Maximum likelihood characterization of distributions

MITIA DUERINCKX[1,*], CHRISTOPHE LEY[1,**] and YVIK SWAN[2]

[1]*Département de Mathématique, Université libre de Bruxelles, Campus Plaine CP 210, Boulevard du triomphe, B-1050 Bruxelles, Belgique. E-mail:* [*]*mitia.duerinckx@ulb.ac.be;* [**]*christophe.ley@ulb.ac.be*
[2]*Université du Luxembourg, Campus Kirchberg, UR en mathématiques, 6, rue Richard Coudenhove-Kalergi, L-1359 Luxembourg, Grand-Duché de Luxembourg. E-mail:* *yvik.swan@uni.lu*

A famous characterization theorem due to C.F. Gauss states that the maximum likelihood estimator (MLE) of the parameter in a location family is the sample mean for all samples of all sample sizes if and only if the family is Gaussian. There exist many extensions of this result in diverse directions, most of them focussing on location and scale families. In this paper, we propose a unified treatment of this literature by providing general MLE characterization theorems for one-parameter group families (with particular attention on location and scale parameters). In doing so, we provide tools for determining whether or not a given such family is MLE-characterizable, and, in case it is, we define the fundamental concept of *minimal necessary sample size* at which a given characterization holds. Many of the cornerstone references on this topic are retrieved and discussed in the light of our findings, and several new characterization theorems are provided. Of particular interest is that one part of our work, namely the introduction of so-called *equivalence classes* for MLE characterizations, is a modernized version of Daniel Bernoulli's viewpoint on maximum likelihood estimation.

Keywords: location parameter; maximum likelihood estimator; minimal necessary sample size; one-parameter group family; scale parameter; score function

1. Introduction

In probability and statistics, a characterization theorem occurs whenever a given law or a given class of laws is the only one which satisfies a certain property. While probabilistic characterization theorems are concerned with distributional aspects of (functions of) random variables, statistical characterization theorems rather deal with properties of *statistics*, that is, measurable functions of a set of independent random variables (observations) following a certain distribution. Examples of probabilistic characterization theorems include:

- Stein-type characterizations of probability laws, inspired from the classical Stein and Chen characterizations of the normal and the Poisson distributions, see Stein [43], Chen [10];
- maximum entropy characterizations, see Cover and Thomas [12], Chapter 11;
- conditioning characterizations of probability laws such as, for example, determining the marginal distributions of the random components X and Y of the vector $(X, Y)'$ if only the conditional distribution of X given $X + Y$ is known, see Patil and Seshadri [37].

The class of statistical characterization theorems includes:

- characterizations of probability distributions by means of order statistics, see Galambos [19] and Kotz [29] for an overview on the vast literature on this subject;
- Cramér-type characterizations, see Cramér [13];
- characterizations of probability laws by means of one linear statistic, of identically distributed statistics or of the independence between two statistics, see Lukacs [33].

Besides their evident mathematical interest per se, characterization theorems also provide a better understanding of the distributions under investigation and sometimes offer unexpected handles to innovations which might not have been uncovered otherwise. For instance, Chen and Stein's characterizations are at the heart of the celebrated *Stein's method*, see the recent Chen, Goldstein and Shao [11] and Ross [42] for an overview; maximum entropy characterizations are closely related to the development of important tools in information theory (see Akaike [2,3]), Bayesian probability theory (see Jaynes [25]) or even econometrics (see Wu [48], Park and Bera [36]; characterizations of probability distributions by means of order statistics are, according to Teicher [46], "harbingers of [...] characterization theorems", and have been extensively studied around the middle of the twentieth century by the likes of Kac, Kagan, Linnik, Lukacs or Rao; Cramér-type characterizations are currently the object of active research, see Bourguin and Tudor [6]. This list is by no means exhaustive and for further information about characterization theorems as a whole, we refer to the extensive and still relevant monograph Kagan, Linnik and Rao [27], or to Kotz [29], Bondesson [5], Haikady [22] and the references therein.

In this paper, we focus on a family of characterization theorems which lie at the intersection between probabilistic and statistical characterizations, the so-called *MLE characterizations*.

1.1. A brief history of MLE characterizations

We call MLE characterization the characterization of a (family of) probability distribution(s) via the structure of the Maximum Likelihood Estimator (MLE) of a certain parameter of interest (location, scale, etc.).

The first occurrence of such a theorem is in Gauss [20], where Gauss showed that the normal (a.k.a. the Gaussian) is the only location family for which the sample mean $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ is "always" the MLE of the location parameter. More specifically, Gauss [20] proved that, in a location family $g(x - \theta)$ with differentiable density g, the MLE for θ is the sample mean for all samples $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$ of all sample sizes n if, and only if, $g(x) = \kappa_\lambda e^{-\lambda x^2/2}$ for $\lambda > 0$ some constant and κ_λ the adequate normalizing constant. Discussed as early as in Poincaré [38], this important result, which we will throughout call *Gauss' MLE characterization*, has attracted much attention over the past century and has spawned a spree of papers about MLE characterization theorems, the main contributions (extensions and improvements on different levels, see below) being due to Teicher [46], Ghosh and Rao [21], Kagan, Linnik and Rao [27], Findeisen [18], Marshall and Olkin [34] and Azzalini and Genton [4]. (See also Hürlimann [24] for an alternative approach to this topic.) For more information on Gauss' original argument, we refer the reader to the accounts of Hald [23], pages 354 and 355, and Chatterjee [9], pages 225–227. See Norden [35] or the fascinating Stigler [45] for an interesting discussion on MLEs and the origins of this fundamental concept.

The successive refinements of Gauss' MLE characterization contain improvements on two distinct levels. Firstly, several authors have worked towards weakening the regularity assumptions on the class of distributions considered; for instance, Gauss requires differentiability of g, while Teicher [46] only requires continuity. Secondly, many authors have aimed at lowering the sample size necessary for the characterization to hold (i.e., the "always"-statement); for instance, Gauss requires that the sample mean be MLE for the location parameter for *all* sample sizes simultaneously, Teicher [46] only requires that it be MLE for samples of sizes 2 and 3 at the same time, while Azzalini and Genton [4] only need that it be so for a single fixed sample size $n \geq 3$. Note that Azzalini and Genton [4] also construct explicit examples of non-Gaussian distributions for which the sample mean is the MLE of the location parameter for the sample size $n = 2$. We already draw the reader's attention to the fact that Azzalini and Genton's [4] result does not supersede Teicher's [46], since the former require more stringent conditions (namely differentiability of the g's) than the latter. We will provide a more detailed discussion on this interesting fact again at the end of Section 4.

Aside from these "technical" improvements, the literature on MLE characterization theorems also contains evolutions in different directions which have resulted in a new stream of research, namely that of discovering new (i.e., different from Gauss' MLE characterization) MLE characterization theorems. These can be subdivided into two categories. On the one hand, MLE characterizations with respect to the location parameter but for other forms than the sample mean have been shown to hold for densities other than the Gaussian. On the other hand, MLE characterizations with respect to other parameters of interest than the location parameter have also been considered. Teicher [46] shows that if, under some regularity assumptions and for *all* sample sizes, the MLE for the scale parameter of a scale target distribution is the sample mean $\bar{x}$ then the target is the exponential distribution, while if it corresponds to the square root of the sample arithmetic mean of squares $(\frac{1}{n} \sum_{i=1}^{n} x_i^2)^{1/2}$, then the target is the standard normal distribution. Following suit on Teicher's work, Kagan, Linnik and Rao [27] establish that the sample median is the MLE for the location parameter for all samples of size $n = 4$ if and only if the parent distribution is the Laplace law. Also, in Ghosh and Rao [21], it is shown that there exist distributions other than the Laplace for which the sample median at $n = 2$ or $n = 3$ is MLE. Ferguson [17] generalizes Teicher's [46] location-based MLE characterization from the Gaussian to a one-parameter generalized normal distribution, and Marshall and Olkin [34] generalize Teicher's [46] scale-based MLE characterization of the exponential distribution to the gamma distribution with shape parameter $\alpha > 0$ by replacing $\bar{x}$ as MLE for the scale parameter with $\bar{x}/\alpha$.

There also exist contributions by Buczolich and Székely [7] where they investigate situations in which a weighted average of ordered sample elements can be an MLE of the location parameter. MLE characterizations in the multivariate setup have been proposed, inter alia, in Marshall and Olkin [34] and Azzalini and Genton [4]. In parallel to all these "linear" MLE characterizations there have also been a number of developments regarding MLE characterizations for spherical distributions, that is distributions taking their values only on the unit hypersphere in higher dimensions, see Duerinckx and Ley [16] and the references therein.

Finally, there exists a different stream of MLE characterization research, inspired by Poincaré [38], in which one relaxes the assumptions made on the role of the parameter θ and seeks to understand the conditions under which the MLE for this θ is an arithmetic mean. This

approach to MLE characterization theorems is known as *Gauss' principle*; quoting Campbell [8] "*by Gauss' principle we shall mean that a distribution should be chosen so that the maximum likelihood estimate of the parameter θ is the same as the arithmetic mean estimate given by* $n^{-1} \sum_{k=1}^{n} T(x_k)$", with $\theta = \mathrm{E}(T(X))$ for some known function T. When T is the identity function and one considers one-parameter location families then we recover the MLE characterization problem solved by Gauss [20]. For general one-parameter families the corresponding problem was solved in Poincaré [38]. Campbell [8] broadens Poincaré's conclusion to general functions T and to the multivariate setup, while Bondesson [5] further extends the above works to exponential families with nuisance parameters. As pointed out by their authors, many of these results remain valid for discrete distributions; for other MLE characterizations of discrete probability laws, we also refer to Puig [39] and Puig and Valero [41].

1.2. Applications of MLE characterizations

Perhaps the most remarkable application of MLE characterizations is to be found in the very origins of this field of study, whose forefathers sought to define families of probability distributions which were "natural" for a given important problem. For instance, the Gaussian distribution was uncovered by Gauss through his effort of finding the location family for which the sample mean $\bar{x}$ is a most probable value for θ, the location parameter. Similarly, Poincaré in his "Calcul des Probabilités" (2nd ed., 1912) derived in Chapter 10 (pages 147–168 in the 1896 edition) a particular case of the *exponential families of distributions* (see Lehmann and Casella [30], Section 1.5) by asking for which distributions $\bar{x}$ is the MLE of θ, without specifying the role of the parameter. We refer to the historical remarks at the end of Bondesson [5] and the references therein for more information on both the works of Gauss and Poincaré. Following Gauss' ideas, von Mises [47] defined the circular (i.e., spherical in dimension two) analogue of the Gaussian distribution by looking for the circular distribution whose circular location parameter always has the circular sample mean as MLE; this led to the now famous *Fisher–von Mises–Langevin* (abbreviated FvML) distribution on spheres.

Of an entirely different nature is Campbell's [8] use of MLE characterizations. In his paper, Campbell establishes an equivalence between MLE characterizations in the spirit of Gauss' principle and the minimum discrimination information estimation of probabilities. More recently, Puig [40] has applied MLE characterizations in order to characterize the Harmonic Law as the only statistical model on the positive real half-line that satisfies a certain number of requirements. As a last example, we cite the work of Ley and Paindaveine [32] who solved a long-standing problem on skew-symmetric distributions through an argument resting on Gauss' MLE characterization.

1.3. Purpose of the paper

As is perhaps intuitively clear, many of the results on MLE characterizations stem from a common origin. As a matter of fact many authors often follow the same "smart path" that can be summarized in three steps: (a) choose the role of the parameter of interest θ (location or scale);

(b) choose a remarkable form for the MLE for θ (e.g., sample mean, variance or median); (c) use the freedom of choice in the samples as well as the sample size (two samples of respective sizes 2 and 3, one sample of size 3, all samples of all sizes, ...) to obtain the largest class of distributions satisfying certain assumptions (continuous at a single point, continuous, differentiable, ...) which share this specific MLE. While similar, the arguments leading to the different results nevertheless are largely ad hoc and rest upon crafty manipulations of the explicit given form of the MLE. Moreover, step (c) contains assumptions on the minimal sample size and on the properties of the distributions being characterized, the necessity of which is barely addressed.

The purpose of the present paper is to unify this important literature by explaining the mechanism behind this "smart path" in the more general context of one-parameter *group families* (as defined in Section 6) although still with particular focus on location and scale families. In doing so, we will introduce the concept of minimal covering sample size (MCSS), a quantity whose value depends on the structure of the support of the target distribution and from which one deduces the a priori minimal necessary sample size (MNSS) at which an MLE characterization holds for a given family of distributions (see Duerinckx and Ley [16] where this notion was introduced). As we shall see, the MCSS and MNSS explain many of the differences in the "always"-statements appearing throughout the literature on MLE characterization. Moreover our unified perspective on MLE characterizations not only permits to identify the minimal sufficient conditions under which group families are characterized by their MLEs (MLE-characterizable) but also provide tools for (easily) constructing new MLE characterizations of many important families of distributions.

In a nutshell, our goal is to (i) propose a unified perspective on MLE characterizations for one-parameter group families, (ii) answer the question of which such families are MLE-characterizable and find their MNSS, (iii) retrieve, improve on and better understand existing results via our general analysis, and (iv) construct new MLE characterization results. Our contribution to the important literature on MLE characterizations unifies the cornerstone references Teicher [46], Ferguson [17], Kagan, Linnik and Rao [27], Marshall and Olkin [34], Azzalini and Genton [4] (to cite but these) and complements the understanding brought by the seminal works of Poincaré [38], Campbell [8] and Bondesson [5].

1.4. Outline of the paper

In Section 2 we describe the framework of our study, give all necessary notations and introduce the so-called *equivalence classes*. In Section 3 we establish and interpret the above-mentioned notion of MCSS which will be central to this paper. In Section 4, we derive the MLE characterization for univariate location families, while in Section 5 we proceed in a similar way with univariate scale families. In Section 6 we obtain MLE characterizations for general one-parameter group families, allowing us to study other roles of the parameter (e.g., skewness). In Section 7 we apply our findings to particular families of distributions. We conclude the paper with a discussion, in Section 8, of the different possible extensions that are yet to be explored.

2. Notations and generalities on ML estimators, equivalence classes

Throughout we consider observations $\mathbf{X}^{(n)} = (X_1, X_2, \ldots, X_n)$ that are sampled independently from a distribution P_θ^f (with density f) which we suppose to be entirely known up to a parameter $\theta \in \Theta \subset \mathbb{R}$. The true parameter value $\theta_0 \in \Theta$ is estimated by ML estimation on basis of $\mathbf{X}^{(n)}$. As explained in the Introduction, our aim consists in determining which classes of distributions are identifiable by means of the MLE of the parameter of interest θ, a parameter that can, in principle, be of any nature (i.e., location, scale, etc.). On the target family of distributions $\{P_\theta^f : \theta \in \Theta\}$ we make the following general assumptions:

- (A1) The parameter space Θ contains an open set Θ_0 of which the true parameter θ_0 is an interior point.
- (A2) For each $\theta \in \Theta_0$, the distribution P_θ^f has support S independent of θ.
- (A3) For all $1 \le i \le n$ the random variable X_i has a density $f(x_i; \theta)$ with respect to the (dominating) Lebesgue measure.
- (A4) For $\theta \neq \theta' \in \Theta_0$, we have $P_\theta^f \neq P_{\theta'}^f$.

These assumptions are taken from Lehmann and Casella [30], page 444.

Remark 2.1. Although typically we will be concerned with either *location families* with densities of the form $f(x; \theta) = f(x - \theta)$ for $\theta \in \mathbb{R}$ the location parameter, or *scale families* with densities of the form $f(x; \theta) = \theta f(x\theta)$ for $\theta \in \mathbb{R}_0^+$ the scale parameter, other roles for θ (skewness, tail behavior...) can also be considered; see Section 6 where we detail our approach for one-parameter group families.

Remark 2.2. While throughout the paper we restrict our attention to the univariate setting, it is easy to see that our arguments are in some cases transposable word-by-word to the multivariate case. We discuss this matter briefly in Section 8.

Remark 2.3. Assumption (A2) implies that only densities with full support $\mathbb{R}$ may be considered for ML estimation of a location parameter, while only densities with support either $\mathbb{R}$, $\mathbb{R}^+$ or $\mathbb{R}^-$ may be considered for ML estimation of a scale parameter. Despite the fact that these restrictions are natural in the present context they can, if deemed necessary, be lifted. We will briefly discuss this topic in Section 8.

We define, for a fixed sample size $n \geq 1$, the MLE of the parameter θ as (if it exists) the measurable function

$$\hat{\theta}_f^{(n)} : S^n := S \times \cdots \times S \to \Theta_0 : \mathbf{x}^{(n)} := (x_1, \ldots, x_n) \mapsto \hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})$$

for which

$$\prod_{i=1}^n f(x_i; \hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})) \geq \prod_{i=1}^n f(x_i; \theta) \tag{1}$$

for all $\theta \in \Theta_0$ and all samples $\mathbf{x}^{(n)} \in S^n$ of size n. It is not trivial to provide minimal conditions on f under which $\hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})$ exists, is uniquely defined and satisfies the necessary measurability conditions. Consistency of the MLE is also a delicate matter and further regularity conditions are required for the problem to make sense. As in Cramér [14,15] one may suppose that, for almost all x, the density $f(x; \theta)$ is differentiable with respect to θ. This allows to define the MLE as the solution of the local likelihood equation

$$\sum_{i=1}^{n} \varphi_f(x_i; \theta) = 0, \tag{2}$$

where

$$\varphi_f(x; \theta) := \frac{\partial}{\partial \theta} \log f(x; \theta)$$

is the *score function* of the density f associated with the parameter θ (we set this function to 0 outside the support of f). The solution to (2) has, at least asymptotically, the required properties (see Lehmann and Casella [30], Theorem 6.3.7). Furthermore, this way of proceeding allows for a simple sufficient condition for uniqueness of the MLE: the mapping $x \mapsto \varphi_f(x; \theta)$ has to be strictly monotone and to cross the x-axis. Note that this requirement coincides with *strong unimodality* or *log-concavity* of the density f when θ is a location parameter (see Lehmann and Casella [30], Exercise 6.3.15).

Remark 2.4. Although in general there is no explicit expression for the MLE of a given parametric family, there exist several important distributions that not only satisfy all the above requirements but also allow for MLEs which take on a remarkable form. Taking $f = \phi$ the standard normal density, the MLE for the location parameter is $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$, the sample arithmetic mean, while that for the scale parameter is $(\frac{1}{n}\sum_{i=1}^{n} x_i^2)^{1/2}$, the square root of the sample arithmetic mean of squares. Taking f the exponential density, the MLE for the scale parameter becomes $\bar{x}$.

As outlined in Section 1.3, our objective in this paper is to determine minimal conditions under which a given form of MLE for a given type of parameter identifies a specific probability distribution. Upon making this statement there immediately arises a trivial identification problem (due to our definition of maximum likelihood estimators) which we first need to evacuate. Indeed if $\hat{\theta}_f^{(n)}$ maximizes the f-likelihood function, then it also maximizes the g-likelihood function for any function $g = cf^d$ with $d > 0$ and c a normalizing constant. In fact, from our definitions (1) and (2), it is immediate that any two parametric densities $f(x; \theta)$ and $g(x; \theta)$ with same support S and such that

$$\varphi_g(x; \theta) = d\varphi_f(x; \theta) \qquad \forall x \in S \tag{3}$$

for some $d > 0$ share the same MLE for θ. This seemingly trivial and innocuous observation leads us to the introduction of a concept that happens to be fundamental for this paper and for MLE characterizations in general: the concept of (parameter-dependent) equivalence classes (e.c. hereafter), meaning that two densities are equivalent if their score functions satisfy (3). It is obvious that e.c.'s constitute a partition of the space of distributions. Stated simply, without

further conditions or specifications on the density associated with the target distribution P_θ^f (the target density f) and on the functions g, MLE characterization theorems identify an e.c. of distributions rather than a single well-specified distribution.

Remark 2.5. Considering, for example, Gauss' MLE characterization, both Teicher [46] and Azzalini and Genton [4], to cite but these, identify the Gaussian distribution with respect to its location parameter only up to an unknown variance. On the contrary, when dealing with scale characterizations, Teicher [46] imposes a tail-constraint in order to be able to single out the standard exponential and the standard Gaussian distribution.

The use of e.c.'s are far from new in the context of ML estimation. Indeed Daniel Bernoulli, who is considered to be one of the first authors to introduce the idea underlying the concept of maximum likelihood, noted as early as in 1778 that the roots of his "likelihood function" would not change by squaring the semi-circular density he used (see Stigler [44], Chapter 16). Further ideas of Bernoulli can be found in Kendall [28], Section 19, and they clearly demonstrate that the equivalence (3) can be viewed as a modern expression of Bernoulli's early thoughts.

In what follows, we shall state our results in the most general possible way, without (at least in the main results) considering additional identification constraints. As will become clear from the subsequent sections, the nature of the parameter of interest θ heavily influences the partition of the space of distributions, so that, for each type of parameter, one needs to identify the e.c.'s (see the beginnings of Sections 4 and 5 for an illustration).

The framework we have developed so far allows us to reformulate the question underpinning the present article in a more transparent form, namely "*Do there exist two distinct e.c.'s $\mathcal{F}(\theta)$ and $\mathcal{G}(\theta)$ such that the distributions P_θ^f and P_θ^g for $f \in \mathcal{F}(\theta)$ and $g \in \mathcal{G}(\theta)$ share a given MLE of the parameter of interest θ*"? As is perhaps already clear, the answer to this question is negative – at least in the interesting cases. To see why this ought to be the case let $\mathcal{F}(\theta)$ and $\mathcal{G}(\theta)$ be two e.c.'s which share an MLE for some sample size $n \geq 1$, in other words suppose that, for some $n \geq 1$, some $f \in \mathcal{F}(\theta)$ and some $g \in \mathcal{G}(\theta)$, the estimator $\hat{\theta}_g^{(n)}$ coincides with $\hat{\theta}_f^{(n)}$, that is,

$$\sum_{i=1}^{n} \log g\big(x_i; \hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})\big) \geq \sum_{i=1}^{n} \log g(x_i; \theta) \tag{4}$$

for all $\theta \in \Theta_0$ and all samples $\mathbf{x}^{(n)} \in S^n$ of size n. Then clearly the only way for f and g to satisfy (4) for all samples of size n is that they be strongly related to one another; as we will see, under reasonable conditions, one can go one step further and deduce that if (4) holds true for a "sufficiently large n" then f and g must belong to the same e.c. This intuition is the heart of all the literature on MLE characterizations.

3. The minimal covering sample size

The first step towards establishing our general MLE characterizations is to gain a better understanding of the meaning of "sufficiently large n" for (4) to induce a characterization theorem. We

use the notation H_n to denote the hyperplane

$$H_n = \left\{ \mathbf{b}^{(n)} = (b_1, \ldots, b_n) \in \mathbb{R}^n \left| \sum_{j=1}^{n} b_j = 0 \right. \right\}$$

and associate, with any parametric distribution P_θ^f satisfying the requirements of Section 2, the collection(s) of sets

$$B_n^{f,\theta}(A) = \left\{ \mathbf{b}^{(n)} \in H_n | \exists \mathbf{x}^{(n)} \in A^n \text{ with } b_j = \varphi_f(x_j; \theta) \text{ for all } 1 \leq j \leq n \right\}, \qquad (5)$$

where $A^n = A \times \cdots \times A$ is the n-fold cartesian product of $A \subseteq S$, the support of the target f, and $\theta \in \Theta_0$. The interplay between the sets $B_n^{f,\theta}$ and the hyperplanes H_n determines the minimal sample size n for a characterization theorem to hold. The following lemma is crucial to our approach.

Lemma 3.1. *Suppose that for some $\theta \in \Theta_0$ the mapping $x \mapsto \varphi_f(x; \theta)$ is strictly monotone over some interval $\mathcal{X} \subset S$ and that the (restricted) image $\mathcal{I}_{f,\theta}(\mathcal{X}) := \{\varphi_f(x; \theta) | x \in \mathcal{X}\}$ is of the form $(-P_{f,\theta}^-, P_{f,\theta}^+)$, for positive constants $P_{f,\theta}^-, P_{f,\theta}^+$ (possibly infinite). Then, for all $n \geq 1$, $B_n^{f,\theta}(\mathcal{X}) = H_n \cap (\mathcal{I}_{f,\theta}(\mathcal{X}))^n$. Also, letting*

$$N_{f,\theta} = \begin{cases} 2, & \text{if } P_{f,\theta}^- = P_{f,\theta}^+, \\ \left\lceil \dfrac{\max(P_{f,\theta}^+, P_{f,\theta}^-)}{\min(P_{f,\theta}^+, P_{f,\theta}^-)} + 1 \right\rceil, & \text{if } P_{f,\theta}^- \neq P_{f,\theta}^+ \text{ and } P_{f,\theta}^-, P_{f,\theta}^+ < +\infty, \\ \infty, & \text{otherwise,} \end{cases} \qquad (6)$$

we get:

- *for $n < N_{f,\theta}$, the orthogonal projections $\Pi_{x_j}(B_n^{f,\theta}(\mathcal{X})) \subsetneq (-P_{f,\theta}^-, P_{f,\theta}^+)$ for all $j = 1, \ldots, n$;*
- *for all $n \geq N_{f,\theta}$, $\Pi_{x_j}(B_n^{f,\theta}(\mathcal{X})) = (-P_{f,\theta}^-, P_{f,\theta}^+)$ for all $j = 1, \ldots, n$.*

The number $N_{f,\theta}$ is called the minimal covering sample size (MCSS) associated to the interval $\mathcal{X}$.

The MCSS, which is by definition greater than 2, is the smallest possible value of the sample size n that ensures that all the projections of $B_n^{f,\theta}(\mathcal{X})$ onto the distinct subspaces generated by each observation x_j, $j = 1, \ldots, n$, cover entirely $\mathcal{I}_{f,\theta}(\mathcal{X})$. For ease of reference we will say, whenever this property holds true, that $B_n^{f,\theta}(\mathcal{X})$ is *projectable*.

Example 3.1. Take $f(x; \theta) = \phi(x - \theta)$ the standard Gaussian density with $\theta \in \mathbb{R}$ a location parameter. We have $S = \mathbb{R}$, $\varphi_\phi(x; \theta) = x - \theta$ (and $\hat{\theta}_\phi^{(n)} = \bar{x}$, the sample mean). Then $\varphi_\phi(x; \theta)$ is invertible over $\mathbb{R}$, $\mathcal{I}_{\phi,\theta}(\mathbb{R}) = \mathbb{R}$ (for all θ) and easy computations show that $B_n^{\phi,\theta}(\mathbb{R}) = H_n$ for all n. Note that we always have $\Pi_{x_j}(B_n^{\phi,\theta}(\mathbb{R})) = \mathbb{R}$; in other words, $B_n^{\phi,\theta}(\mathbb{R})$ is projectable for all $n \geq 2$ (and all θ) and hence MCSS $= 2$.
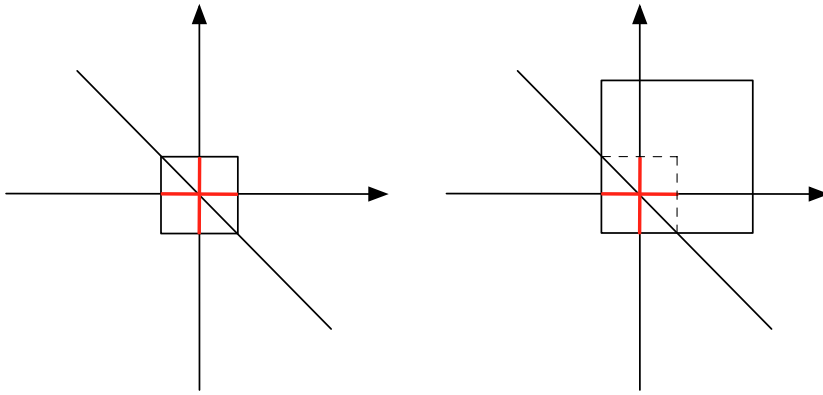
Figure 1. The sets $B_2^{f,\theta}(\mathbb{R})$ when $\mathcal{I}_{f,\theta} = [-1,1]$ (left plot) and $\mathcal{I}_{f,\theta} = [-1,3]$ (right plot). The red lines indicate the marginal projections.

Example 3.2. Take $\varphi_f(\cdot\,;\theta)$ to be monotone on $S = \mathbb{R}$ with symmetric image $(-1,1)$, say, independently of θ. Then clearly $B_2^{f,\theta}(\mathbb{R})$ is the intersection between the line $H_2 \equiv x + y = 0$ (see Figure 1) and the square $(-1,1)^2$, while $B_3^f(\mathbb{R})$ is the intersection between the plane $H_3 \equiv x + y + z = 0$ and the cube $(-1,1)^3$. Consequently, the coordinates of points on $B_2^{f,\theta}(\mathbb{R})$ and $B_3^{f,\theta}(\mathbb{R})$ cover the full image $(-1,1)$. Hence, MCSS $= 2$.

Example 3.3. Take $\varphi_f(\cdot\,;\theta)$ to be monotone on $S = \mathbb{R}$ with skewed image $(-1,3)$, say, independently of θ. Then, while $B_2^{f,\theta}(\mathbb{R})$ and $B_3^{f,\theta}(\mathbb{R})$ remain defined as in Example 3.2 (with $(-1,3)$ replacing $(-1,1)$), coordinates of points in these domains do not cover the full image. In fact, in $B_2^{f,\theta}(\mathbb{R})$ these coordinates only cover the interval $(-1,1)$ (see Figure 1), in $B_3^{f,\theta}(\mathbb{R})$ these coordinates only cover the interval $(-1,2)$ and it is only from $n \geq 4$ onwards that the coordinates of points in $B_n^{f,\theta}(\mathbb{R})$ cover the full image. Hence, MCSS $= 4$.

Example 3.4. Take $f(x;\theta) = \theta\phi(\theta x)$ the Gaussian density with $\theta \in \mathbb{R}_0^+$ a scale parameter. We have $S = \mathbb{R}$, $\varphi_\phi(x;\theta) = \frac{1}{\theta}(1 - \theta^2 x^2)$. Then $\varphi_\phi(\cdot\,;\theta)$ is invertible over $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$, separately, and $\mathcal{I}_{f,\theta}(\mathbb{R}_0^\pm) = (-\infty, 1/\theta)$. Note that we have $\Pi_{x_j}(B_n^{f,\theta}(\mathbb{R}_0^\pm)) = (-(n-1)/\theta, 1/\theta)$ for all j, all θ and all $n \geq 2$. In other words, $B_n^{f,\theta}(\mathbb{R}_0^\pm)$ is only asymptotically projectable. Hence, MCSS $= +\infty$.

Proof of Lemma 3.1. The assumptions on the image of $\varphi_f(\cdot\,;\theta)$ guarantee the existence of a point in $\mathcal{X}$ where φ_f crosses the x-axis so that $B_1^{f,\theta}(\mathcal{X}) = \{0\}$. Also, for all $n \geq 1$, we have $B_n^{f,\theta}(\mathcal{X}) = H_n \cap (\mathcal{I}_{f,\theta}(\mathcal{X}))^n$; this follows by definition. Regarding the MCSS, first take $P_{f,\theta}^- = P_{f,\theta}^+ = P$ (possibly infinite). Then, for all $n \geq 2$, $H_n \cap (-P,P)^n$ contains for each of the n coordinates the full interval $(-P,P)$ (see Figure 1); hence MCSS $= 2$. Next suppose that $P_{f,\theta}^- < P_{f,\theta}^+ < \infty$ and consider a "worst-case scenario" by taking a point at the extreme of

$H_n \cap (\mathcal{I}_{f,\theta}(\mathcal{X}))^n$, with one coordinate set to $b_1 = P_{f,\theta}^+ - \varepsilon$ for some $\varepsilon > 0$. Then, in order to construct a sample satisfying $\sum_{i=1}^n b_i = 0$, it is necessary to choose the remaining $n-1$-tuple $(b_2, \ldots, b_n)$ so as to satisfy $\sum_{i=2}^n b_i = -b_1$. Since the best choice in this respect consists in setting all b_i near the other extremum $-P_{f,\theta}^- + \varepsilon'$ for $\varepsilon' > 0$, we see that, depending on the magnitude of the ratio $P_{f,\theta}^+ / P_{f,\theta}^-$, a given sample size n may not be large enough for the equality to hold. In order to palliate this it suffices to take $N_{f,\theta}$ to be the smallest natural number such that $P_{f,\theta}^+ - (N_{f,\theta} - 1) P_{f,\theta}^- \le 0$, that is, $N_{f,\theta} = \lceil P_{f,\theta}^+ / P_{f,\theta}^- + 1 \rceil$. The case $P_{f,\theta}^+ < P_{f,\theta}^- < \infty$ follows along the same lines, and hence

$$\mathrm{MCSS} = \left\lceil \frac{\max(P_{f,\theta}^+, P_{f,\theta}^-)}{\min(P_{f,\theta}^+, P_{f,\theta}^-)} + 1 \right\rceil.$$

The same argumentation applies in the case where either one of $P_{f,\theta}^+$ or $P_{f,\theta}^-$ is infinite, this time with $\mathrm{MCSS} = +\infty$. This concludes the proof. $\qquad\square$

Now look at the connection between the MCSS and MLE characterizations. Let f and g be two representatives of distinct e.c.'s with f the target density. Under the assumption of θ-differentiability of g, the defining equation (4) can be re-expressed as

$$\sum_{i=1}^n \varphi_g\big(x_i; \hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})\big) = 0 \qquad \text{for all } \mathbf{x}^{(n)} \in S^n,$$

which in turn can be rewritten as

$$\sum_{i=1}^n \varphi_g(x_i; \theta) = 0 \qquad \text{for all } \theta \text{ and all } \mathbf{x}^{(n)} \text{ such that } \sum_{i=1}^n \varphi_f(x_i; \theta) = 0 \qquad (7)$$

(here θ and $\mathbf{x}^{(n)}$ are interdependent) or, equivalently,

$$\sum_{i=1}^n h(y_i; \theta) = 0 \qquad \text{for all } \theta \text{ and all } \mathbf{y}^{(n)} \in B_n^{f,\theta}(\mathcal{X}) \qquad (8)$$

for each interval $\mathcal{X}$ on which $y \mapsto \varphi_f(y; \theta)$ is invertible for all θ, with $h(y; \theta) = \varphi_g(\varphi_f^{-1}(y; \theta); \theta)$. Equation (8) completely identifies the function h (and hence the interconnection between g and f), at least when $B_n^{f,\theta}(\mathcal{X})$ is sufficiently rich. This richness depends strongly on the MCSS introduced in Lemma 3.1. Indeed supposing that (8) is only valid for a sample size smaller than the MCSS implies that portions of the images $\mathcal{I}_{f,\theta}(\mathcal{X})$ cannot be reached, so that h cannot be identified over its entire support. This necessarily implies that MNSS $\ge$ MCSS. It is, however, pointless to try to solve (8) in all generality and it is now necessary to specify the role of θ in order to pursue. We will do so in detail in the next sections, first in the case of location parameters; our arguments will afterwards adapt directly to other parameter choices.

4. MLE characterization for location parameter families

We start by identifying the e.c.'s for θ a location parameter. In such a case, $\Theta_0 = S = \mathbb{R}$, and the location score functions are of the form $\varphi_f(x; \theta) = \varphi_f(x - \theta) = -f'(x - \theta)/f(x - \theta)$ over $\mathbb{R}$, so that equation (3) turns into a simple first-order differential equation whose solution yields $g(x) = c(f(x))^d$ for some $d > 0$ and c the normalizing constant. Thus, all densities which are linked one to another via that relationship belong to a same e.c. We here attract the reader's attention to the fact that, for $f = \phi$ the standard Gaussian density, such transformations reduce to a non-specification of the variance, which is clearly in line with Gauss' MLE characterization as stated by Teicher [46] or Azzalini and Genton [4].

Our first main theorem is, in essence, a generalization of Gauss' MLE characterization from the Gaussian distribution to the entire class of log-concave distributions with continuous score function.

Theorem 4.1. *Let $\mathcal{F}$(loc) and $\mathcal{G}$(loc) be two distinct location-based e.c.'s and let their respective representatives f and g be two continuously differentiable densities with full support $\mathbb{R}$. Let $x \mapsto \varphi_f(x) = -f'(x)/f(x)$ be the location score function of f. If φ_f is invertible over $\mathbb{R}$ and crosses the x-axis then there exists $N \in \mathbb{N}$ such that, for any $n \geq N$, we have $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of size n if and only if there exist constants $c, d \in \mathbb{R}_0^+$ such that $g(x) = c(f(x))^d$ for all $x \in \mathbb{R}$, that is, if and only if $\mathcal{F}$(loc) $= \mathcal{G}$(loc). The smallest integer for which this holds (the minimal necessary sample size) is $MNSS = \max\{N_f, 3\}$, with N_f the MCSS as defined in Lemma 3.1.*

Proof. The sufficient condition is trivial. To prove necessity first note how our assumptions on f ensure that the score function φ_f is strictly increasing on the whole real line $\mathbb{R}$ and has a unique root. This allows us to write the image $\text{Im}(\varphi_f)$ as $(-P_f^-, P_f^+)$ with $0 < P_f^-, P_f^+ \leq \infty$. The differentiability of g and the nature of the parameter θ permit us to rewrite, for any admissible θ, (7) as

$$\sum_{i=1}^n \varphi_g(x_i - \theta) = 0 \qquad \text{for all } \mathbf{x}^{(n)} \in \mathbb{R}^n \text{ such that } \sum_{i=1}^n \varphi_f(x_i - \theta) = 0, \tag{9}$$

where $\varphi_g(x) = -g'(x)/g(x)$. Using the strict monotonicity of φ_f one then concludes that (9) is equivalent to requiring that g satisfies

$$\sum_{i=1}^n h(b_i) = 0 \qquad \text{for all } (b_1, \ldots, b_n) \in B_n^{f,\theta}(\mathbb{R}), \tag{10}$$

where $h = \varphi_g \circ \varphi_f^{-1}$ and $b_i = \varphi_f(x_i - \theta)$, $i = 1, \ldots, n$, as in (5). In what follows, we shall use our liberty of choice among all n-tuples $\mathbf{b}^{(n)} \in B_n^{f,\theta}(\mathbb{R})$ in order to gain sufficient information on h to conclude.

First, suppose that $P_{f,\theta}^- = P_{f,\theta}^+ = P$, hence that the image of φ_f is symmetric. We know from Lemma 3.1 that the corresponding MCSS equals 2, hence that two observations suffice to make

$B_n^{f,\theta}(\mathbb{R})$ projectable. Therefore, for any $n \geq 2$, we can always build an n-tuple $b_1, \ldots, b_n$ such that $b_2 = -b_1$ for all $b_1 \in (-P, P)$ and $b_i = 0$ for $i = 3, \ldots, n$. From (10) we then deduce that h satisfies the equality $h(-a) = -h(a)$ for all $a \in (-P, P)$, hence that h is odd on $(-P, P)$. Evidently this leaves h undetermined, hence the MNSS must at least equal 3. For $n \geq 3$, choose an n-tuple such that $b_3 = -b_1 - b_2$ and $b_i = 0$ for $i = 4, \ldots, n$, for $b_1, b_2 \in (-P, P)$ such that $b_1 + b_2 \in (-P, P)$. Using (10) combined with the antisymmetry of h we deduce that this function must satisfy

$$h(b) + h(c) = h(b + c) \tag{11}$$

for all $b, c \in (-P, P)$ such that $b + c \in (-P, P)$. One recognizes in (11) a (restricted) form of the celebrated *Cauchy functional equation*. Assume that $P < \infty$; then $h(P/2)$, say, is finite and standard arguments (see, e.g., Aczél and Dhombres [1]), imply that our solution h satisfies $h(uP/2) = uh(P/2)$ for all $u \in (-2, 2)$ and we conclude that $h(x) = dx$ for all $x \in (-P, P)$, with $d = h(P/2)/(P/2) \in \mathbb{R}$. Considering $x = \varphi_f(y)$ for $y \in \varphi_f^{-1}(-P, P) = \mathbb{R}$, we obtain that $\varphi_g(y) = d\varphi_f(y)$. Solving this first-order differential equation gives $g(y) = c(f(y))^d$ for all $y \in \mathbb{R}$, with c a constant. In order for the function g to be integrable over $\mathbb{R}$, the constant d must be strictly positive; in order for g to be positive and integrate to 1, the constant c must be a normalizing constant. Thus, for $P < \infty$, the problem is solved. For $P = \infty$, the situation becomes even simpler as (11) is then *precisely* the Cauchy functional equation, and one may immediately draw the same conclusion as for finite P.

Let us now consider the case where φ_f has a skewed image and set $P = \min(P_f^-, P_f^+)$ (note that P is necessarily finite as otherwise $\mathrm{Im}(\varphi_f)$ would be symmetric). First restricting our attention to $(-P, P)$, we can repeat the above arguments to deduce that $g(y) = c(f(y))^d$ for all $y \in \varphi_f^{-1}(-P, P) \subsetneq \mathbb{R}$. We thus further need to investigate the behavior of h on the remaining part of $\mathrm{Im}(\varphi_f)$ which, for the sake of simplicity, we denote as $\mathrm{Out}(P)$ (it is either $(-P_f^-, -P)$ or (P, P_f^+)). To this end, we precisely need to know the MCSS and hence call upon Lemma 3.1. Fixing $n \geq N_f$ and taking a sample $(b_1, \ldots, b_n)$ such that $\sum_{i=1}^n b_i = 0$ with $b_1 \in \mathrm{Out}(P)$ and $(b_2, \ldots, b_n) \in (-P, P)^{n-1}$, we can apply (10) to get $h(b_1) + \sum_{i=2}^n h(b_i) = 0$ and hence, from our knowledge about the behavior of h on $(-P, P)$, we deduce that

$$h(b_1) = -\sum_{i=2}^n b_i h(P)/P = b_1 h(P)/P,$$

since $h(P)$ is necessarily finite. Consequently we get $h(y) = y h(P)/P$ for all $y \in (-P, P) \cup \mathrm{Out}(P) = \mathrm{Im}(\varphi_f)$ and $g(y) = c(f(y))^d$ for all $y \in \mathbb{R}$, and the conclusion follows.

The proof of the theorem is nearly complete: all that remains is to show that the MNSS $=$ $\max\{3, N_f\}$ is minimal and sufficient. The latter is immediate since if the result holds true for any sample of size $N = \mathrm{MNSS}$ then, for any larger sample size $M > \mathrm{MNSS}$, one can always consider $\mathbf{x}^{(M)}$ such that $\varphi_f(\mathbf{x}^{(M)}) = \mathbf{b}^{(M)} \in B_M^{f,\theta}(\mathbb{R})$ is of the form $(b_1, \ldots, b_N, 0, \ldots, 0)$ and $(b_1, \ldots, b_N) \in B_N^{f,\theta}(\mathbb{R})$, and work as above to characterize the density. To prove the minimality of the MNSS it suffices to exhibit specific counter-examples. This is done in Examples 4.1 and 4.2 below. $\qquad\square$

Example 4.1. To see that $N = 3$ is minimal when $\text{Im}(\varphi_f)$ is symmetric, we need to construct two distributions g_1, g_2 which share f's MLE for all samples of size 2. Construct g_1 as in the proof of Theorem 4.1. To construct g_2, it suffices to replace the function h from (10) with any odd function and to solve the resulting equation in g (while ensuring integrability of g). If, for example, we choose $h(x) = dx^3$, then we readily obtain $g(y) = c\exp(-d\int_{-\infty}^{x}(\varphi_f(y))^3\,dy)$; this is however not a density for all f, though a good choice for $f = \phi$ the Gaussian (for which $\varphi_\phi(x) = x$). Another way of proceeding is to work as in Azzalini and Genton [4] and to choose $h(y) = y + w'(y)$ for some differentiable even function w.

Example 4.2. Suppose that $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of size n for some $n < N_f$ when $N_f > 3$. Then, as is clear from Lemma 3.1, the whole domain (i.e., $\mathbb{R}$) of f is not identified by our technique and it suffices to choose any density which is equal to f on the maximal identifiable subdomain but differs elsewhere. Expressed in terms of h for the case $P_f^- < P_f^+$, we can only identify h on some interval $(-P_f^-, P_f^- + a(n))$, say, with $0 < a(n) < P_f^+ - P_f^-$. On the remaining part $(P_f^- + a(n), P_f^+)$, h is undetermined and hence can take any possible form, implying that the relationship between g and f can only be established on the part $\varphi_f^{-1}(-P_f^-, P_f^- + a(n)) \subsetneq \mathbb{R}$.

As in Azzalini and Genton [4], it is sufficient to require in Theorem 4.1 that g be continuously differentiable at a single point for everything to run smoothly. Pursuing in this vein, it is of course natural to enquire whether the result still holds if no such regularity assumption is imposed on g, that is, if we only suppose that the target density f is differentiable but g is a priori not. Put simply the question becomes that of enquiring whether the condition

$$\sum_{i=1}^{n} \log g\left(x_i - \hat{\theta}_f^{(n)}(\mathbf{x}^{(n)})\right) \geq \sum_{i=1}^{n} \log g(x_i - \theta) \tag{12}$$

for all $\mathbf{x}^{(n)} \in \mathbb{R}^n$ and all $\theta \in \mathbb{R}$ suffices to determine g. This is the approach adopted, for example, in Teicher [46], Kagan, Linnik and Rao [27] or Marshall and Olkin [34], where it is shown that having the likelihood condition (12) with $\hat{\theta}_f^{(n)}$ the sample mean implies g is the Gaussian as soon as the result holds for all samples of sizes 2 and 3 simultaneously. Interestingly, in our framework, this arguably more general assumption on g comes with a cost: our method of proof then necessitates imposing more restrictive assumptions on f and requiring the likelihood equations to hold for two sample sizes *simultaneously*.

Theorem 4.2. *Let $\mathcal{F}(\text{loc})$ and $\mathcal{G}(\text{loc})$ be two distinct location-based e.c.'s and let their respective representatives f and g be two continuous densities with full support $\mathbb{R}$. Suppose that f is symmetric and continuously differentiable, and assume that its location score function $\varphi_f(x)$ is invertible over $\mathbb{R}$ and crosses the x-axis. Then we have $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of sizes 2 and $n \geq 3$ simultaneously if and only if there exist constants $c, d \in \mathbb{R}_0^+$ such that $g(x) = c(f(x))^d$ for all $x \in \mathbb{R}$, that is, if and only if $\mathcal{F}(\text{loc}) = \mathcal{G}(\text{loc})$.*

Proof. Our proof, which extends that of Teicher [46] from the Gaussian case to the entire class of symmetric log-concave densities f, proceeds in two main steps: we first show that our as-

sumptions on g in fact entail that g is continuously differentiable, and then conclude by applying Theorem 4.1. The additional sample size $n = 2$ needed here stems from the first step.

Condition (12) can be rewritten as

$$\sum_{i=1}^{n} \log g(y_i) \geq \sum_{i=1}^{n} \log g(y_i - \theta)$$

for all $\theta \in \mathbb{R}$ and $y_1, \ldots, y_n$ satisfying $\sum_{i=1}^{n} \varphi_f(y_i) = 0$. The latter expression in turn is equivalent to

$$\sum_{i=1}^{n-1} \log g(y_i) + \log g\left(\varphi_f^{-1}\left(-\sum_{j=1}^{n-1} \varphi_f(y_j) \right) \right)$$

$$\geq \sum_{i=1}^{n-1} \log g(y_i - \theta) + \log g\left(\varphi_f^{-1}\left(-\sum_{j=1}^{n-1} \varphi_f(y_j) \right) - \theta \right). \tag{13}$$

Arguing as in Teicher [46], it is sensible to confine our attention at first to symmetric densities g. Using the assumed symmetric nature of f (and hence the oddness of φ_f^{-1}), considering the sample size $n = 2$ and setting observation y_1 equal to some $y \in \mathbb{R}$, (13) simplifies into

$$2 \log g(y) \geq \log g(y - \theta) + \log g(y + \theta) \tag{14}$$

for all $y, \theta \in \mathbb{R}$. Since $\log g$ is everywhere finite, concave according to (14), and inherits measurability from g, it is an *a.e.*-continuously differentiable function. Arrived at this point, we may apply Theorem 4.1 to conclude (note that the oddness of φ_f makes Theorem 4.1 hold with MNSS equal to 3).

Finally, for non-necessarily symmetric densities g, we can follow exactly the argumentation from Teicher [46] and derive that the previously obtained solution is the only one, hence the claim holds. □

We stress the fact that, as in Teicher [46], we may further weaken our assumptions on g by only requiring that it is lower semi-continuous at the origin and need not have full support $\mathbb{R}$. Indeed, as shown in Teicher's proof, continuity and a.e.-positivity ensue from the above arguments.

One may wonder whether the symmetry assumption on the target density f is necessary or whether this second general location MLE characterization theorem may in fact hold for the entire class of log-concave densities as well. Our method of proof indeed requires this assumption so as to enable us to deal with such quantities as $\varphi_f^{-1}(-\sum_{j=1}^{n-1} \varphi_f(y_j))$ in (13); without any assumption on f, for $n = 2$, this expression does not simplify into the agreeable form $-y_1$. Similarly, one may wonder whether it is necessary to suppose the result to hold for two sample sizes simultaneously or whether one single sample size $N \geq 3$ might not suffice. We leave as open problems the question whether these assumptions are necessary or simply sufficient.

Finally, our Theorems 4.1 and 4.2 do not cover target densities whose location score function is monotone but not invertible over the entire real line, that is, piecewise constant. We do not consider explicitly such setups here since they do not assure that the MLE is defined in a unique way.

The strategy we however suggest consists in applying our results on the monotonicity intervals, to draw the necessary conclusions and express g in terms of f on those intervals. If we add the condition of monotonicity of φ_g, the equality $\varphi_g(x) = d\varphi_f(x)$ has to hold over the entire support $\mathbb{R}$ as monotonicity imposes φ_g to be constant outside the above-mentioned intervals. Since we here do not implicitly use the intervals where φ_f is constant, there might exist better strategies, and consequently the smallest possible sample size we obtain by following this scheme is an upper bound for the true MNSS. The most extreme situation takes place when the target is a Laplace distribution, in which case $\varphi_f(x) = \text{sign}(x)$; we refer to Kagan, Linnik and Rao [27] for a treatment of this particular distribution. We will return to these matters briefly in Section 8.

5. MLE characterization for scale parameter families

As for location parameter families, we start by identifying the e.c.'s when θ plays the role of a scale parameter. In such a setup, $\Theta_0 = \mathbb{R}_0^+$, $S = \mathbb{R}$, $\mathbb{R}_0^+$ or $\mathbb{R}_0^-$ in view of assumption (A2), and the scale score functions are of the form $\varphi_f(x; \theta) = \frac{1}{\theta}\psi_f(\theta x) := \frac{1}{\theta}(1 + \theta x f'(\theta x)/f(\theta x))$ over S, so that equation (3) turns into another quite simple first-order differential equation whose solution leads to $g(x) = c|x|^{d-1}(f(x))^d$ for some $d > 0$ (such that g is integrable) and c the normalizing constant. This relationship defines the scale-based e.c.'s. It is to be noted that $c = d = 1$ when the origin belongs to the support, that is, when $S = \mathbb{R}$, in which case the e.c.'s reduce to singletons $\{f\}$.

Our main scale MLE characterization theorem is the exact equivalent of Theorem 4.1 with ψ_f replacing φ_f, hence its proof is omitted.

Theorem 5.1. *Let $\mathcal{F}(\text{sca})$ and $\mathcal{G}(\text{sca})$ be two distinct scale-based e.c.'s and let their respective representatives f and g be two continuously differentiable densities with common support S (either $\mathbb{R}$, $\mathbb{R}_0^+$ or $\mathbb{R}_0^-$). Let $\psi_f(x) = 1 + xf'(x)/f(x)$ be the scale score function of f. If ψ_f is invertible over S and crosses the x-axis then there exists $N \in \mathbb{N}$ such that, for any $n \geq N$, we have $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of size n if and only if there exist constants $c, d \in \mathbb{R}_0^+$ such that $g(x) = c|x|^{d-1}(f(x))^d$ for all $x \in S$ (with $c = d = 1$ for $S = \mathbb{R}$), that is, if and only if $\mathcal{F}(\text{sca}) = \mathcal{G}(\text{sca})$. The smallest integer for which this holds is MNSS $= \max\{N_f, 3\}$, with N_f the MCSS as defined in Lemma 3.1.*

As in the case of a location parameter, requiring differentiability of the g's is not indispensable. One could indeed restrict the class of target distributions under consideration, as in Teicher [46]. We leave this as an easy exercise.

When dealing with scale families it is natural to work as in Teicher [46] and add a scale-identification condition of the form

$$\lim_{x \to 0} g(\lambda x)/g(x) = \lim_{x \to 0} f(\lambda x)/f(x) \qquad \forall \lambda > 0. \tag{15}$$

Imposing this condition in Theorem 5.1 allows, at least when the limit is finite, positive and does not equal $1/\lambda$ (which occurs only for pathological cases; we leave as an exercise to the reader to see why this type of limiting behavior precludes identifications), to deduce that $c = d = 1$

for $S = \mathbb{R}_0^+$ and $\mathbb{R}_0^-$, and hence $g = f$ in all cases. Interestingly Teicher already remarks that this "*seemingly ad hoc condition appears to be crucial*"; this is clearly the case for a complete identification of the family of densities which share a scale MLE.

The invertibility condition imposed on ψ_f is as natural in a scale family context as the invertibility condition on φ_f in a location family setup (see Lehmann and Casella [30], page 502). Unfortunately it suffers from one major drawback for $S = \mathbb{R}$: requiring invertibility of ψ_f over the whole real line forces us to discard several interesting cases such as, for example, the standard normal density ϕ, for which $\psi_\phi(x) = 1 - x^2$ is only invertible over the positive and negative real half-lines, respectively. More generally any symmetric density f for which φ_f is invertible over $\mathbb{R}$ will suffer from that same problem and hence will not be characterizable by means of Theorem 5.1. This flaw is nevertheless easily fixed, since Lemma 3.1 is applicable even if ψ_f is only invertible over portions of its support. This leads to our next general result (whose proof is omitted).

Theorem 5.2. *Let $\mathcal{F}$(sca) and $\mathcal{G}$(sca) be two distinct scale-based e.c.'s and let their respective representatives f and g be two continuously differentiable densities with full support $\mathbb{R}$. Let the scale score function $\psi_f(x) = 1 + xf'(x)/f(x)$ be invertible and cross the x-axis over $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$, respectively. Then there exists $N \in \mathbb{N}$ such that, for any $n \geq N$, we have $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of size n if and only if $g(x) = f(x)$ for all $x \in \mathbb{R}$. Moreover the MNSS is given by $\max\{MNSS_-, MNSS_+\}$, where $MNSS_-$ and $MNSS_+$, respectively, stand for the MNSS required on each half-line.*

It should be noted that the scale condition (15) is not necessary here since we are working on the entire support $S = \mathbb{R}$ which imposes that $d = 1$ as otherwise the non-vanishing density g would vanish at 0.

Finally note that the separation of the two real half-lines is tailored for scale families because both $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$ are invariant under the action of the scale parameter, which permits us to work on each half-line separately and put the ends together by continuity. The same would not hold true for location families due to a lack of invariance, that is, we could not "glue together" location characterizations valid on complementary subsets of the support.

6. MLE characterization for one-parameter group families

The relevance of our approach is not confined to location and scale families, but can be used for other θ-parameter families with θ neither a location nor a scale parameter. In this section, we shall consider general one-parameter group families and provide them with MLE characterization results. Group families play a central role in statistics as they contain several well-known parametric families (location, scale, several types of skew distributions as shown in Ley and Paindaveine [31], ...) and allow for significant simplifications of the data under investigation (see Lehmann and Casella [30], Section 1.4, for more details). To the best of the authors' knowledge, there exist no MLE characterizations for group families other than the location and scale families.

A univariate group family of distributions is obtained by subjecting a scalar random variable with a fixed distribution to a suitable family of transformations. More prosaically, let X be a random variable with density f defined on its support S and consider a *transformation group* $\mathcal{H}$ (meaning that it is closed under both composition and inversion) of monotone increasing functions $H_\theta : D \subseteq \mathbb{R} \to S$ depending on a single real parameter $\theta \in \Theta_0$. The family of random variables $\{H_\theta^{-1}(X), H_\theta \in \mathcal{H}\}$ is called a *group family*. These variables possess densities of the form

$$f_{\mathcal{H}}(x; \theta) := H_\theta'(x) f\big(H_\theta(x)\big), \tag{16}$$

where H_θ' stands for the derivative of the mapping $x \mapsto H_\theta(x)$ (which we therefore also suppose everywhere differentiable); their support D does not depend on θ. We call θ a $\mathcal{H}$-parameter for $f(x; \theta)$. The most prominent examples are of course $\mathcal{H}_{\mathrm{loc}} := \{H_\theta(x) = x - \theta, x, \theta \in \mathbb{R}\}$, leading to location families, and $\mathcal{H}_{\mathrm{sca}} := \{H_\theta(x) = \theta x, x \in S, \theta \in \mathbb{R}_0^+\}$ for $S = \mathbb{R}, \mathbb{R}_0^+$ and $\mathbb{R}_0^-$, yielding scale families. For further examples, we refer to Lehmann and Casella [30], Section 1.4, and the references therein.

Let us now determine the e.c.'s for $\mathcal{H}$-parameter families. Assuming that the mappings $\theta \mapsto H_\theta(x)$ and $\theta \mapsto H_\theta'(x)$ are differentiable, the $\mathcal{H}$-score function associated with densities of the form (16) corresponds to

$$\varphi_f^{\mathcal{H}}(x; \theta) := \frac{\partial_\theta H_\theta'(x)}{H_\theta'(x)} + \frac{\partial_\theta H_\theta(x) f'(H_\theta(x))}{f(H_\theta(x))} \tag{17}$$

over D (it is set to 0 outside D). Extracting e.c.'s from equation (3) is all but evident here, as (i) there is no structural reason for $\partial_\theta H_\theta(x)$ to cross the x-axis so as to allow to fix d to 1 as in the scale case over $\mathbb{R}$, and (ii) the generality of the model hampers a clear understanding of the role of θ inside the densities. Especially the latter point is crucial, as e.c.'s cannot depend on the parameter. We thus need to further specify the form of $f_{\mathcal{H}}(x; \theta)$ or, more exactly, the form of the transformations in $\mathcal{H}$ and thus the action of θ inside the densities. We choose to restrict our attention to transformations H_θ satisfying the following two factorizations:

$$\begin{cases} \partial_\theta H_\theta(x) = T(\theta) U_1\big(H_\theta(x)\big), \\ \dfrac{\partial_\theta H_\theta'(x)}{H_\theta'(x)} = T(\theta) U_2\big(H_\theta(x)\big), \end{cases}$$

where T, U_1 and U_2 are real-valued functions. At first sight, such restrictions might seem severe, but there exist numerous one-parameter transformations enjoying these factorizations, including:

- transformations of the form $H_\theta(x) = a_1(x) + a_2(\theta)$ defined over the entire real line, with a_1 a monotone increasing differentiable function over $\mathbb{R}$ and a_2 any real-valued differentiable function. These transformations lead to "generalized location families" and satisfy the above factorizations with $T(\theta) = a_2'(\theta)$, $U_1(x) = 1$ and $U_2(x) = 0$;
- transformations of the form $H_\theta(x) = a_1(x) a_2(\theta)$ defined over $\mathbb{R}, \mathbb{R}_0^+$ or $\mathbb{R}_0^-$, with a_1 a monotone increasing differentiable function over the corresponding domain and a_2 a positive real-valued differentiable function. These transformations lead to "generalized scale

families" and satisfy the above factorizations with $T(\theta) = a_2'(\theta)/a_2(\theta)$, $U_1(x) = x$ and $U_2(x) = 1$;

- transformations of the form $H_\theta(x) = \sinh(\operatorname{arcsinh}(x) + \theta)$ defined over $\mathbb{R}$. These are the so-called sinh–arcsinh transformations put to use in Jones and Pewsey [26] in order to define sinh–arcsinh distributions which allow to cope for both skewness and kurtosis. The above factorizations are verified for $T(\theta) = 1$, $U_1(x) = \sqrt{1 + x^2}$ and $U_2(x) = x/\sqrt{1 + x^2}$.

Under these premisses, equation (3) becomes

$$d\left(U_2\big(H_\theta(x)\big) + U_1\big(H_\theta(x)\big) \frac{f'(H_\theta(x))}{f(H_\theta(x))} \right)$$

$$= U_2\big(H_\theta(x)\big) + U_1\big(H_\theta(x)\big) \frac{g'(H_\theta(x))}{g(H_\theta(x))} \qquad \forall x \in D,$$

which can be rewritten as

$$d\left(U_2(x) + U_1(x) \frac{f'(x)}{f(x)} \right) = U_2(x) + U_1(x) \frac{g'(x)}{g(x)} \qquad \forall x \in S.$$

This first-order differentiable equation admits as solution $g(x) = c \exp((d - 1) \int^x U_2(y)/U_1(y)\,\mathrm{d}y)(f(x))^d$ for some $d > 0$ (such that g is integrable) and c a normalizing constant. This relationship establishes the $\mathcal{H}$-based e.c.'s. As for the scale case, $c = d = 1$ when there exists $x_0 \in S$ such that $U_1(x_0) = 0$, yielding e.c.'s constituted of singletons $\{f\}$.

For each transformation group $\mathcal{H}$, we obtain the following MLE characterization theorem for one-parameter group families. The proof of this result contains nothing new and is thus omitted.

Theorem 6.1. *Let $\mathcal{F}(\mathcal{H})$ and $\mathcal{G}(\mathcal{H})$ be two distinct $\mathcal{H}$-based e.c.'s and let their respective representatives f and g be two continuously differentiable densities with common full support S. Let $\varphi_f^{\mathcal{H}} := U_2(x) + U_1(x)f'(x)/f(x)$ be the $\mathcal{H}$-score function of f. If $\varphi_f^{\mathcal{H}}$ is invertible over S then there exists $N \in \mathbb{N}$ such that, for any $n \geq N$, we have $\hat{\theta}_f^{(n)} = \hat{\theta}_g^{(n)}$ for all samples of size n if and only if there exist constants $c, d \in \mathbb{R}_0^+$ such that $g(x) = c \exp((d - 1) \int^x U_2(y)/U_1(y)\,\mathrm{d}y)(f(x))^d$ for all $x \in S$ (with $c = d = 1$ if there exists $x_0 \in S$ such that $U_1(x_0) = 0$), that is, if and only if $\mathcal{F}(\mathcal{H}) = \mathcal{G}(\mathcal{H})$. The smallest integer for which this holds is $MNSS = \max\{N_f, 3\}$, with N_f the MCSS as defined in Lemma 3.1.*

Aside from location- and scale-based characterizations (or variations thereof) which are already available from Theorems 4.1 and 5.1, Theorem 6.1 allows, inter alia, to characterize asymmetric distributions (namely the sinh–arcsinh distributions of Jones and Pewsey [26]) with respect to their skewness parameter.

7. Examples

In this section, we analyze and discuss several examples of absolutely continuous distributions in light of the findings of the previous sections. We indicate, in each case, the corresponding

MNSS. As we shall see, we hereby retrieve a wide variety of existing results, and obtain several new ones. We stress that, in each case discussed below, the minimal sample size provided is optimal in the sense that counter-examples can be constructed if the results only are supposed to hold true for smaller sample sizes. Moreover, we attract the reader's attention to the fact that, in some examples of scale-based characterizations, we need the scale-identification condition (15), whereas in others it is superfluous, as explained in Section 5.

For the sake of clarity, we will adopt in this section the commonly used notations $\hat{\mu}_f^{(n)}$ and $\hat{\sigma}_f^{(n)}$ for location and scale ML estimators.

7.1. The Gaussian distribution

Consider the Gaussian distribution whose MLE characterizations for both the location and the scale parameter have been extensively discussed in the literature. For ϕ the standard Gaussian density, we get $\varphi_\phi(x) = x$ which is invertible over $\mathbb{R}$ and has image $\mathrm{Im}(\varphi_\phi) = \mathbb{R}$. As already mentioned several times, the MLE $\hat{\mu}_\phi^{(n)}$ is given by the sample arithmetic mean $\bar{x}$. Thus, Theorems 4.1 and 4.2 apply, with MNSS = 3 since $P_\phi^+ = P_\phi^- = \infty$. The first corresponds to Azzalini and Genton [4], Theorem 1, the second to Teicher [46], Theorem 1. Regarding the scale characterization for $\hat{\sigma}_\phi^{(n)} = (n^{-1} \sum_{i=1}^n x_i^2)^{1/2}$, direct calculations reveal that $\psi_\phi(x) = 1 - x^2$ which is invertible over both $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$ and maps both domains onto $(-\infty, 1)$. The conditions of Theorem 5.2 are thus fulfilled and yield that the MNSS equals ∞. Hence we retrieve Teicher [46], Theorem 3.

7.2. The gamma distribution

Consider the gamma distribution with tail parameter $\alpha > 0$, whose density is given by

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x) \mathbb{I}_{(0,\infty)}(x),$$

where $\mathbb{I}_A$ represents the indicator function of the set A. The exponential density is a special case of gamma densities obtained by setting $\alpha = 1$. Gamma distributions are not natural location families; this can be seen, for instance, by considering the exponential case, where $\varphi_f(x) = 1$ and hence the location likelihood equations make no sense. Within the framework of the current paper, we anyway do not provide a location-based MLE characterization of gamma densities, since their support is only $\mathbb{R}_0^+$ instead of $\mathbb{R}$. On the contrary, gamma densities allow for agreeable scale characterizations. Indeed easy computations show that $\varphi_f(x) = (-\alpha + 1)/x + 1$, $\psi_f(x) = \alpha - x$, which is thus invertible over $\mathbb{R}_0^+$, $\mathrm{Im}(\psi_f) = (-\infty, \alpha)$ and $\hat{\sigma}_f^{(n)} = \alpha^{-1}\bar{x}$. We can therefore use Theorem 5.1 in combination with the scale-identification condition (15) to obtain that the gamma distribution with shape α is characterizable w.r.t. its scale MLE $\alpha^{-1}\bar{x}$ for an infinite MNSS. We hereby recover Teicher [46], Theorem 2, and the univariate case of Marshall and Olkin [34], Theorem 5.1.

7.3. The generalized Gaussian distribution

Consider the one-parameter generalization of the normal distribution proposed in Ferguson [17], with density

$$f(x) = \frac{|\gamma|\alpha^\alpha}{\Gamma(\alpha)} \exp(\alpha\gamma x - \alpha \exp(\gamma x)),$$

where $\alpha > 0$ and γ, the additional parameter, differs from zero (Ferguson has proved that, for $\gamma \to 0$, this density converges to the Gaussian). This probability law is in fact strongly related to the gamma distribution, as it is defined as $\gamma^{-1}\log(X)$ with $X \sim \text{Gamma}(\alpha)$. Now, direct calculations yield $\varphi_f(x) = -\alpha\gamma(1 - \exp(\gamma x))$, invertible over $\mathbb{R}$, $\text{Im}(\varphi_f) = \text{sign}(\gamma)(-\alpha|\gamma|, \infty)$ and $\hat{\mu}_f^{(n)} = \gamma^{-1}\log(n^{-1}\sum_{i=1}^n \exp(\gamma x_i))$. Hence, from Theorem 4.1, we deduce that these distributions can be characterized in terms of their location parameter, with MNSS equal to ∞; we retrieve Ferguson [17], Theorem 5. Concerning the scale part, $\psi_f(x) = \alpha\gamma x(1 - \exp(\gamma x)) + 1$ is not invertible over the whole real line, but invertible over both $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$, and maps both half-lines onto $(-\infty, 1)$. Consequently, Theorem 5.2 reveals that this distribution admits as well a scale MLE characterization result, with infinite MNSS.

7.4. The Laplace distribution

Consider the Laplace distribution with density

$$f(x) = \exp(-|x|)/2.$$

One easily obtains $\varphi_f(x) = \text{sign}(x)$ and $\psi_f(x) = -x\,\text{sign}(x) + 1$. While the former function is clearly not invertible at all (but allows for a location MLE characterization; see the end of Section 5), the latter is invertible on both $\mathbb{R}_0^-$ and $\mathbb{R}_0^+$ with $\text{Im}(\psi_f) = (-\infty, 1)$. Hence Theorem 5.2 applies and reveals that the Laplace distribution is also MLE-characterizable w.r.t. its scale parameter (with infinite MNSS), which complements the existing results on MLE characterizations of the Laplace distribution from Ghosh and Rao [21], Kagan, Linnik and Rao [27], Marshall and Olkin [34].

Corollary 7.1. *The statistic*

$$\hat{\sigma}_f^{(n)} = \left(n^{-1}\sum_{i=1}^n |x_i|\right)^{-1}$$

is the MLE of the scale parameter σ within scale families over $\mathbb{R}$ for all samples of all sample sizes if and only if the samples are drawn from a Laplace distribution.

For the sake of readability we will, here and in the sequel, content ourselves with such informal statements of our characterization results; rigorous statements are straightforward adaptations of the corresponding theorems from the previous sections.

7.5. The Weibull distribution

Consider the Weibull distribution with density

$$f(x) = kx^{k-1}\exp(-x^k)\mathbb{I}_{(0,\infty)}(x),$$

where $k > 0$ is the shape parameter. As for gamma distributions, we do not provide a location-based MLE characterization for this distribution on the positive real half-line. Regarding the scale part, we have $\varphi_f(x) = -\frac{k-1}{x} + kx^{k-1}$, $\psi_f(x) = k(1 - x^k)$, clearly invertible over $\mathbb{R}_0^+$, and $\mathrm{Im}(\psi_f) = (-\infty, k)$. Thus, all conditions for Theorem 5.1 are satisfied, from which we derive, under the scale-identification condition (15), the following, to the best of our knowledge new, MLE characterization of the Weibull distribution.

Corollary 7.2. *Let condition* (15) *hold. Then the statistic*

$$\hat{\sigma}_f^{(n)} = \left(n^{-1}\sum_{i=1}^{n} x_i^k\right)^{-1/k}$$

is the MLE of the scale parameter σ within scale families over $\mathbb{R}_0^+$ for all samples of all sample sizes if and only if the samples are drawn from a Weibull distribution with shape parameter k.

7.6. The Gumbel distribution

Consider the Gumbel distribution with density

$$f(x) = \exp(-x - \exp(-x)).$$

Straightforward manipulations yield $\varphi_f(x) = 1 - \exp(-x)$, invertible over $\mathbb{R}$ and $\mathrm{Im}(\varphi_f) = (-\infty, 1)$. Thus, all conditions for Theorem 4.1 are satisfied, from which we derive the following, to the best of our knowledge new, MLE characterization of the Gumbel distribution (actually, of the *power-Gumbel distribution*).

Corollary 7.3. *The statistic*

$$\hat{\mu}_f^{(n)} = \log\left[\left(n^{-1}\sum_{i=1}^{n}\exp(-x_i)\right)^{-1}\right]$$

is the MLE of the location parameter μ within location families over $\mathbb{R}$ for all samples of all sample sizes if and only if the samples are drawn from a power-Gumbel distribution with density $c\exp(-dx - d\exp(-x))$ for $c, d \in \mathbb{R}_0^+$.

As for the scale part, it follows that $\psi_f(x) = x(-1 + \exp(-x)) + 1$, non-invertible over $\mathbb{R}$ but invertible over both $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$, and $\mathrm{Im}(\psi_f) = (-\infty, 1)$. Consequently, Theorem 5.2 applies

and shows that the Gumbel distribution (here not a general power-Gumbel distribution) allows as well for a MLE characterization with respect to its scale parameter (with corresponding MNSS equal to ∞).

7.7. The Student distribution

Consider the Student distribution with $\nu > 0$ degrees of freedom, with density

$$f(x) = \kappa(\nu)\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$$

for $\kappa(\nu)$ the appropriate normalizing constant. Then although the support of f is the whole real line, the location score function $\varphi_f(x) = (\nu+1)\frac{x}{\nu+x^2}$ is not invertible and thus we cannot provide a location characterization. On the other hand straightforward computations yield

$$\psi_f(x) = 1 - (\nu+1)\frac{x^2}{\nu+x^2} = -\nu + \frac{\nu(\nu+1)}{\nu+x^2}.$$

This function is invertible over both the positive and the negative real half-line with $\mathrm{Im}(\psi_f) = (-\nu, 1)$ and thus the Student distribution with ν degrees of freedom is by virtue of Theorem 5.2 scale-characterizable with

$$\mathrm{MNSS} = \begin{cases} \left\lceil 1 + \dfrac{1}{\nu}\right\rceil, & \text{if } \nu < 1, \\ 3, & \text{if } \nu = 1, \\ \lceil 1 + \nu \rceil, & \text{if } \nu > 1. \end{cases}$$

This result generalizes the scale characterization of the Gaussian distribution, which is a particular case of the Student distributions when ν tends to infinity. Note that the expression above then indeed yields an infinite MNSS. Moreover, the Cauchy distribution, obtained for $\nu = 1$, is MLE-characterizable w.r.t. its scale parameter with an MNSS of 3.

7.8. The logistic distribution

Consider the logistic distribution, whose density is given by

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

Straightforward manipulations yield $\varphi_f(x) = \tanh(x/2)$, which is invertible over $\mathbb{R}$ and $\mathrm{Im}(\varphi_f) = (-1, 1)$. Theorem 4.1 applies and yields the (to the best of our knowledge) first MLE characterization of the (power-)logistic distribution with respect to its location parameter (with corresponding MNSS equal to 3).

Further, $\psi_f(x) = 1 - x\tanh(x/2)$ is not invertible over the whole real line, but is invertible over both $\mathbb{R}_0^+$ and $\mathbb{R}_0^-$, and maps both half-lines onto $(-\infty, 1)$. Consequently, Theorem 5.2

applies and yields the (to the best of our knowledge) first scale MLE characterization of the logistic distribution, with infinite MNSS.

7.9. The sinh–arcsinh skew-normal distribution

As a final example, we consider the sinh–arcsinh skew-normal distribution of Jones and Pewsey [26] whose density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{(1 + \sinh^2(\operatorname{arcsinh}(x) + \delta))^{1/2}}{(1 + x^2)^{1/2}} e^{-\sinh^2(\operatorname{arcsinh}(x) + \delta)/2},$$

where $\delta \in \mathbb{R}$ is a skewness parameter regulating the asymmetric nature of the distribution. Clearly, for $\delta = 0$, corresponding to the symmetric situation, one retrieves the standard normal distribution. Now, straightforward but tedious calculations provide us with expressions for φ_f and ψ_f which can both be seen to be non-invertible. Hence, no location-based nor scale-based characterizations can be obtained. However, the sinh–arcsinh skew-normal distribution can be characterized w.r.t. its skewness parameter. As shown in Section 6, the sinh–arcsinh transform belongs to the class of transforms leading to group families. Consequently, its skewness score function is given by

$$\varphi_f^{\mathcal{H}}(x) = U_2(x) + U_1(x) \frac{\phi'(x)}{\phi(x)} = \frac{-x^3}{(1 + x^2)^{1/2}}$$

with ϕ the standard Gaussian density. This mapping is invertible over $\mathbb{R}$ with symmetric image $\mathbb{R}$. Theorem 6.1 therefore applies and yields the (to the best of our knowledge) first MLE characterization of the sinh–arcsinh skew-normal distribution (with respect to its skewness parameter) with an MNSS equal to 3.

8. Discussion and open problems

In this article, we have provided a unified treatment of the topic of MLE characterizations for one-parameter group families of absolutely continuous distributions satisfying certain regularity conditions. A natural question of interest is then in how far our methodology can be adapted to other distributions which do not satisfy these assumptions. Of particular interest are (i) parametric families whose score function is either not invertible or not differentiable at a countable number of points (such as, e.g., the Laplace distribution w.r.t. its location parameter), (ii) families depending on more than one parameter and (iii) discrete families. Although we will not cover these questions in full here, we conclude the paper by providing a number of intuitions on these questions; in all cases it seems clear that our methodology provides – at least in principle – the path towards a satisfactory answer.

Regarding the first point, an interesting issue to investigate is how the non-invertibility of the score function influences the MNSS. Indeed in the case of a Laplace target the MNSS is known to be equal to 4 (see Ghosh and Rao [21], Kagan, Linnik and Rao [27]). This increase is due to the fact that the Laplace score function only takes on two distinct non-zero values so that

having (9) for three sample points forces one of the observations to be 0 (otherwise the equality cannot hold) and therefore the case $n = 3$ provides no more information than the case $n = 2$ (and thus MNSS ≥ 4). It would of course be interesting to understand the influence of the number of distinct values taken by a given score function on the corresponding MNSS. One would, moreover, need to deal in this case with commensurability issues in order for the corresponding identity (9) to hold; this would most certainly lead to interesting discussions. Aside from these issues, however, the question of characterizability is, to the best of our understanding, covered by our approach (see the end of Section 4).

Regarding the second point, it seems straightforward (but clearly requires some care) to extend our method to a multi-dimensional location parameter, as is already done in Marshall and Olkin [34] for a Gaussian target density. In a nutshell, it suffices to project the now multi-dimensional location score function onto distinct-directional unit vectors and then proceed "as in the univariate case". On the contrary, dealing with a high-dimensional scale parameter seems more difficult, as the scale parameter becomes a matrix-valued scatter or shape parameter. One possibility could be to try to adapt Marshall and Olkin's [34] working scheme, who have been able to provide an MLE characterization for the scatter parameter of a multinormal distribution. Along these lines a final issue that we have not considered is that of MLE characterizations of univariate target distributions with respect to multivariate parameters (such as the Gaussian in terms of its two parameters (μ, σ)). In support of our optimism for these multivariate setups, see Duerinckx and Ley [16] where our methodology was successfully applied to the (perhaps more complex) case of the spherical location parameter families.

Finally concerning the discrete setup, it seems clear that our approach again yields in principle a satisfactory answer for discrete group families, although this will require a certain amount of work. We defer the systematic treatment of this interesting question to later publications.

We conclude this paper by an intriguing question suggested by an anonymous referee, which is the following: do characterization results survive mixtures of distributions? More concretely, if a given distribution is MLE-characterizable w.r.t. a given parameter of interest, under which conditions will mixtures of this distribution remain MLE-characterizable? This question is all but straightforward to answer. Indeed, we have seen in Section 7 that the Student distribution is only characterizable w.r.t. its scale parameter, whereas the logistic distribution admits an MLE characterization w.r.t. both the location and scale parameter; the Student and the logistic distribution are both scale mixtures of the Gaussian distribution, which is MLE-characterizable both as a location and scale family. We leave this question as an open problem.

Acknowledgements

References

[1] Aczél, J. and Dhombres, J. (1989). *Functional Equations in Several Variables with Applications to Mathematics*, *Information Theory and to the Natural and Social Sciences*. *Encyclopedia of Mathematics and Its Applications* **31**. Cambridge: Cambridge Univ. Press. MR1004465

[2] Akaike, H. (1977). On entropy maximization principle. In *Applications of Statistics* (P.R. Krishnaiah, ed.) 27–41. Amsterdam, The Netherlands: North-Holland.

[3] Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann*. *Inst*. *Statist*. *Math*. **30** 9–14. MR0507075

[4] Azzalini, A. and Genton, M.G. (2007). On Gauss's characterization of the normal distribution. *Bernoulli* **13** 169–174. MR2307401

[5] Bondesson, L. (1997). A generalization of Poincaré's characterization of exponential families. *J. Statist*. *Plann*. *Inference* **63** 147–155. MR1491575

[6] Bourguin, S. and Tudor, C.A. (2011). Cramér theorem for gamma random variables. *Electron*. *Commun*. *Probab*. **16** 365–378. MR2819659

[7] Buczolich, Z. and Székely, G.J. (1989). When is a weighted average of ordered sample elements a maximum likelihood estimator of the location parameter? *Adv. in Appl*. *Math*. **10** 439–456. MR1023943

[8] Campbell, L.L. (1970). Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities. *Ann*. *Math*. *Statist*. **41** 1011–1015. MR0267698

[9] Chatterjee, S.K. (2003). *Statistical Thought*: *A Perspective and History*. Oxford: Oxford Univ. Press. MR1999246

[10] Chen, L.H.Y. (1975). Poisson approximation for dependent trials. *Ann*. *Probab*. **3** 534–545. MR0428387

[11] Chen, L.H.Y., Goldstein, L. and Shao, Q.M. (2010). *Normal Approximation by Stein's Method*. *Springer Series in Probability and Its Applications*. New York: Springer.

[12] Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley. MR2239987

[13] Cramér, H. (1936). Über eine Eigenschaft der Normalen Verteilungsfunktion. *Math. Z*. **41** 405–414. MR1545629

[14] Cramér, H. (1946). A contribution to the theory of statistical estimation. *Skand*. *Aktuarietidskr*. **29** 85–94. MR0017505

[15] Cramér, H. (1946). *Mathematical Methods of Statistics*. *Princeton Mathematical Series* **9**. Princeton, NJ: Princeton Univ. Press. MR0016588

[16] Duerinckx, M. and Ley, C. (2012). Maximum likelihood characterization of rotationally symmetric distributions on the sphere. *Sankhyā Ser. A* **74** 249–262. MR3021559

[17] Ferguson, T.S. (1962). Location and scale parameters in exponential families of distributions. *Ann*. *Math*. *Statist*. **33** 986–1001. MR0141184

[18] Findeisen, P. (1982). Die Charakterisierung der Normalverteilung nach Gauß. *Metrika* **29** 55–63. MR0661707

[19] Galambos, J. (1972). Characterization of certain populations by independence of order statistics. *J. Appl*. *Probab*. **9** 224–230. MR0292127

[20] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. *Cambridge Library Collection*. Cambridge: Cambridge Univ. Press. Reprint of the 1809 original.

[21] Ghosh, J.K. and Rao, C.R. (1971). A note on some translation parameter families of densities for which the median is an m.l.e. *Sankhyā Ser. A* **33** 91–93.

[22] Haikady, N.N. (2006). *Characterizations of Probability Distributions*, part a ed. *Springer Handbook of Engineering Statistics*. London: Springer.

[23] Hald, A. (1998). *A History of Mathematical Statistics from* 1750 *to* 1930. *Wiley Series in Probability and Statistics*: *Texts and References Section*. New York: Wiley. MR1619032

[24] Hürlimann, W. (1998). On the characterization of maximum likelihood estimators for location-scale families. *Comm*. *Statist*. *Theory Methods* **27** 495–508. MR1621418

[25] Jaynes, E.T. (1957). Information theory and statistical mechanics. *Phys*. *Rev*. (2) **106** 620–630. MR0087305

[26] Jones, M.C. and Pewsey, A. (2009). Sinh–arcsinh distributions. *Biometrika* **96** 761–780. MR2564489

[27] Kagan, A.M., Linnik, Yu.V. and Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*. New York: Wiley. MR0346969

[28] Kendall, M.G. (1961). Studies in the history of probability and statistics. XI. Daniel Bernoulli on maximum likelihood. *Biometrika* **48** 1–2. MR0124988

[29] Kotz, S. (1974). Characterizations of statistical distributions: A supplement to recent surveys. *Int*. *Statist*. *Rev*. **42** 39–65. MR0353532

[30] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR1639875

[31] Ley, C. and Paindaveine, D. (2010). Multivariate skewing mechanisms: A unified perspective based on the transformation approach. *Statist*. *Probab*. *Lett*. **80** 1685–1694. MR2734229

[32] Ley, C. and Paindaveine, D. (2010). On the singularity of multivariate skew-symmetric models. *J*. *Multivariate Anal*. **101** 1434–1444. MR2609504

[33] Lukacs, E. (1956). Characterization of populations by properties of suitable statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol*. *II* 195–214. Berkeley, CA: Univ. California Press. MR0084892

[34] Marshall, A.W. and Olkin, I. (1993). Maximum likelihood characterizations of distributions. *Statist*. *Sinica* **3** 157–171. MR1219297

[35] Norden, R.H. (1972). A survey of maximum likelihood estimation. *Int*. *Statist*. *Rev*. **40** 329–254.

[36] Park, S.Y. and Bera, A.K. (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *J*. *Econometrics* **150** 219–230. MR2535518

[37] Patil, G.P. and Seshadri, V. (1964). Characterization theorems for some univariate probability distributions. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **26** 286–292. MR0171340

[38] Poincaré, H. (1912). *Calcul des Probabilités*. Paris: Carré-Naud.

[39] Puig, P. (2003). Characterizing additively closed discrete models by a property of their maximum likelihood estimators, with an application to generalized Hermite distributions. *J*. *Amer*. *Statist*. *Assoc*. **98** 687–692. MR2011682

[40] Puig, P. (2008). A note on the harmonic law: A two-parameter family of distributions for ratios. *Statist*. *Probab*. *Lett*. **78** 320–326. MR2387553

[41] Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications. *J*. *Amer*. *Statist*. *Assoc*. **101** 332–340. MR2268050

[42] Ross, N. (2011). Fundamentals of Stein's method. *Probab*. *Surv*. **8** 210–293. MR2861132

[43] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol*. *II*: *Probability Theory* 583–602. Berkeley, CA: Univ. California Press. MR0402873

[44] Stigler, S.M. (1999). *Statistics on the Table*: *The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard Univ. Press. MR1712969

[45] Stigler, S.M. (2007). The epic story of maximum likelihood. *Statist*. *Sci*. **22** 598–620. MR2410255

[46] Teicher, H. (1961). Maximum likelihood characterization of distributions. *Ann*. *Math*. *Statist*. **32** 1214–1222. MR0130726

[47] von Mises, R. (1918). Über die Ganzzahligkeit der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift* **19** 490–500.

[48] Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *J. Econometrics* **115** 347–354. MR1984780