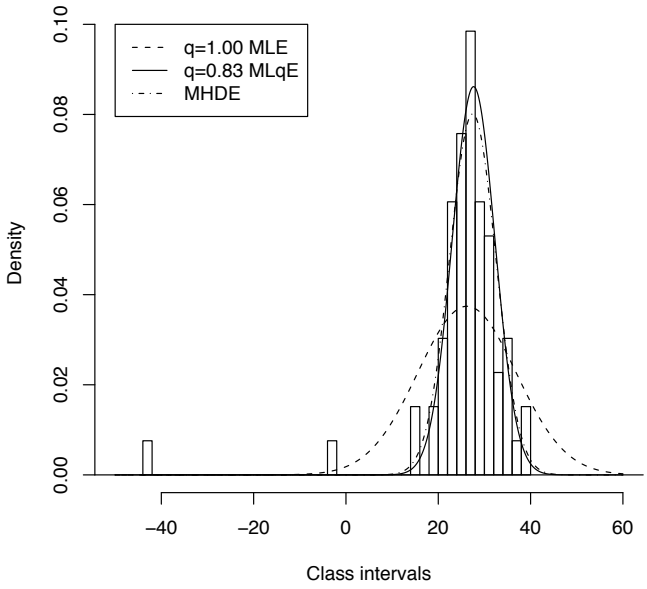
and Hwang (1993), Basu et al. (1998) and Bhandari et al. (2006) analyzed this dataset under a normal model $N(\mu, \sigma)$, as it will be the case here. Since the data present strong outliers at -44 and -2, the selection of q is performed by the criterion function based on the estimated convergence rate of the MLqE. Table 5.2 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 estimates. 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 (1993), p.254, and Basu et al. (1998), 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 (Bhandari et al. (2006), p. 105). Without the outliers, MLqE adapts well to the data and selects $\widehat{q}$ near 1, resulting in estimates close to those of the MLE and giving about the same efficiency. A visual representation is given in Fig. 5.5, 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.

5.6.2 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 $\tau = 0.05$ for $n = 10, 25, 50, 100$. When contamination is included, the samples

(a)



(b)

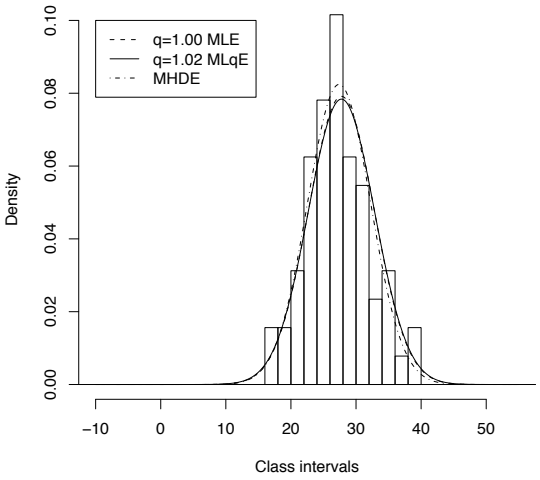


Figure 5.5: Histograms of the Newcomb data with outliers (a) and without (b) with normal densities fitted using maximum Lq-likelihood (MLqE), maximum likelihood (MLE) and minimum Hellinger distance (MHD) estimates. The distortion parameter for MLqE is computed using Windham's criterion.

With outliers

	$\widehat{q} = 0.83$	$q = 1$	$q = 1/2(*)$	MHDE
$\widehat{\mu}$	27.65(0.63)	26.21(1.32)	27.25(0.65)	27.46
$\widehat{\sigma}$	4.63(0.52)	10.66(3.52)	4.34(1.15)	4.98

W/o outliers

	$\widehat{q} = 1.02$	$q = 1$	$q = 1/2(*)$	MHDE
$\widehat{\mu}$	27.76(0.64)	27.75(0.64)	27.25(0.65)	27.40
$\widehat{\sigma}$	5.09(0.45)	5.04(0.46)	4.34(1.15)	4.84

Table 5.2: 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.

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 bandwidth $h = c_n s_n$, 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. (2006)). 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 Table 5.3 suggest that the MLqE performed well whether or not contamination was present. For each simulation setting, we report mean squared error of the estimates, computed as $B^{-1}\sum_b(\widehat{\theta}_b - \theta_0)$, $\theta_0^{\intercal} = (0, 1)$, and 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 $\widehat{q}$ 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 performed 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 bandwidth 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 5.4 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 Windham's criterion, the gain is more modest than the case when the asymptotic criterion is used.

Furthermore, Fig. 5.3 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-adjusted relative efficiency between MLqE and MLE in Eq. (5.4.18). One can see that the true optimal values of q based on the Monte Carlo simulations tend to be greater than the maximum for the solid line.

	$N(0,1)$				$0.95N(0,1)+0.05N(4,1)$			
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 5.3: Monte Carlo squared bias, variance and mean square error of the MLqE, MHDE and MLE of μ and σ for sample sizes 15, 25, 50, 100 and 250 under clear and contaminated normal model.

	$n =$	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 5.4: Monte Carlo relative efficiency between MLqE and MLE for various sample sizes. Asymptotic bias-adjusted relative efficiency (Adj-Eff) and Windham criterion are employed for choosing q.

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 that situation. 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. 5.6, we plot the Monte Carlo estimates of the optimal q 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 bootstrap approximates better the true optima and does sensibly better than the asymptotic criterion for sample sizes of 25 or larger.

5.6.3 Linear Regression

Regression and prediction problems are among the most important in parametric statistics. For the computation of regression coefficients, the MLq method requires a choice of a proper distributional assumption of the error term. In this section, we consider the model

$$Y = X\beta + \varepsilon, \tag{5.6.1}$$

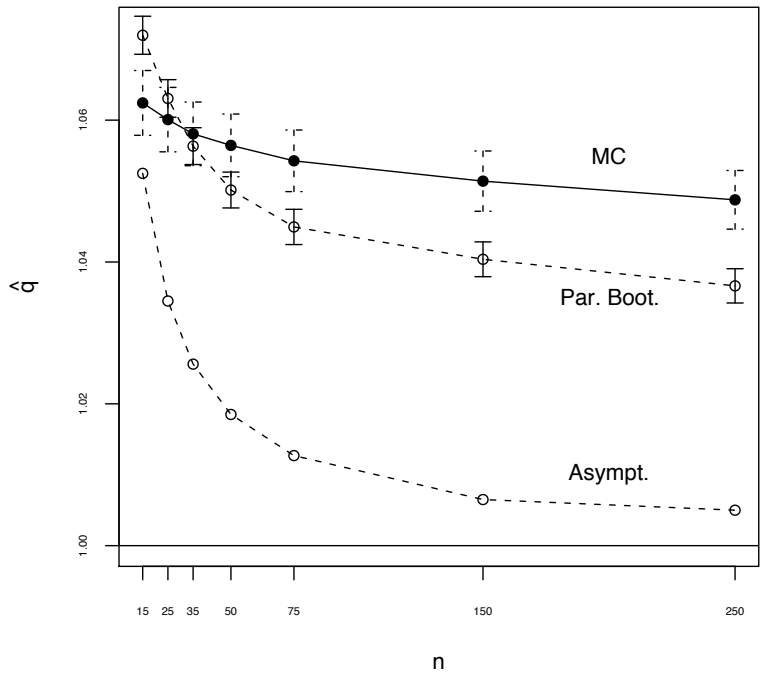


Figure 5.6: True values of q (MC) and estimates of the optimal q via parametric bootstrap (Par.Boot) and minimization of Eq. (5.4.18) (Asympt.). The vertical segments represent 95% confidence intervals.

where Y is a univariate response, X is a p-dimensional vector of predictors and ε is the error term. We consider the case where the residuals are i.i.d. realizations of $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Given a sample (Y_i, X_i), $i = 1, \ldots, n$, in this case the estimating Eq. (5.3.3) amounts to solve

$$\sum_{i=1}^{n} \phi(\epsilon_i; \sigma_\varepsilon, \beta) \left(\frac{n}{2} \log 2\pi\sigma_\varepsilon^2 - \sum_{i=1}^{n} \frac{\epsilon_i^2}{2\sigma_\varepsilon^2} \right) = 0, \tag{5.6.2}$$

where $\phi(\cdot)$ is the density of a univariate normal distribution and $\epsilon_i = Y_i - X_i\beta$. Note that, when the error is assumed to be normal, the MLq estimation task is represented by simultaneous estimation of both regression coefficients and variance of the error term. For simplicity, in our simulations we fix σ_ε.

We present a simulation study of the prediction error of MLqE, focusing on the following aspects:

(i) Sample size and parameter space dimensionality (sparseness);

(ii) Fraction of observations that disagree with the model assumptions (robustness).

All of our experiments are carried out using the following experimental protocol. First, we set a random seed and initialize the simulations by generating design points randomly drawn from the unit hypercube $[-1, 1]^p$. The entries of the true vector of coefficients β are assigned by sampling p points at random in the interval $[-10, 10]$. The values of X and β are kept fixed during the simulations. Then, samples (Y_i, X_i), $i = 1, \ldots, n$ are generated according to two different model specifications: (i) clear normal model; (ii) normal model with contamination. The performance is evaluated by generating B independent out-of-sample observations and computing the empirical squared error

$$PE = \frac{1}{B} \sum_{j=1}^{B} \left(Y_j^{\text{test}} - Y_j^{\text{pred}} \right)^2, \tag{5.6.3}$$

where Y_j^{pred} are predicted values depending on the testing obsevations for the predictors X_j^{test} and on the estimates of the coefficients. Next, we present 2 ex-

periments where we compare the MLqE predictions with other methods. Namely, we consider: Ordinary Least Squares (OLS), Support Vector Machine (SVM) ϵ-regression, ridge regression, L_1-regression (or Least Absolute Deviation Regression (LAD)) and Least Trimmed Square (LTS). The tuning parameters for SVM are chosen by 5-fold cross validation. For ridge regression, we tune the ridge parameter using Generalized Cross-Validation (GCV).

	n	MLqE	$[\widehat{q}]$	OLS	SVM	Ridge	LAD
	25	4.629	0.990	4.628	4.785	4.632	4.654
		(0.040)	(0.050)	(0.043)	(0.046)	(0.043)	(0.043)
$p = 2$	50	4.292	0.99	4.292	4.369	4.295	4.302
		(0.028)	(0.050)	(0.028)	(0.029)	(0.028)	(0.028)
	250	4.031	0.986	4.031	4.046	4.031	4.032
		(0.012)	(0.040)	(0.012)	(0.012)	(0.012)	(0.012)
	25	7.171	0.993	7.17	7.697	7.099	7.488
		(0.086)	(0.010)	(0.097)	(0.084)	(0.098)	(0.133)
$p = 10$	50	5.177	0.995	5.176	5.460	5.172	5.258
		(0.037)	(0.010)	0.037	(0.040)	(0.037)	(0.038)
	250	4.215	0.994	4.215	4.244	4.214	4.223
		(0.015)	(0.010)	(0.014)	(0.014)	(0.015)	(0.014)

Table 5.5: Clear normal model with $\varepsilon \sim N(0, 2)$. Prediction error over 250 realizations for: MLqE with estimates of q, Ordinary Least Squares (OLS), Support Vector Machines (SVM), Ridge Regression (Ridge), Least Absolute Deviation regression (LAD).

Experiment 1 In our first experiment, we generate data from an uncontaminated normal with gaussian error $\varepsilon \sim N(0, 2)$. The size of the testing sample is $B = 250$. The distortion parameter for MLqE is computed using criterion in Eq. (5.5.13). The aim of this experiment is to study the effect of sample size and number of independent variables on the prediction accuracy, under correct model specification.

Experiment 2 In the second experiment we study the effect of contamination by introduction an exogenous distribution. Given $0 < \delta < 1$, we sample $(1 -$

$\delta)n$ observations from our usual model $Y \sim N(X\beta, 2)$ and a smaller number of obervations δn from a model with different slope $Y \sim N(2X\beta, 2)$. In Fig. 5.7 (a), we show a typical realization from a sample of size 100. For this set of experiments the performance is measured in terms of mean squared error with respect the main model.

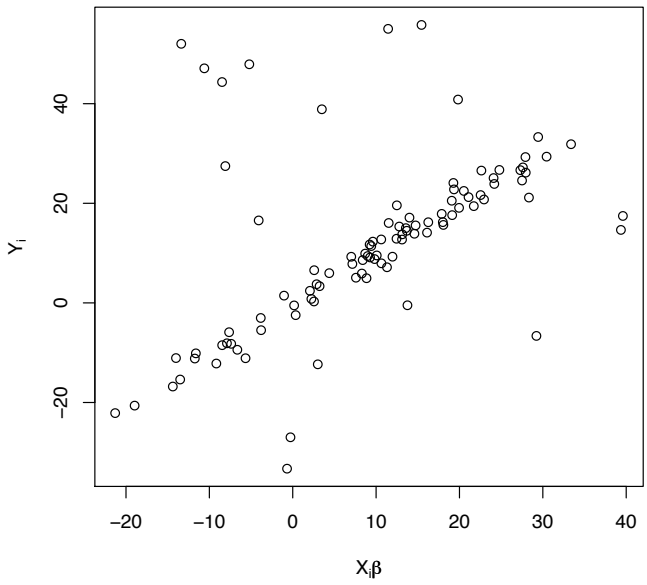


Figure 5.7: Typical realization of $(1-\delta)n = 80$ observations from $Y \sim N(X\beta, 2)$ and $\delta n = 20$ observations from the model $Y \sim N(2X\beta, 2)$

5.6.4 Discussion

In Table 5.4, we present the results of Experiments 1 and 2, under correct model specification. When the model assumptions reasonably hold and the sample size is moderate or large, the performance of the MLqE is similar to that of OLS and Ridge regression. An exception is the case when a sparse design with $n = 25$ is small and $p = 10$ is considered and q is chosen according to criterion (5.5.13), where ridge regression slightly improves upon MLqE. When observations are generated from a

		MLqE	$[\widehat{q}]$	OLS	SVM	Ridge	LAD
$\delta = 0.2$	$p = 5$	0.420	0.601	2.631	0.640	2.404	0.549
	$p = 10$	0.483	0.660	4.371	0.983	3.547	0.657
$\delta = 0.4$	$p = 5$	0.664	0.614	10.839	5.485	10.163	10.316
	$p = 10$	2.843	0.641	15.113	13.355	13.141	15.547

Table 5.6: MSE computed using 250 realizations of samples of size 100 from the contaminated model $Y \sim (1-\delta)N(X\beta, 2) + \delta N(2X\beta, 2)$. Results for: MLqE with average estimate of q, Ordinary Least Squares (OLS), Support Vector Machines (SVM), Ridge Regression (Ridge), Least Absolute Deviation regression (LAD).

well-specified model, the goal of efficiency is prioritized and the empirical bound characterizing criterion (5.5.13) becomes tight. Values that maximize (5.5.13) tend to settle in a neighborhood of 1. Recall that for q approaching 1, the MLqE is actually the OLS. We remark that for all considered sample sizes and parameter dimensions, the MLqE outperforms sensibly both the SVM regression and LAD.

In Table 5.6 we present results from Experiment 3, showing the effect of perturbations applied to the normal noise model. For any portion of "bad data" , the MLqE balances the trade-off between robustness and efficiency. Specifically, the MLqE (i) clearly outperforms nonrobust procedures such as Ridge Regression and OLS, (ii) in addition it retains enough efficiency to outperform most of other specifically designed nonrobust methods. When the portion of bad data is small ($\delta = 0.2$), the MLqE did similarly to robust estimation via LAD . Among the methods that are not specifically designed for robustness, the SVM appears the one with best performance, doing better than OLS and Ridge regression.

The results are interesting when we increase the amount of bad data to almost half ($\delta = 0.4$). The MLqE for $q \to 1/2$ has robustness similar to Helliger distance estimator yielding a breakdown point of 50%. Therefore it is not surprising that it can handle well a large fraction of bad data when $p = 5$. Moreover, unexpectatedly, in higher dimensions ($p = 10$) the MLqE still performs well whereas LAD cannot tolerate so many bad data. Similarly, although the SVM is fairly robust when $\delta = 0.2$, it clearly falls apart when the portion of outlier increases.

5.7 Re-weighting algorithm for multivariate normal

The multivariate normal distribution has the following pdf:

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{5.7.1}$$

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

$$\ell_i := \log f(\mathbf{x}_i;\boldsymbol{\mu},\Sigma) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{5.7.2}$$

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 (5.7.2) is

$$(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}(\mathbf{x}_i-\boldsymbol{\mu}) = (\mathbf{z}_i-\boldsymbol{\mu}^*)'\boldsymbol{\Lambda}^{-1}(\mathbf{z}_i-\boldsymbol{\mu}^*). \tag{5.7.3}$$

Thus, we can rewrite (5.7.2) as:

$$\ell_i = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{p}\log(\lambda_j) - \sum_{j=1}^{p}\frac{(z_{ij}-\mu_j^*)^2}{2\lambda_j}, \tag{5.7.4}$$

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

$$\sum_{i=1}^{n} v_i\frac{\partial\ell_i}{\partial\mu_k} = \sum_{i=1}^{n} v_i\frac{(z_{ik}-\mu_j^*)}{\lambda_k},\ k=1,\ldots,p \tag{5.7.5}$$

and

$$\sum_{i=1}^{n} v_i\frac{\partial\ell_i}{\partial\lambda_k} = -\sum_{i=1}^{n}\frac{v_i}{2\lambda_k} + \sum_{i=1}^{n} v_i\frac{(z_{ik}-\mu_j^*)^2}{2\lambda_k},\ k=1,\ldots,p. \tag{5.7.6}$$

Equating (5.7.5) and (5.7.6) 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

$$\widehat{\boldsymbol{\mu}} = \sum_{i=1}^n v_i \mathbf{x}_i \quad \text{and} \quad \widehat{\Sigma} = \sum_{i=1}^n v_i (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})'(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}). \tag{5.7.7}$$

Chapter 6

An application to Extreme Quantile Estimation in Finance

Estimating financial risk is a critical issue for banks and insurance companies. Recently, quantile estimation based on Extreme Value Theory (EVT) has found a successful domain of application in such a context, outperforming other methods. Given a parametric model provided by EVT, a natural approach is Maximum Likelihood estimation. Although the resulting estimator is asymptotically efficient, often the number of observations available to estimate the parameters of the EVT models is too small to make the large sample property trustworthy. In this chapter, we study a new estimator of the parameters, the Maximum Lq-Likelihood estimator (MLqE), introduced in Chapter 3. We show that the MLqE outperforms the standard MLE, when estimating tail probabilities and quantiles of the Generalized Extreme Value (GEV) and the Generalized Pareto (GP) distributions. First, we assess the relative efficiency between the the MLqE and the MLE for various sample sizes, using Monte Carlo simulations. Second, we analyze the performance of the MLqE for extreme quantile estimation using real-world financial data. The MLqE is characterized by a distortion parameter q and extends the traditional log-likelihood maximization procedure. When $q \to 1$, the new estimator approaches the traditional Maximum Likelihood Estimator (MLE), recovering its desirable asymptotic properties; when $q \neq 1$ and the sample size is moderate or small, the MLqE successfully trades bias for variance, resulting in an overall gain in terms of

accuracy (Mean Squared Error).

6.1 Introduction

Recent financial crises and new regulations for banks and insurance companies[1] have prompted intermediaries to regularly compute statistical tail-related measures of risk. One of the most popular measures of financial risk is the Value-at-Risk (VaR), usually defined as the α-th quantile of the distribution of losses (negative returns). Although the appropriateness of VaR as a risk measure (Artzner et al., 1999) has been recently questioned, it is still the most widely used for risk management, asset allocation and risk-adjusted performance evaluation. Various methods have been proposed to estimate VaR: historical approach, parametric quantile estimators (e.g., Normal or t-Student parametric models), variance-covariance models and Monte Carlo methods are the most commonly used techniques. Recently, Extreme Value Theory (EVT) has found extensive application in finance to estimate tail-related risk measures, as it has been shown to provide estimators that perform best overall in predicting Value-at-Risk (Brooks et al., 2005; Kuester et al., 2006).

EVT is supported by a sound statistical theory and it relies on the asymptotic properties of the distributions of sample extrema. Specifically, the two prevailing parametric approaches for modelling extreme events are the Peaks-Over-Threshold (POT) and Block Maxima (BM) methods (e.g., Gilli and Kellezi (2006); McNeil et al. (2005)). The POT method exploits the Generalized Pareto (GP) distribution for modelling the exceedances over a certain threshold, while the BM method relies on the Generalized Extreme Value (GEV) distribution to model the maximum value that a variable takes in a given period of time (block).

Although maximum likelihood is the most popular estimation approach in this context, mainly due to its asymptotic properties and ease of implementation[2], often the number of observations available to estimate GEV and GP parameters

[1] Basel II for banks, Solvency II for insurance companies and IFRS 32 and 39 for all financial companies.

[2] Other methods include the method of moments, the method of probability-weighted moments and the elemental percentile method. The reader is referred to Hosking and Wallis (1987), Grimshaw (1993) and Castillo et al. (1997).

is too small to guarantee the desirable large sample properties of the Maximum Likelihood Estimator (MLE); thus, inference might not be trustworthy. Our investigation aims to address this issue by studying, for the first time in the EVT context, the performance of a new estimator of the parameters: the Maximum Lq-Likelihood Estimator (MLqE), recently proposed by Ferrari and Yang (2007). The MLqE is based on the information measure introduced by Havrda and Charvát (1967) and generalizes the traditional log-likelihood maximization procedure: it preserves the desirable asymptotic properties of the traditional MLE, while it allows for a peculiar type of distortion introduced by the extra parameter q, resulting in a gain in terms of precision (Mean Squared Error) when the sample size is moderate or small.

The objective of this chapter is to study the behavior of the new estimator on both simulated data and on real-world time series for extreme quantile estimation. First, we show that the new estimator is more efficient than the standard MLE when the goal is to estimate the tail probability of the GP and GEV distributions. The comparison is carried out through Monte Carlo simulations, where the performance of the two estimators is evaluated for different choices of the tail probability and sample size. We show that when the distortion parameter q is properly chosen, the Mean Squared Error of the MLqE is sensibly smaller than that of MLE. Second, we focus on extreme quantile estimation, assessing the performance of MLqE on a financial stock market index for both GEV and GP distributions. The comparison with the MLE indicates that choices of the distortion parameter q smaller than 1 can dramatically reduce the generalization error.

The chapter is organized as follows. In Section 6.2, we describe the two main parametric approaches for risk estimation based on EVT; in section 6.3 we introduce the Maximum Lq-Likelihood Estimator. In Section 6.4 we present a Monte Carlo simulation study to explore the relative efficiency between the MLqE and the MLE in a finite-sample situation. Section 6.5 describes a hold-out validation procedure applied to real-world financial data and compares the generalization error of the new estimator with that of MLE. Finally, in section 6.6 we outline the conclusions.

6.2 Extreme Value Theory for tail-related risk measures

Extreme Value Theory has found numerous applications in various fields (Lazar, 2004), including finance. The reader is referred to Embrechts et al. (1997), and Reiss and Thomas (1997) for an overview of the main applications in finance, while a brief description of the two main approaches, namely the Peaks-Over-Threshold and the Block Maxima, is reported below.

6.2.1 Peaks-Over-Threshold

The POT approach considers exceedances over a certain threshold u. Let $\{X_i, 1 \leq i \leq n\}$ be a random sample from a distribution F with mean μ and variance σ^2. An *exceedance* occurs when $X_i > u$ and an *excess over* u is defined by $y = x - u$. The conditional distribution of the exceedances over u, taken at $X > u$ is

$$F_u(y) = P\left(X - u \leq y | X > u\right) = \frac{F(u+y) - F(u)}{1 - F(u)}, \; y \geq 0. \tag{6.2.1}$$

Balkema and de Haan (1974) showed that for a large class of distributions, $F_u(y) \rightarrow G(y)$ as $u \rightarrow \infty$ where $G(y)$ is a *Generalized Pareto (GP)* distribution. A representation of the GP distribution is

$$G(y; \xi, \sigma) = \begin{cases} 1 - \left(1 + \dfrac{\xi}{\sigma} y\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-y/\sigma\right), & \xi = 0, \end{cases} \tag{6.2.2}$$

where

$$y \in \begin{cases} [0, \infty), & \xi \geq 0, \\ [0, -\sigma/\xi], & \xi < 0. \end{cases}$$

The probability density function g is obtained by differentiating with respect to x:

$$g(x; \xi, \sigma) = \begin{cases} \sigma^{-1} \left(1 + \xi \dfrac{(x-u)}{\sigma}\right)^{-(1/\xi + 1)}, & \xi \neq 0, \\ \sigma^{-1} \exp\left(-(x-u)/\sigma\right), & \xi = 0. \end{cases} \tag{6.2.3}$$

The shape parameter ξ can be positive, negative or zero and provides an indication on the heaviness of the tail. The GP can represent different distributions depending on the value taken by ξ. In particular, when $\xi > 0$, we obtain the ordinary Pareto distribution which is suitable for modelling heavy tailed distributions such as financial returns. When $\xi = 0$ and $\xi < 0$ we have respectively the exponential and the Pareto II type distributions.

From eq.(6.2.1) one can obtain the following equality for values of x larger than u:

$$1 - F(x) = (1 - F(u))(1 - F_u(x - u)). \tag{6.2.4}$$

Given a sufficiently high threshold value, $F_u(x-u)$ can be estimated using the plug-in estimate based on GP distribution and $F(u)$ can be estimated using the sample proportion of observations. Thus, from eq.(6.2.4) one can write the tail estimator of $F(x)$. Inverting the expression for the tail gives the estimating equation of the Value-at-Risk[3]

$$\widehat{VaR}_{1-\alpha} = u + \frac{\hat{\sigma}}{\hat{\xi}} \left(\frac{n}{N_u} \alpha^{-\hat{\xi}} - 1 \right), \tag{6.2.5}$$

where N_u denotes the observed number of exceedances over the threshold u. The reader is referred to McNeil et al. (2005) for a complete mathematical treatment of the POT approach for Value-at-Risk estimation.

Note that the asymptotic result poses some applicability constraints. In fact, the threshold u has to be large in order for the Generalized Pareto approximation to hold; as a consequence, few exceedances would be left. Thus, if an excessively high threshold is chosen, the plug-in estimator might be inaccurate with high variance. Furthermore, the asymptotic properties of the Maximum Likelihood estimator would hardly hold. Conversely, a low threshold would inevitably induce bias.

[3]The Value-at-Risk is usually defined as the α-th quantile of the distribution of losses, or the negative returns. Namely, $\text{VaR}_{1-\alpha} := \inf\{x \in \Re : P(X > x) \leq \alpha\}$ where X is a real-valued random variable representing losses or negative returns and $0 \leq \alpha \leq 1$. Typically, values of interest for α are 0.05 and 0.01.

6.2.2 Block Maxima

The BM method models the maximum value that a variable takes in a given period of time (block). Consider a random variable X with cumulative distribution function $F(x)$ with mean μ and variance σ^2. Let $\{Y_i, 1 \leq i \leq n\}$ be a random sample from the standardized distribution $F\left(\frac{x-\mu}{\sigma}\right)$ and define[4]

$$Y_{n,n} = \max\{Y_1, Y_2, ..., Y_n\}.$$

In addition, let $\{a_n; n \geq 1\}$ and $\{b_n \geq 0; n \geq 1\}$ be sequences of numbers such that

$$P\left(\frac{Y_{n,n} - a_n}{b_n} \leq x\right) \rightarrow G(y), \tag{6.2.6}$$

as $n \rightarrow \infty$ for some nondegenerate distribution G. Fisher and Tippett (1928), and Gnedenko (1943) showed that G belongs to one of the following three extreme value distributions:

Gumbel: $\Lambda(y) = \exp(-\exp(-y)),\ -\infty \leq y \leq \infty$

Fréchet: $\Phi(y;\alpha) = \begin{cases} 0, & y \leq 0 \\ \exp(-y^{-\alpha}), & y > 0, \quad \alpha > 0 \end{cases}$

Weibull: $\Psi(y;\alpha) = \begin{cases} \exp(-(-y)^{\alpha}), & y \leq 0 \quad \alpha > 0 \\ 1, & y > 0. \end{cases}$

Later, Jenkinson (1955) and von Mises (1954) suggested a re-parametrization of the above expressions by setting $\xi = \alpha^{-1}$ for the Fréchet distribution and $\xi = -\alpha^{-1}$ for the Weibull distribution. Thus, Gumbel, Fréchet and Weibull can be represented in a unified parametric model, known as the Generalized Extreme Value distribution (GEV), where ξ represents the shape parameter and gives an indication about the heaviness of the tail of the distribution.

[4]We could study as well the minimum rather than the maximum and the results for one of the two can be immediately transferred using the relationship $Y_{1,n} = -max\{-Y_1, -Y_2, ..., -Y_n\}$.

The following characterization, which includes also the location and scale parameters μ and σ, is most commonly used:

$$H(x;\xi,\mu,\sigma) = \begin{cases} \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right], & \text{if } \xi \neq 0, \quad 1+\xi\frac{x-\mu}{\sigma} > 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right], & \text{if } \xi = 0. \end{cases} \tag{6.2.7}$$

The probability density function is then:

$$h(x;\xi,\mu,\sigma) = \begin{cases} \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi-1}\exp\left(-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{1/\xi}\right) & \text{if } \xi \neq 0, \\ \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left(\exp\left(-\frac{x-\mu}{\sigma}\right)\right), & \text{if } \xi = 0, \end{cases} \tag{6.2.8}$$

with $1+\xi\frac{x-\mu}{\sigma} > 0$. We remark that the asymptotic results just described only guarantee that Y is *approximately* distributed according to a GEV distribution. Hence, the accuracy of such an approximation relies strongly on the size of the blocks from which the maxima are computed.

The block maxima approach allows one to compute the so-called return-level, that is the level expected to be exceeded in one out of the k periods of length n. Given a block size large enough to hold the GEV approximation, the return level can be computed by inverting eq.(6.2.7) and thus obtaining

$$U_k = H^{-1}\left(1-\frac{1}{k};\xi,\sigma,\mu\right). \tag{6.2.9}$$

Substituting the parameter estimates, we have

$$\widehat{U}_k = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1-\left(-\log\left(1-\frac{1}{k}\right)\right)^{-\hat{\xi}}\right), & \text{if } \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log\left(-\log\left(1-\frac{1}{k}\right)\right), & \text{if } \hat{\xi} = 0. \end{cases} \tag{6.2.10}$$

6.3 The Maximum L_q-Likelihood Method

Let $f(x;\theta_0)$ be the GP density in eq.(6.2.3) or the GEV density in eq.(6.2.8), where $\theta_0 = (\theta_{01},...,\theta_{0p}) \in \Theta$ denotes the vector of parameters to estimate (p=2 for GP and p=3 for GEV). Given a random sample $X_1,...,X_n$ from $f(x;\theta_0)$, the Maximum

Likelihood Estimator is

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log\left[f(X_i;\theta)\right]. \tag{6.3.1}$$

Maximum Likelihood is the standard approach in parametric estimation, mainly due to the desirable asymptotic properties of consistency, efficiency and normality. In particular, under some regularity conditions (e.g., see Van der Vaart (1998), Ferguson (1996)), we have that $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, V\right)$ as $n \to \infty$, where V represents the inverse of the Fisher information matrix. However, when applying POT and BM methods, inference based on the asymptotic results of MLE can be unreliable. Note that in the POT method, an increasingly high threshold u is needed to guarantee the convergence to the GP distribution. Similarly, in the BM method, a large block size is necessary in order to hold the GEV distribution (see sections 2.1 and 2.2). Clearly, if we choose high thresholds or large block sizes, the number of observations left for parameter estimation is small, making the asymptotic properties of MLE not trustworthy.

Recently, in order to handle this issue, Ferrari and Yang (2007) introduced an estimator based on the generalized information measure introduced by Havrda and Charvát (1967) [5], the Maximum L_q-Likelihood Estimator (MLqE). The MLqE of θ_0 is defined as

$$\widetilde{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} L_q\left[f(X_i;\theta)\right], \tag{6.3.2}$$

where

$$L_q(z) = \begin{cases} \dfrac{z^{1-q}-1}{1-q} & if\ q \neq 1, \\ \log z & if\ q = 1. \end{cases} \tag{6.3.3}$$

The function L_q represents a Box-Cox transformation in statistics and in other contexts it is often called deformed logarithm of order q. The estimates of the

[5] Such information measures, usually called α-order entropies (or q-entropies in physics), which relaxes the additivity assumption that characterizes Shannon's information. In recent years α-order entropies have found successful applications in different fields, such as finance, biomedical sciences, environmental sciences and linguistics (e.g., see Gell-Mann (2004)).

parameters are computed by solving the following system of equations:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} L_q \left[f(X_i; \theta) \right] = 0, \;\; j = 1, 2, ..., p. \tag{6.3.4}$$

When q is a fixed constant, $\widetilde{\theta}_n$ belongs to the class of *M-estimators*. Under some regularity conditions, such estimators have well known asymptotic proprieties such as asymptotic normality (e.g., see Van der Vaart (1998) and Huber (1981a)). The classical regularity requirements concern mainly the smoothness of the objective function under exam (in this case, $L_q \left[f(x; \theta) \right]$) with respect to its parameters and the boundedness in probability of the third derivative of the objective function in a neighborhood of θ_0.

The MLqE can be considered as a generalization of the traditional MLE. For values of q arbitrarily close to 1, we have that $L_q(\cdot) \rightarrow \log(\cdot)$ and the MLqE approaches the classical MLE. However, an advantage is obtained by having q slightly different from 1: in this situation the MLqE allows trading bias for variance and provides more accurate estimates when the sample size is small. Chosing $q \neq 1$ corresponds to assign a different weight to the observations in the sample based on the rarity of their occurrence. In particular, when $q < 1$ the role played by extreme observations, which are the most influential on the estimates, is reduced. Consequently, when setting $q < 1$ the variability is reduced by increasing the bias, which can result in an overall gain in terms of Mean Squared Error, as we shall see in section 4. Conversely, if $q > 1$ the role of the observations corresponding to density values close to zero is accentuated.

The peculiar type of distortion introduced by our estimator allows a gain in terms of precision (Mean Squared Error) by reducing the variance when both the sample size and the tail probability to be estimated are small. Conversely, when the sample size is large, reducing the amount of bias allows for the recovery of a number of desirable large sample properties such as efficiency and consistency. Hence, the MLq procedure extends the classic method resulting in a general inferential procedure that inherits most of the desirable features of traditional maximum likelihood methods and at the same time gains some new properties that can be

usefully exploited in *ad hoc* estimation settings. The following sections report empirical results supporting the use of such estimator in the EVT framework.

6.4 Finite-sample efficiency of MLqE: Monte Carlo simulations

In this section, we compare the relative efficiency between the MLqE and the MLE on simulated data from both GEV and GP distributions[6]. Our first aim is to investigate whether the MLqE can outperform, in terms of Mean Squared Error, the classical MLE when estimating small tail probabilities. The estimates of the tail probability are obtained by using the so-called *plug-in* approach, where the point estimate of the unknown parameter is substituted into the distribution of interest.

Let $F(x;\theta)$ be the cumulative distribution function for either GEV or GP distributions. The true parameter is denoted by θ_0 and the true tail probability by α (in particular, $\alpha = 1 - F(x;\theta_0)$ if the right tail is considered, and $\alpha = F(x;\theta_0)$ otherwise). Further, let $\widehat{\alpha}_n$ and $\widetilde{\alpha}_n$ be the plug-in estimates of α, obtained respectively via the ML and the MLq methods.

The relative performance of the two estimators is measured by taking the ratio between the two Mean Squared Errors:

$$R_n = \frac{MSE\,(\widehat{\alpha}_n)}{MSE\,(\widetilde{\alpha}_n)} = \frac{E(\widehat{\alpha}_n - \alpha)^2}{E(\widetilde{\alpha}_n - \alpha)^2} = \frac{(E(\widehat{\alpha}_n) - \alpha)^2 + Var(\widehat{\alpha}_n)}{(E(\widetilde{\alpha}_n) - \alpha)^2 + Var(\widetilde{\alpha}_n)}. \tag{6.4.1}$$

As pointed out by the error decomposition in the above expression, we are interested in the relative trade-off between bias and variance of the two estimators, for *a given sample size*. The simulations are then carried out as follows:

- For any given sample size n, a number $B = 1000$ of random samples $X_1, ..., X_n$ are generated from either GEV or GP with parameter vector θ_0.

[6]The analyses presented in sections 4 and 5 are performed using the statistical computing envirnoment R (R Development Core Team (2006)). In the routines described we use functions from the Extreme Value Theory package *evir* (McNeil and Stephenson (2007)).

- For each sample, $\widehat{\alpha}_{n,b}$ and $\widetilde{\alpha}_{n,b}$, $b = 1, ..., B$, the ML and MLq estimates of the tail probability α are obtained. The estimates of the parameters for both estimators are computed by solving numerically the L_q-likelihood equations (6.3.4). The optimization is performed by using a variable metric algorithm (e.g., see Givens and Hoeting (2005)), where the MLE estimates $\widehat{\theta}_{n,b}$ are chosen as starting values.

- Finally, the relative performance between the two estimators is evaluated by the ratio

$$\widehat{R}_n = \frac{\widehat{\mu}}{\widetilde{\mu}} = \frac{\sum_{k=1}^{B}(\hat{\alpha}_{n,k} - \alpha)^2/B}{\sum_{k=1}^{B}(\widetilde{\alpha}_{n,k} - \alpha)^2/B}$$

  where $\widehat{\mu}$ and $\widetilde{\mu}$ represent the Monte Carlo estimates of the Mean Squared Error for MLE and MLqE, respectively. Furthermore, the standard error of $\widehat{R}_n$ is computed via the multivariate Delta Method as

$$se\left(\widehat{R}_n\right) = B^{-1/2}\left(\frac{\widehat{\sigma}_{11}}{\widetilde{\mu}^2} - 2\widehat{\sigma}_{12}\frac{\widehat{\mu}}{\widetilde{\mu}^3} + \widehat{\sigma}_{22}\frac{\widehat{\mu}^2}{\widetilde{\mu}^4}\right)^{1/2} \tag{6.4.2}$$

  where $\widehat{\sigma}_{11}$, $\widehat{\sigma}_{22}$ and $\widehat{\sigma}_{12}$ denote respectively the Monte Carlo estimates for the variances and the covariance of the squared errors (see Appendix 1 for the details of the calculation).

The procedure described above is repeated for several samples sizes (ranging from 5 to 200) and different choices of the true tail probability α and the distortion parameter q. The simulations discussed in the remainder of this section are obtained by sampling from a GEV distribution with parameters

$$\theta_0 = (\xi_0, \mu_0, \sigma_0) = (0.1, 0.05, 0.015),$$

and from a GP distribution with parameters

$$\theta_0 = (\xi_0, \sigma_0) = (0.5, 1).$$

We remark that the parameter values[7] are chosen to be similar in size to the estimates for various stock indexes computed by Gilli and Kellezi (2006) and McNeil

[7]The value of the shape parameter ξ, which determines heaviness of the tail, is critical for both GEV and GP distributions. Since financial returns are usually heavy tailed distributions Cont (2001), they can be suitably represented by considering $\xi > 0$.

et al. (2005) using the traditional ML method. Nevertheless, we also performed simulations using other parameter settings, obtaining similar results.

Fig. 6.1 and 6.2 show the results for the GP distribution. In particular, Fig. 6.1 shows the performance of the MLqE when q is 0.94 for different values of the tail probability α. For small and moderate sample sizes, we have that $\widehat{R}_n > 1$ and the MLqE is clearly more accurate than MLE. From Fig 6.2, one can see that MLq estimates are more precise not only for small but even for larger sample sizes (up to 200). Moreover, for a given tail probability the gain is more accentuated when q is smaller.

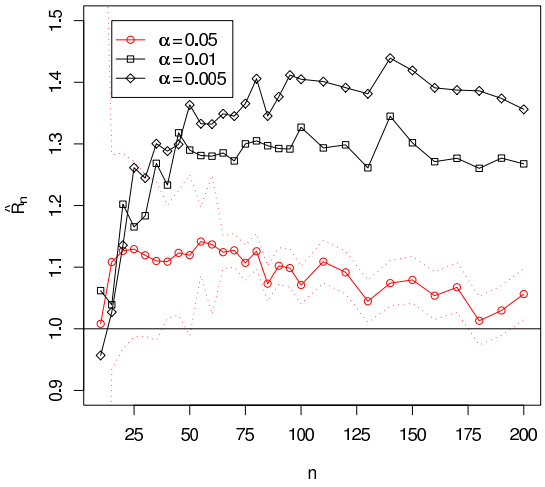


Figure 6.1: GP distribution. Monte Carlo Mean Squared Error ratio computed from $B = 1000$ samples of size n, for $\alpha = 0.05, 0.01, 0.005$ and $q = 0.94$. The dashed lines represent 95% confidence bands for the case when $\alpha = 0.05$.

Fig. 6.3 and 6.4 present the case of the GEV distribution. Similarly to the GP distribution, Fig. 6.3 points out that MLqE is more accurate than the MLE for moderate or small sample sizes. Moreover, the gain appears to be more evident for smaller values of α. Actually, note that when α is 0.05, the MLqE outperforms the MLE in accuracy only for sample sizes smaller than 80, while this is not the case when α equal to 0.01. In figure 4 we can see that the relative performance

of MLqE versus MLE improves when the tail size becomes smaller ($\alpha = 0.005$) and the parameter q decreases from 0.95 to 0.93. Recall that decreasing the distortion parameter q is equivalent to downweighting extreme observations that can be dramatically influential on the accuracy of the estimates when the size of α is small.

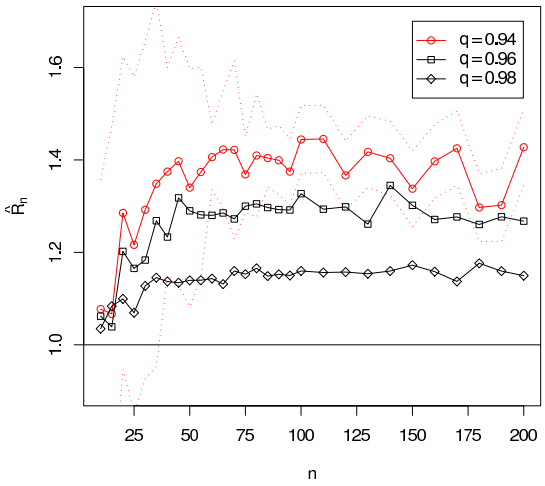


Figure 6.2: GP distribution. Monte Carlo Mean Squared Error ratio computed from $B = 1000$ samples of size n, for various values of the distortion parameter ($q = 0.94, 0.96, 0.98$) and true tail probability $\alpha = 0.01$.

In general, if q is fixed, it is important to note that as the sample size gets larger, the bias component of the error becomes more relevant than the variance component and the MLE will always tend to dominate MLqE due to its asymptotic properties. This observation has suggested that a value of q closer to 1 should be preferred when the sample size increases.

6.5 Forecasting financial empirical quantiles

The simulation results have encouraged a further investigation on real-world financial data, where Extreme Value Theory plays a crucial role in forecasting the

empirical quantiles. The analyses presented in the following sections have been carried out on publicly available financial data[8]: the daily log-returns of the Standard & Poor's 500 index (S&P500) from January 1960 to June 1993. Extreme value analysis on these data set has been previously discussed in literature (e.g., see McNeil and Frey (2000), and Knight et al. (2003)). The summary statistics for this data set are given in table 1. A plot of the time series is provided in figure 5.

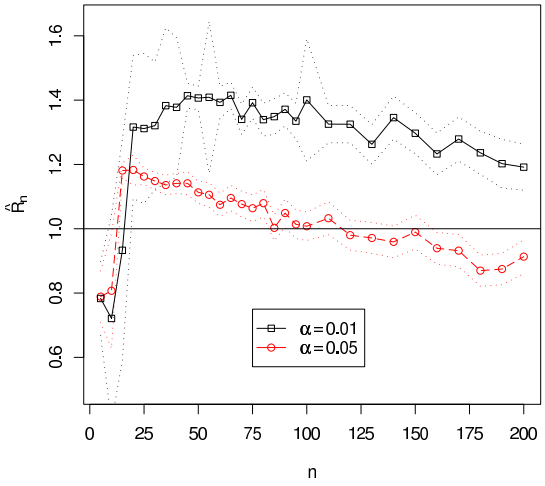


Figure 6.3: GEV distribution. Monte Carlo Mean Squared Error ratio computed from $B = 1000$ samples of size n, for two values of the true tail probability ($\alpha = 0.01, 0.05$) and distortion parameter $q = 0.95$. The dashed lines represent 95% confidence bands.

Sample Size	Min	Max	Mean	St.Dev.	Skewness	Kurtosis
8414	-20.388	9.099	0.028	0.871	-1.510	44.300

Table 6.1: Descriptive statistics of the log-return series of S&P500 index.

This data set presents features that commonly characterize the distribution of financial log-returns. In particular, note that the distribution of returns for the S&P500 index is remarkably skewed with large kurtosis. In the remainder of this

[8]http://www.ma.hw.ac.uk/ mcneil/data.html

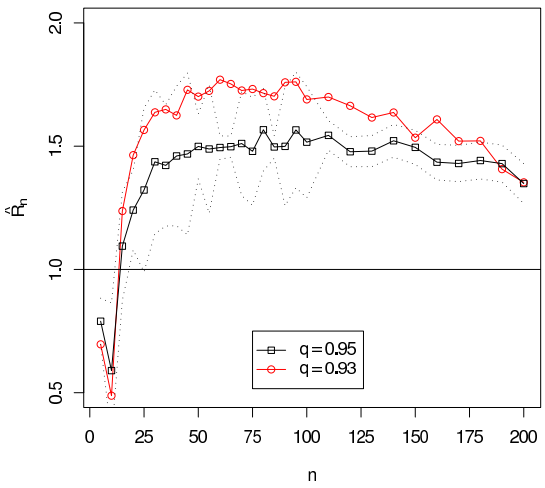


Figure 6.4: GEV distribution. Monte Carlo Mean Squared Error ratio computed from $B = 1000$ samples of size n, for two values of the distortion parameter ($q = 0.93, 0.95$) and true tail probability $\alpha = 0.005$. The dashed lines represent 95% confidence bands for the case when $q = 0.95$.

section we consider the commonly employed hold-out procedure to estimate the generalization error of the estimates. We use such a measure to (i) compare the relative performance between MLqE and MLE when predicting empirical quantiles of one of the extreme value distributions (GEV or GP) and (ii) study the performance of MLqE, relatively to the tail size α and the distortion parameter q.

6.5.1 Hold-out validation procedure

The comparison between the MLqE and the MLE is carried out using an estimate of the generalization error (Hastie et al. (2001)), obtained via a *repeated hold-out* procedure. First, from the original dataset of the log-returns we take the block maxima (for the BM model) or the exceedances over a certain threshold (for the POT model). Then, on the filtered data, the following steps are performed:

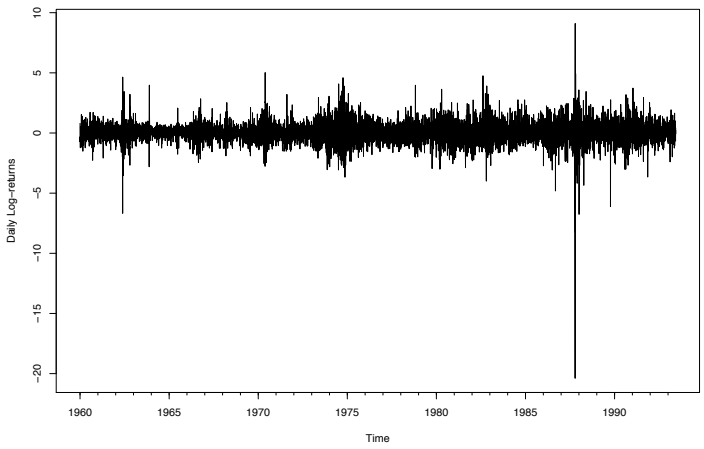


Figure 6.5: Daily returns of the S&P500 index.

(i) The data are randomly divided into a training set of size $n^{(tr)}$ and a testing set of size $n^{(ts)} = n - n^{(tr)}$, where n is the size of the filtered sample. The training and testing samples are chosen such that $n^{(ts)} = n^{(tr)}$.

(ii) The ML and MLq estimates of the quantile τ, denoted by $\widehat{\tau}^{(tr)}$, are computed from the training set.

(iii) The sample quantile, $t^{(ts)}$, is computed from the testing set.

Steps (i),(ii) and (iii) are repeated for $B = 500$ times and then the performance of the estimator is evaluated by

$$\widetilde{\mathcal{E}} = B^{-1} \sum_{b=1}^{B} \left(\widehat{\tau}_b^{(tr)} - t_b^{(ts)} \right)^2. \tag{6.5.1}$$

Finally, the standard error of $\widetilde{\mathcal{E}}$ is calculated using nonparametric bootstrap, based on 2000 replications. The analysis is carried out for both left and right tails of the distributions of returns.

6.5.2 Empirical results on financial data

In the filtering phase, 100 observations are extracted from the S&P500 log-return time series. In the BM model, the original sample is divided in $n = 100$ blocks, obtaining a block size reasonably large in order for the GEV asymptotic approximation to apply (e.g., see Gilli and Kellezi (2006)). In the POT model, although some data driven procedures have been proposed (Lazar, 2004), there seems to be no universal agreement on the choice of the threshold value to employ. However, Monte Carlo studies (McNeil and Frey, 2000) have shown that for heavy tailed distributions a threshold corresponding to about $n = 100$ exceedances performs well in terms of Mean Squared Error[9].

Tables 6.2 and 6.3 report the empirical results for the BM and POT approaches, for different quantiles and choices of the distortion parameter. Column 3 and 5 report the generalization error $\widetilde{\mathcal{E}}$ for the left tail and the right tail, while columns 4 and 6 report the percent gain (or loss) in terms of prediction error of the MLqE over that of MLE. The results for the MLE are reported in the row corresponding to $q = 1$, since the two estimators are the same for such a value.

For the BM method, a substantial improvement is obtained when $q < 1$. In all the cases, the improvement is relevant when the distortion parameter decreases to $q = 0.95$. Furthermore, we notice that the gain deriving from the MLq method is more evident on the left tail, which is usually of major interest in risk analysis as it represents the losses. Actually, it is known that equity times series usually show a loss/gain asymmetry (Cont, 2001) with left-skewed distributions, as shown in Table 6.1 for the data sets under examination. Finally, as expected, as the distortion parameter approaches 1, the usual MLE is recovered and the performance of the two estimators becomes similar.

Table 6.3 shows the results corresponding to the POT method. The analysis on the left tail confirms the considerations previously discussed for the BM method. However, the performance on the right tail shows only little or no improvement with respect the standard approach, when considering the 90th percentile. Nevertheless, the analysis clearly points out that the MLqE can be considered as a

[9]This choice is also confirmed by preliminary exploratory analyses carried out by using the graphical tools contained in the R package POT (Ribatet, 2006).

valid alternative to the MLE when computing the value at risk of financial losses, especially if interested in estimating extreme quantiles.

For the right tail, our results show that the MLqE performs better in the case of the Block Maxima method. However, in the case of the left tail we cannot conclude the same, as the estimated error for MLqE is smaller for the POT model when the 95th percentile is considered. The behavior observed for the right tail is due to both the filtering method (block maxima vs thresholding) and the characteristics of the empirical distribution of the filtered data. Financial times series are characterized by an asymmetric behavior of gains and losses. Typically, one observes large drawdowns in stock prices but not equally large upward movements. The right tail of the empirical distribution is usually lighter than the left one (Cont 2001), as it is the case in our data. Consequently, in the POT method, the variability of the exceedances over the given threshold is smaller and the fitted GP distribution is less spread out. In such a case, we observe little gain from distorting the logarithmic loss function, as shown in Tables 2 and 3.

Percentile	q	Left Tail		Right Tail	
		$\widetilde{\mathcal{E}}$	% Gain	$\widetilde{\mathcal{E}}$	% Gain
90th	1.000	0.3836(0.0332)	/	0.2759(0.0248)	/
	0.995	0.3642(0.0320)	5.3155	0.2716(0.0244)	1.5981
	0.975	0.2981(0.0273)	28.6603	0.2565(0.0227)	7.5647
	0.950	0.2373(0.0231)	61.6476	0.2429(0.0210)	13.5677
95th	1.000	1.3706(0.1135)	/	0.5583(0.0531)	/
	0.995	1.3213(0.1128)	3.7305	0.5449(0.0523)	2.4618
	0.975	1.1594(0.1025)	18.2168	0.4971(0.0478)	12.3238
	0.950	1.0239(0.0953)	33.8575	0.4505(0.0432)	23.9352

Table 6.2: Block Maxima method. The squared error, $\widetilde{\mathcal{E}}$, is computed for $q = 1, 0.995, 0.975$ and 0.95 (where $q = 1$ corresponds to the MLE) and considering two choices of the tail size. In parenthesis, the bootstrap standard error of $\widetilde{\mathcal{E}}$, computed from 2000 replicates. The percent gain is computed as $(\widetilde{\mathcal{E}}_{MLE}/\widetilde{\mathcal{E}}_{MLqE} - 1) \times 100$.

Percentile	q	Left Tail		Right Tail	
		$\widetilde{\mathcal{E}}$	% Gain	$\widetilde{\mathcal{E}}$	% Gain
90th	1.000	1.4190(0.1307)	/	0.522(0.0622)	/
	0.995	1.3627(0.1277)	4.1327	0.522(0.0618)	0.0094
	0.975	1.1630(0.1107)	22.0158	0.5233(0.0623)	-0.2430
	0.950	0.9671(0.0933)	46.7225	0.5285(0.0631)	-1.2296
95th	1.000	7.0898(0.6909)	/	1.7555(0.2685)	/
	0.995	6.7981(0.6763)	4.2903	1.7452(0.2666)	0.5887
	0.975	5.7576(0.5813)	23.1364	1.7103(0.2757)	2.6421
	0.950	4.7161(0.4806)	50.3295	1.6814(0.2883)	4.4060

Table 6.3: Peaks-Over-Threshold method. Squared error, $\widetilde{\mathcal{E}}$, for $q = 1, .995, .975$ and .95 (when $q = 1$ we are computing the MLE) and two choices of the tail size. In parenthesis, the bootstrap standard error of $\widetilde{\mathcal{E}}$, computed from 2000 replicates. The percent gain is computed as $(\widetilde{\mathcal{E}}_{MLE}/\widetilde{\mathcal{E}}_{MLqE} - 1) \times 100$.

6.6 Discussion and Final Remarks

In this chapter, we have shown that the MLqE can be a valid alternative to the classical MLE when estimating a small tail probability or a large quantile in the context of Extreme Value Theory. The MLqE can be regarded as a natural extension of the classical MLE. Specifically, the distortion parameter q adjusts the relative weight of the information provided by each observation in the sample. If q is close to 1, the estimator preserves the large sample properties of the MLE, while for $q \neq 1$ the trade-off between bias and variance is modified, producing an overall gain in terms of accuracy (Mean Squared Error) when the sample size and/or the tail probability to estimate are small. Such settings are typical in finance, where the attention is often on estimating very small probabilities with a small number of extrema. Although we have considered the MLqE for the specific purpose of Extreme Value Theory estimation, this stream of research seems to be very promising, due to the considerable flexibility of the new estimator to many

classical estimation settings and its finite-sample variance reduction properties. The simulation study has pointed out that the MLqE is more accurate than MLE in estimating tail probabilities for GEV and GP distributions for relatively small and moderate sample sizes. The gain from the MLqE appears to be more remarkable when the target tail probability is smaller. When the sample size is too large relative to the choice of the distortion parameter q, the bias component plays an increasingly relevant role and eventually we observe that the MLqE decreases its accuracy. This indicates that the distortion parameter should approach 1 as the sample size increases in order to preserve the efficiency gain. In addition, smaller values of the distortion parameter q enhance the accuracy attainable in small sample situations by reducing the role played by extreme (and more influential) observations. The findings from the simulation study are also confirmed by the empirical analysis on financial data. We show that for more extreme target quantiles, the MLqE achieves a superior performance in terms of generalization error, when the distortion parameter q is chosen to be smaller than 1.
Even if the arbitrariness of the choice of q could be one of the main criticisms of the new method, we believe that the main strength of the MLqE is the flexibility gained from the choice of such a parameter. Further work is needed on this issue. Currently, two research directions are under investigation on the choice of q: (i) theoretical derivation of optimal values of q based on asymptotic theory, and (ii) data-driven regularization procedures such as cross-validation.

Delta Method Calculation

Consider α, $\widehat{\alpha}_{n,b}$ and $\widetilde{\alpha}_{n,b}$ defined as in Section 6.3. Moreover, let $x_B = B^{-1}\sum_{b=1}^{B}(\hat{\alpha}_{n,b}-\alpha)^2$ and $y_B = B^{-1}\sum_{b=1}^{B}(\widetilde{\alpha}_{n,b}-\alpha)^2$. By the central limit theorem, for large values of B we have that

$$\sqrt{B}\begin{bmatrix} x_B \\ y_B \end{bmatrix} \xrightarrow{\mathscr{D}} N\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right), \tag{6.6.1}$$

where $\mu_1 = MSE\left(\hat{\alpha}_n\right)$ and $\mu_2 = MSE\left(\widetilde{\alpha}_n\right)$. We are interested in the limiting distribution of $g(x_B, y_B) = x_B/y_B$ when $B \to \infty$. By the Delta Method (e.g., see

Ferguson (1996)) we have that

$$\sqrt{B}\ g(x_B, y_B) \xrightarrow{\mathscr{D}} N\left(g(\boldsymbol{\mu}), \nabla g(\boldsymbol{\mu})^T \boldsymbol{\Sigma}\ \nabla g(\boldsymbol{\mu})\right), \text{ as } B \to \infty \tag{6.6.2}$$

where $\nabla g(\cdot)$ is the gradient vector. In this case we have that

$$\nabla g(\boldsymbol{\mu})^T = \left(\frac{\partial}{\partial \mu_1} g(\boldsymbol{\mu}), \frac{\partial}{\partial \mu_2} g(\boldsymbol{\mu})\right)^T = \left(\frac{1}{\mu_1}, -\frac{\mu_1}{\mu_2^2}\right), \tag{6.6.3}$$

and

$$\begin{aligned} \nabla g(\boldsymbol{\mu})^T \boldsymbol{\Sigma}\ \nabla g(\boldsymbol{\mu}) &= \left(\frac{1}{\mu_1}, -\frac{\mu_1}{\mu_2^2}\right) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} 1/\mu_1 \\ -\mu_1/\mu_2^2 \end{pmatrix} \\ &= \frac{\sigma_{11}}{\mu_2^2} - 2\sigma_{12}\frac{\mu_1}{\mu_2^3} + \sigma_{22}\frac{\mu_1^2}{\mu_2^4}. \end{aligned}$$

Therefore, we obtained that

$$\sqrt{B}\begin{pmatrix} x_B \\ y_B \end{pmatrix} \xrightarrow{\mathscr{D}} N\left(\frac{\mu_1}{\mu_2}, \frac{\sigma_{11}}{\mu_2^2} - 2\sigma_{12}\frac{\mu_1}{\mu_2^3} + \sigma_{22}\frac{\mu_1^2}{\mu_2^4}\right), \text{ as } B \to \infty. \tag{6.6.4}$$

Chapter 7

Discussion and areas for future research

7.1 Conclusions

In statistics, a large number of inferential techniques rely, often implicitly, on the notion of information introduced by Shannon (1948) and on its companion divergence measure, the KL divergence (Kullback and Leibler, 1951). The underlying idea of our work is that inferential procedures based on measures of information that generalize the postulates of Shannon entropy can bring novel statistical properties. In particular, we consider the information measure proposed by Havrda and Charvát (1967) and studied later by Tsallis in thermodynamics (e.g., see Tsallis (1988); Tsallis et al. (1998)), which relaxes the postulate of strong additivity of Shannon entropy. The assumption of strong additivity is critical as it shapes completely the logarithmic form of Shannon entropy (see Chapter 2, Theorem 2.1.1).

Our primary aim is to gain more understanding on how dropping strong additivity can be exploited in statistical inference. In this thesis, we begin to explore these issues within a standard parametric framework. We introduced the Maximum Lq-Likelihood method, a novel estimation procedure based on the empirical version of the Havrda-Charvát-Tsallis entropy. The behavior of the resulting estimator – the MLqE – has been analyzed using asymptotic theory, numerical simulations

and real data. In what follows, we discuss our findings on the properties of the MLqE keeping in mind three main aspects: (i) efficiency; (ii) robustness and (iii) dimensionality of the parameter space.

7.1.1 Efficiency and the trade-off between bias and variance

In Chapter 3, we examined the MLqE in the context of exponential families for estimating parameters, tail probabilities and quantiles. The latter two tasks are carried out using the plug-in approach.

One important finding from our study is the connection between sample size and the distortion parameter q. In our asymptotic calculations, we explicitly take into account such a relationship and consider q as a sequence depending on n (see Chapter 3). For estimating the parameters, when $q_n \to 1$, the asymptotic variance of MLqE is shown to be equivalent to that of MLE. However, if we properly choose a (deterministic) q, a benefit from variance reduction by introducing a small bias is observed. When the sample size is small or moderate, the MLqE actually trades bias for variance, obtaining an overall reduction of the MSE, sometimes dramatically. When the sample size is large and q tends to 1, a necessary and sufficient condition to ensure a proper asymptotic normality and efficiency of MLqE is established. For estimating the tail probability, we studied explicitly the relationship between the size of the probability to be estimated and the distortion parameter, showing that for a proper choice of q the variance of MLqE can be much smaller than that of MLE.

The behavior of the MLqE for finite samples is explored using Monte Carlo simulations for various settings including: multivariate normal, GLMs with covariates and exponential tail probability estimation. When estimating the model parameters, in all the situations we observe a relationship between q and the sample size. In particular, we see a clear improvement in terms of MSE with respect to the classical MLE, due to variance reduction. Moreover, when considering a multivariate parameter space, the ratio of the mean square error of the MLE over that of MLqE increases with the number of parameters. Often, the magnitude of

the improvements is large and good performances are obtained in presence of high dimensionality and small sample size n.

Finally, it is important to remark that the reduction of MSE given by the MLqE does not apply indistinctly to all parametric functions. In some cases, the performance does not improve over the MLE. An example is the multivariate normal distribution, where a remarkable gain in accuracy is obtained when estimating the covariance matrix, but little change with respect the MLE is observed for the mean vector.

On the distortion parameter q. Clearly, the choice of q_n is important as it determines the finite sample performance of MLqE. When the focus is on estimating the parameters in an exponential family, we found that $|q_n - 1|n^{1/2} \to 0$ gives the "right" asymptotic normality. In our examples and simulations, choices of q_n with $|1 - q_n|$ between $1/n$ and $1/\sqrt{n}$ usually are seen to produce improvements over the MLE. However, typically only one direction of distortion (i.e., $q > 1$ or $q < 1$) can reduce the variance (and improve the MSE). For a given family, the expression of the asymptotic variance of the estimator can be used to find the right direction of distortion. For tail probability estimation, the optimal choice of q_n does not appear to be a characteristic of the family, but it depends on the function to be estimated. This is clear in our example about the exponential distribution where for the tail q needs to be smaller than one, whereas for estimating the parameter itself we need $q > 1$.

7.1.2 Robustness

In Chapter 5, we focus on the case where q is fixed and give an asymptotic theory to general classes of parametric distributions. In particular, using tools from M-estimation theory, we give results of convergence in probability and in distribution for the MLqE, providing expressions for the asymptotic variance. Within that framework, we examined the situation where the assumed model is affected by outlying observations.

The nature of the MLqE in this situation is best understood by studying its relationship with other familiar divergence measures usually employed in robust

estimation. Actually, we observed that computing the MLqE minimizes the discrepancy between the empirical distribution of the data and a target distribution by means of the general family of divergences indexed by the parameter q, sometimes referred to as power divergences (Cressie and Read, 1984). For special values of q, the power divergence is: (i) KL divergence ($q \to 1$), (ii) Hellinger distance ($q = 1/2$) and (iii) Pearson's Chi-Square divergence ($q = 2$). The first corresponds to maximum likelihood estimation, whereas the latter two are often employed for robust inference.

The MLqE exhibits strong performance in presence of perturbations with respect to the assumed model, conditional on a proper choice of the distortion parameter. Furthermore, this is usually achieved with little loss in efficiency. In this context, q serves for tuning the trade-off between robustness and efficiency. For fixed q, the methodology can be regarded as a robust extension of maximum likelihood estimation for which $q \to 1$. Choices of q near 1 afford considerable robustness while retaining efficiency close to that of maximum likelihood. In this sense, our method shares some of the strengths of other methods such as the Density Power Divergence estimator (DPDE) of Basu et al. (1998). However, whereas the MLqE provides a smooth bridge between Hellinger distance and KL divergence minimization, the DPDE provides a continuum between ℓ_2-norm and KL-divergence.

In our approach, we put forward an important innovation over other commonly used divergence-based estimators, such as the Hellinger distance estimator. For continuous models, Beran's minimum Hellinger distance estimator (Beran, 1977) and other similar methodologies (Lindsay, 1994; Bhandari et al., 2006), involve some degree of nonparametric analysis, with all the complications related to bandwidth selection. In higher dimensions, the bandwidth choice becomes troublesome with serious consequences on the reliability of the estimates. Instead, our method is shown to rely exclusively on the empirical distribution of the data avoiding the need of approximating the "true density" by a nonparametric smoothed version.

7.1.3 Computational aspects

Besides the appealing features related to variance reduction, our methodology has a relevant perk: an easy-to-implement and fast algorithm is readily available for computing the estimates (Chapter 5). Recall that the estimating equations are simply obtained by replacing the logarithm of log-likelihood function in the usual maximum likelihood procedure by the distorted logarithm $L_q(u) = (u^{1-q} - 1)/(1 - q)$. Clearly, the solution of the resulting estimating equations can be nontrivial due to nonlinearity. This issue is crucial when dealing with more complex multiparameter models.

However, the computational complexity can be considerably reduced by observing that the estimation equations are just a weighted version of the familiar score function, with weights dependent on the parameters (see Chapters 3 and 5). Based on this observation, we suggested an algorithm relying on a straightforward re-weighting strategy. The steps of the algorithm sometimes reduce to a simple variable transformation, depending on the weights. In all the considered applications, the algorithm performed extremely well and usually achieved convergence in a few steps. This is due to the fact that when the empirical distribution of the data is close to the true model, the algorithm is approximately linear.

From this perspective, we found points of contact with other techniques such as the robustification method introduced by Windham (1995), the empirical model tilting of Choi et al. (2000) or the bias-correction to maximum likelihood of Firth (1993).

7.2 Current work, possible developments and open problems

In the rest of the chapter, we consider an application of the MLqE in finance, which is currently under development. Next, we present a set of research questions closely related to this thesis and currently under investigation mainly concerning: (i) criteria for the choice of q and (ii) approximations to the bias/variance and to the distribution of MLqE.

7.2.1 An application of MLqE currently under study

The bottom line in many financial analyses is the basic issue of modeling a set of multivariate data. In many situations the data are characterized by their fat tails containing some proportion of extreme observations. The MLq methodology is apt to capture these main characteristics, and to provide a good fit for the bulk of the data. The weights computed by the re-weighting strategy presented in Chapter 5 provide useful information about the probability of the occurrence of atypical observations.

Consider a vector of returns of p risky assets $X^{\intercal} = (X_1, \ldots, X_p)$, following a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A portfolio is defined as random variable $Y = \mathbf{w}'X$, where $\mathbf{w}$ is a p-dimensional vector of real numbers such that $\sum_{j=1}^{p} w_j = 1$, to be called the investment strategy. Clearly, Y has mean $\mathbf{w}'\boldsymbol{\mu}$ and covariance matrix $\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$. Given observations of the random vector X, we are interested in selecting an investment strategy that balances the trade-off between the mean and the variance of the portfolio. The goal is usually stated in terms the following constrained optimization problem:

$$\underset{\mathbf{w}\in\mathbf{R}^p}{\arg\max} \left\{\gamma^{-1}\mathbf{w}'\widehat{\boldsymbol{\mu}} - \mathbf{w}'\widehat{\boldsymbol{\Sigma}}\mathbf{w}\right\}, \text{ subject to } \sum_{j=1}^{p} w_j = 1, \tag{7.2.1}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are MLqE estimates of the true mean vector and covariance matrix and $\gamma > 0$ is the risk aversion constant. Thus, the estimation procedure involves two stages: (i) estimation of the parameters and (ii) computing optimal weights. Note that if $\gamma \to \infty$ (infinite risk aversion), the above minimization task is equivalent to the minimization of the covariance matrix, subject to the constraints on the weights. Tipically, observations of X are assumed to be independently drawn from one model of reference (the multivariate normal model is the standard choice).

In this context, we expect the MLqE to be appealing and produce significant improvements with respect to other robust techniques commonly employed in finance. In particular, we believe that our approach can produce significant advantages when p is large and where methods such as Hellinger distance estimation can fail due to the complexity of the choice of the bandwidth selection.

7.2.2 Finite sample size and criteria for selecting q

Although we gained insight on the role of q and outlined possible criteria for its choice in various estimation settings, more work is needed in this direction. In particular, it is our feeling that a more in-depth theoretical analysis of the connection between q and the sample size is desirable. In general, for the choice of q, we must distinguish the following scenarios:

(i) Estimation of the parameters with no outliers.

(ii) Estimation of tail probability or quantiles with no outliers.

(iii) Estimation of the parameters and tail probabilty in presence of outliers.

For (i) and (ii), an accurate tuning of the distortion parameter is a key issue: a good criterion can lead to the MLqE outperforming the MLE. We found that relying on large sample approximations can be a reasonable strategy. Choices of q based on first-order approximations lead to minimization of the asymptotic MSE. Our simulations and analyses of real data showed that this choice is reasonable even for small samples and but yet is still rather conservative. Certainly, room for improvement is available in this direction.

In case (iii), the choice of q does not need to necessarily rely on delicate asymptotics as for (i) and (ii). In presence of outliers, the issues related to efficiency must be considered at a more coarse-grained level. In Chapter 5 it is seen that even a criterion based on a rough empirical upper bound for relative efficiency works reasonably well. In this context, a more important contribution may be given by a thorough understanding of the relationship between power divergences and the MLqE and in particular by studying the role of the MLqE corrected for its asymptotic bias.

We believe that the following research lines can be helpful for constructing new criteria for the choice of q: (a) higher-order approximations to the asymptotic distribution of the MLqE (or at least to its moments) and (b) finite-sample approximation of bias and variance. Currently, part of our efforts is directed towards studying higher-order asymptotic approximations of bias, variance and distribution of MLqE. In this context, we found that an extremely powerful tool is represented

by the class of saddle-point methods. More insight about this topic and its possible consequences on our research will be provided in the next paragraph.

Higher-order approximations. When q is a sequence depending on n, our asymptotic calculations have shown that the asymptotic variance of MLqE is equivalent to that of MLE. However, all of our numerical results indicate that, for small samples, the accuracy of the MLqE is superior. In order to fill this gap, the development of higher-order asymptotics for the MLqE will be valuable. In this direction, approximations based on the saddle-point method are powerful tools for obtaining accurate expressions for densities and distribution functions.

Saddlepoint methods have been used for approximating the density of the sample mean of i.i.d. random variables and for approximating the density of the MLE in exponential families (see Barndorff-Nielsen (1983); Barndorff-Nielsen and Cox (1994) or Field and Ronchetti (1990)). For example, the approximated density of the MLE for i.i.d. observations generated from a distribution with univariate parameter θ_0 is:

$$f_n(\widehat{\theta}_n|\theta_0) \approx \left(\frac{nA''(\widehat{\theta}_n)}{2\pi}\right)^{1/2} \exp\Big\{(\theta_0 - \widehat{\theta}_n)B(\widehat{\theta}_n) - n[A(\theta_0) - A(\widehat{\theta}_n)]\Big\}, \quad (7.2.2)$$

where B is the sufficient statistic $B = \sum_i b(X_i)$ and A is the cumulant generating function as defined in Chapter 3. The saddle point method yields extremely accurate approximations, with error up to order n^{-1}, uniformly. The error can be further improved to be of order $n^{-3/2}$ when a proper normalization is applied the approximated density (Field and Ronchetti, 1990). In Fig. 7.1, we show the approximated distribution of a MLE for the rate of an exponential distribution. The exact distribution is also plotted for comparison. When $n = 5$ the exact and approximated curves are practically identical.

General results have been obtained for M-estimation, which can prove to be helpful for the case of MLqE. In that context, saddlepoint approximation have been investigated by Hampel (1973), Field and Hampel (1982), and Daniels (1997). Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random vectors with common distribution function and that we are interested in the density f_n

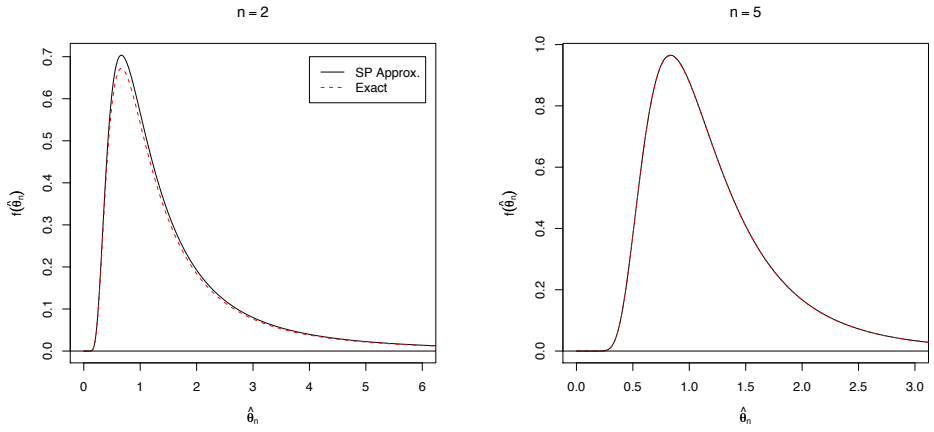


Figure 7.1: Saddle point approximation of the density of the MLE of the rate of an Exp(1). The exact distribution of the MLE (dashed lines) is plotted for comparison.

of a statistic $T_n \in \mathbb{R}^p$, $p \geq 2$, which satisfies the system of equations of the form $\sum_{i=1} \psi(X_i, T_n) = 0$. Field (1982) gave a formula for the the saddlepoint approximation of f_n, say $\widehat{g}_n$, showing that under certain regularity condition

$$f_n = g_n(1 + O(n^{-1})), \quad \text{as} \quad n \to \infty. \tag{7.2.3}$$

As for the MLE, also in this more general case, considering a normalized version of g_n can improve the above approximation to be of order $n^{-3/2}$.

Finite sample approximations. Another useful line of research is represented by working on accurate approximations of finite sample variance/bias of MLqE. Currently, finite sample approximations for M-estimators appears to be a rather challenging topic. However, useful tools can by provided by the advances in empirical process theory (see e.g., Pollard (1990); Van der Vaart and Wellner (1996); Koul and Koul (2002)). Possibly, recent understanding of the behavior of the empirical measure for finite sample sizes may help to derive reliable approximations to the finite sample bias and variance of the MLqE.

An other important consideration for both small sample approximations and asymptotic derivation regards the specific form of the variance of the MLqE com-

pared to that of MLE. In all our examples, we observed that the ratio between the two is a function of q that does not dependent on the true value of parameters. We feel that this issue deserves a more detailed treatment and general results for both $n \to \infty$ and finite samples will be valuable. In particular, if we had reason to say that this fact holds true even in finite samples, for a given family of distributions, q could be computed numerically simply by Monte Carlo simulations.

Finally, although in this chapter we focused on the specific problem of the choice of q, there are other new streams of research that can naturally derive from our work. For instance, at its current state the methodology requires distributional specifications. An interesting development of the estimation method can be obtained by considering a nonparametric Lq-likelihood function that does not rely on distributional assumptions, but otherwise analogous to its parametric counterpart. For instance, similarly to Owen (1990, 1991, 2001), one can consider optimizing the empirical Lq-likelihood function subject to moment conditions. In a different direction, a research ideas can be obtained by using the q-entropy for modeling general dependence structures in a dataset. In this sense, for example, the q-entropy can be employed for developing measures for testing independence.

The statistical usage of generalized information measures opens a number of interesting and challenging problems. Hopefully, the methodology proposed in this work can act as a catalyst for new understanding of the implications that general characterization of information have on statistical inference.

Bibliography

Abe, S. (2001). *Nonextensive Statistical Mechanics and Its Applications*. Springer-Verlag Inc.

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

Aczél, J. D. and Z. Daróczy (1975). On measures of information and their generalizations. *Publications Mathematicae 10*, 171–190.

Agostinelli, C. (2002). Robust testing hypotheses via weighted likelihood function. *Statistica (Bologna) 62*(1), 87–110.

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

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 716–723.

Altun, Y. and A. Smola (2006). Unifying divergence minimization and statistical inference via convex duality. *in COLT*.

Andricioaei, I. and J. E. Straub (1996). Generalized simulated annealing algorithms using tsallis statistics: Methodology, optimization and applications to atomic clusters. *Physical Review 53*, R3055.

Anteneodo, C., C. Tsallis, and A. S. Martinez (2002). Risk aversion in economic transactions. *Europhys. Lett*, 635.

Arndt, C. (2001). *Information Measures: Information and Its Decription in Science and Engineering*. Springer-Verlag Inc.

Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1999). Coherent measures of risk. *Mathematical Finance 9*, 203–228.

Balkema, A. A. and L. de Haan (1974). Residual life time at great age. *The Annals of Probability 2*, 792–804.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*, 343–365.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and Asymptotics*. Chapman & Hall Ltd.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *Special Commemorative Issue: Information Theory: 1948-1998 IEEE. Trans. Inform. Theory*, 2743–2760.

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

Beck, C. and F. Schlögl (1993). *Thermodynamics of Chaotic Systems: An Introduction*. Cambridge, England: Cambridge University Press.

Bell, A. (2003). Asymptotic theory of weighted maximum likelihood estimation for growth models. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003))*.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics 5*, 445–463.

Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli 12*(2), 299–331.

Bhandari, S. K., A. Basu, and S. Sarkar (2006). Robust inference in parametric models using the family of generalized negative exponential disparities. *Australian & New Zealand Journal of Statistics 48*(1), 95–114.

Borland, L., A. Plastino, and C. C Tsallis (1998). Information gain within nonextensive thermostatistics. *Journal of Mathematical Physics 39*, 6490.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*, 353–360.

Brooks, C., J. Clare, J. Dalla Molle, and G. Persand (2005). A comparison of extreme value theory approaches for determining value at risk. *Journal of Empirical Finance 22*, 1–22.

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

Castillo, E., J. María, and A. S. Hadi (1997). Fitting continuous bivariate distributions to data. *The Statistician: Journal of the Institute of Statisticians 46*, 355–369.

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

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1,2*, 223–236.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. Wiley Series in Telecommunications.

Cressie, N. and T. R. C. Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B: Methodological 46*, 440–464.

Csiszár, I. (1967). Information-type measures and difference of probability distributions and indirect observations. *Studia Scientiarum Matematicarum Hungarica 2*, 299–318.

Daniels, H. E. (1997). Saddlepoint approximations in statistics (Pkg: P171-200). In S. a. Kotz and N. L. a. Johnson (Eds.), *Breakthroughs in Statistics, Volume III*, pp. 177–200. Springer-Verlag Inc.

Duarte-Queiros, S. M., C. Anteneodo, and C. Tsallis (2005). *Power-law distributions in economics: a nonextensive statistical approach, in Noise and Fluctuations in Econophysics and Finance, Proc. of SPIE 5848, 151*. SPIE, Bellingham, WA.

Ebank, B., P. Sahoo, and S. Wolfgang (1998). *Characterizations of Information Measures*. World Scientific.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). *Modelling extremal events for Insurance and Finance*. Applications of Mathematics. Springer.

Fan, J. and I. Gijbels (Eds.) (1996). *Local Polynomial Modeling and its Applications*. London: Chapman & Hall.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall Ltd.

Ferrari, D. and Y. Yang (2007). Estimation of tail probability via the maximum Lq-likelihood method. *Technical Report 659, School of Statistics, University of Minnesota*.

Field, C. and E. Ronchetti (1990). *Small Sample Asymptotics*. Institute of Mathematical Statistics.

Field, C. A. and F. R. Hampel (1982). Small-sample asymptotic distributions of M-estimators of location. *Biometrika 69*, 29–46.

Firth, D. (1993). Bias reduction of maximum likelihood estimates (Corr: 95V82 p667). *Biometrika 80*, 27–38.

Fisher, R. and L. C. Tippett (1928). Limiting forms of the frequency distribution of largest or smallest member of a sample. *Proceedings of the Cambridge philosophical society 24*, 180–190.

Furuichi, S. (2005). On uniqueness theorems for Tsallis entropy and Tsallis relative entropy. *Information Theory, IEEE Transactions on 51*, 3638–3645.

Gell-Mann, M. (Ed.) (2004). *Nonextensive Entropy, Interdisciplinary Applications*. New York: Oxford University Press.

Gell-Mann, M. and S. Lloyd (1996). Information measures, effective complexity, and total information. *Complexity 2*(1), 44–52.

Gilli, M. and E. Kellezi (2006). An application of extreme value theory for measuring financial risk. *Computational Economics 2-3*, 207–228.

Givens, G. H. and J. A. Hoeting (2005). *Computational Statistics*. New Jersey: Wiley.

Gnedenko, B. V. (1943). Sur la distribution limite du terme d'une serie aleatoire. *Annals of Mathematics 44*, 423–453.

Goldfarb, D. (1970). A family of variable-metric method derived by variational means. *Mathematics of Computation 24*, 23–26.

Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics 35*, 185–191.

Hampel, F. R. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 27*, 87–104.

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

Hand, D., F. Daly, A. Lunn, K. McConway, and E. Ostrowski (1994). *A Handbook of Small Data Sets*. London: Chapman and Hall.

Hastie, T. and R. Tibshirani (1987). Non-parametric logistic and proportional odds regression. *Applied Statistics 36*, 260–276.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-color Illustrations*. Springer-Verlag Inc.

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

Hosking, J. R. M. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics 29*, 339–349.

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

Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85–98.

Huber, P. J. (1981a). *Robust Statistics*. Wiley Series in Probability. John Wiley and Sons.

Huber, P. J. (1981b). *Robust Statistics*. John Wiley & Sons.

Jakulin, A. and T. Bratko (2003, sep). Analyzing attribute dependencies. In N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski (Eds.), *Proc. of Principles of Knowledge Discovery in Data (PKDD)*, Volume 2838 of *LNAI*, pp. 229–240. Springer-Verlag.

Jakulin, A. and T. Bratko (2004). Quantifying and visualizing attribute interactions: An approach based on entropy.

James, W. and C. Stein (1961). Estimation with quadratic loss. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 361–379. University of California Press.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev. 106*, 620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Phys. Rev. 108*, 171.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society 81*, 158–172.

Johnson, L. W. and R. D. Riess (1982). *Numerical Analysis*. Reading: Addison-Wesley.

Kinchin, A. J. (1953). *The concept of entropy in the theory of probability, in Mathematical Foundations of Information theory*, Volume 22. New York: Dover.

Knight, J., S. Satchell, and G. Wang (2003). Value at risk linear exponent (varlinex) forecasts. *Quantitative Finance 3*, 332–344.

Koul, H. L. and H. L. Koul (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*. Springer-Verlag Inc.

Kuester, K., S. Mittnik, and M. Paolella (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics 4,1*, 53–89.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Lazar, N. A. (2004). *Statistics of Extremes: Theory and Applications*. England: Wiley.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation.* Springer-Verlag Inc.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics 22*, 1081–1114.

Magnus, J. R. and H. Neudecker (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics 7*, 381–394.

Markatou, M., A. Basu, and B. G. Lindsay (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association 93*, 740–750.

McCulloch, C. E. (1982). Symmetric matrix derivatives with applications. *Journal of the American Statistical Association 77*, 679–682.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika 19*, 97–116.

McNeil, A. and R. Frey (2000). Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach. *Journal of Empirical Finance 7*, 271–300.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance, New Jersey.

McNeil, A. and A. Stephenson (2007). *evir: Extreme Values in R*. R package version 1.5. S original (EVIS) by Alexander McNeil and R port by Alec Stephenson.

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

Newton, M. and A. Raftery (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of Royal Statistical Society Series B 56*, 3–48.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*, 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics 19*, 1725–1747.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall Ltd.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. New York: Chapman and Hall.

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

Plastino, A. and A. R. Plastino (1999). Tsallis entropy and Jaynes' information theory formalism. *Brazilian Journal of Physics 29*(1).

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics.

Prakasa Rao, B. L. S. (1991). Asymptotic theory of weighted maximum likelihood estimation for growth models. In N. U. Prabhu and I. V. Basawa (Eds.), *Statistical Inference in Stochastic Processes*, pp. 183–208. Marcel Dekker Inc.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rajagopal, A. and S. Abe (2002). Statistical mechanical foundations of power-law distributions. *Invited talk at Anomalous Distributions, Nonlinear Dynamics and Nonextensivity, Santa Fe, USA, November 6-9*.

Rao, C. R. (1994). Criteria of estimation in large samples. In *Selected papers of C.R. Rao, Vol. 2*, pp. 331–352. Wiley Eastern Ltd.

Reiss, R. and M. Thomas (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Basel: Birkhauser Verlag.

Rényi, A. (1961). On measures of entropy and information. *Proc. 4th Berk. Symp. Math. Statist. and Prob. 1*, 547–461.

Ribatet, M. A. (2006). A users guide to the pot package (version 1.0).

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics 43*(3), 274–285.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 197–206. University of California Press.

Suyari, H. (2004). Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for nonextensive entropy. *IEEE Trans. Inf. Theory 50*, 1783–1787.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*, 559–567.

Tsallis, C. (1988, July). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics 52*(1-2), 479–487.

Tsallis, C., R. S. Mendesc, and A. R. Plastino (1998). The role of constraints within generalized nonextensive statistics. *Physica A: Statistical and Theoretical Physics 261*, 534–554.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag Inc.

von Mises, R. (1954). *La distribution de la plus grande de n valeurs, in Selected Papers, Volume II*, Volume 44. American Mathematical Society, Providence, RI.

Wang, X., C. van Eeden, and J. V. Zidek (2004). Asymptotic properties of maximum weighted likelihood estimators. *Journal of Planning and Statistical Inference 119*(2), 37–54.

Wang, X. and J. V. Zidek (2005a). Derivation of mixture distributions and weighted likelihood function as minimizers of KL-divergence subject to constraints. *Annals of the Institute of Statistical Mathematics 57*(4), 687–701.

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

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