Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

When the score function is the identity function - A tale of characterizations of the normal distribution

Christophe Ley

Department of Applied Mathematics, Computer Science and Statistics, Campus Sterre, Krijgslaan 281, Ghent 9000, Belgium

ARTICLE INFO

Article history:
Received 11 May 2020
Revised 1 October 2020
Accepted 5 October 2020
Available online xxx

MSC:
Primary 62E10
Secondary 60E99

Keywords:
Maximum likelihood characterization
Score function
Skew-symmetric distributions
Stein characterization
Variance bounds

ABSTRACT

The normal distribution is well-known for several results that it is the only to fulfil. Much less well-known is the fact that many of these characterizations follow from the fact that the derivative of the log-density of the normal distribution is the (negative) identity function. This *a priori* very simple yet surprising observation allows a deeper understanding of existing characterizations and paves the way for an immediate extension of various seemingly normal-based characterizations to a general density by replacing the (negative) identity function in these results with the derivative of that log-density.

1. Introduction

The normal or Gaussian distribution is the most popular probability law in statistics and probability. The reasons for this popularity are manifold, including the nice bell curve shape, the simple form of the density

$$x \mapsto \phi_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$

with easily interpretable location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$, the ensuing mathematical tractability, the straightforward extension to the multivariate normal density (which we shall however not deal with in this paper) or the fact of being the limit distribution in the Central Limit Theorem. Besides these major appeals, the normal distribution is also famous for satisfying various characterizations, the latter being theoretical results that only one distribution (or one class of distributions) fulfils. Carl Friedrich Gauss himself has obtained the normal density by searching for a probability distribution where the maximum likelihood estimator of the location parameter *always* (see Section 2.1 for a precise meaning) coincides with the most intuitive estimator, namely the sample average. Numerous other characterizations of this popular distribution have followed, and in general it took the researchers decades to extend them to other distributions, often in an ad hoc way.

For the sake of historical correctness, it is necessary to recall that the "Gaussian distribution" is a perfect example of Stigler's law of eponymy (Stigler, 1980) because it got first introduced by de Moivre in 1738. For more information, we refer the interested reader to the insightful paper Le Cam (1986).

E-mail address: christophe.ley@ugent.be

https://doi.org/10.1016/j.ecosta.2020.10.001

In the present paper, we will show that an apparently inessential characterization of the normal distribution turns out to be a crucial building block in several more famous characterizations. This characterization is the fact that $(\log \phi_{0,1}(x))' = -x$ or, equivalently, $\frac{d}{d\mu}(\log \phi_{\mu,\sigma}(x)) = \frac{x-\mu}{\sigma^2}$. In the former notation we speak of the derivative of the log-density, while the second case features the location score function (we will refer to both settings as the "identity function" or "location score function"). It is straightforward to see that the normal distribution is the only one for which these results hold. We shall show in the remainder of this paper that this particular characterization of the normal distribution via the identity function lies at the core of many characterizations that convey its special role to the normal distribution. We will illustrate this fact by means of 4 totally unrelated examples from the literature, namely the maximum likelihood characterization (Section 2.1), a singular Fisher information matrix characterization within skew-symmetric distributions (Section 2.2), Stein characterizations (Section 2.3) and a characterization related to variance bounds (Section 2.4). In each case, we indicate where the identity function plays its role and how, by replacing it with $(\log p(x))'$ for some general density p, the characterization that seemed tailor-made for the normal distribution can in fact be extended to other distributions. For some examples we bring in some innovative viewpoint in the proofs, for others not. We wish to stress that the examples are not the goal of this paper, but they serve the purpose of illustrating the global vision of our paper, namely that recognizing $-x$ as $(\log \phi_{0,1}(x))'$

- allows a better understanding of various characterization results of the normal distribution;
- yields a simple tool for extending a result for the normal to virtually any distribution.

The reason why we wish to stress the $(\log \phi_{0,1}(x))' = -x$ characterization is because it goes unnoticed. If one were to find $(\log(p(x)))'$ in the examples shown in Sections 2.1-2.4, it would be obvious, and not surprising, that the result holds for the density p and can be generalized to another density q by replacing p with q. In case of $-x$, the underlying density $\phi_{0,1}(x)$ is hidden (of course, not every $-x$ is associated with the normal distribution).

Finally, bearing in mind that the issues we just discussed are based on the location score function, we explain in Section 3 how alternative characterizations can be obtained by rather looking at the scale score function. We conclude the paper with final comments in Section 4.

2. Four characterizations from different research domains

2.1. Maximum likelihood characterization

We call *location MLE characterization* the characterization of a probability distribution via the structure of the Maximum Likelihood Estimator (MLE) of the location parameter. Gauss (1809) showed that, in a location family $p(x-\mu)$ with differentiable density p, the MLE for μ is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ for all samples $(x_1, \ldots, x_n)$ of all sample sizes n if, and only if, p is the normal density. This result has been successively refined in two directions. On the one hand, several authors have worked towards weakening the regularity assumptions on the class of densities p considered; for instance Gauss requires differentiability while Teicher (1961) only requires continuity. On the other hand, many authors have lowered the sample size necessary for the characterization to hold, in other words, the "always"-statement from the Introduction. For instance, Gauss requires that the sample mean be MLE for the location parameter for all sample sizes simultaneously, Teicher (1961) only needs that it be MLE for samples of sizes 2 and 3 at the same time, while Azzalini and Genton (2007) only require that it be so for a single fixed sample size $n \geq 3$. We refer the reader to Section 1.1 of Duerinckx et al. (2014) for a complete list of references on the topic. MLE characterizations are useful in various aspects. First, they allow a statistician to relate a distribution to a particular estimator; in other words, if a statistician wishes to use a certain estimator which is part of an MLE characterization, then he/she is strongly advised to work with the related distribution. Moreover, these characterizations have proved helpful in theoretical developments, see for instance Section 5 of Ley and Paindaveine (2010).

Location MLE characterizations for other distributions (Laplace, Gumbel, among others) were first proposed on a case-by-case ad hoc basis, before Duerinckx et al. (2014) unified all existing results and established the most general MLE characterization results. We shall here not delve into the technical details of Duerinckx et al. (2014), but rather take a look at the proof by Azzalini and Genton (2007) and explain how the identification of the identity function and a subsequent replacement with $\varphi_p(x) = (\log p(x))'$ directly leads to a generalization of location MLE characterizations.

Letting $g(x-\mu)$ denote a density over $\mathbb{R}$ with location parameter $\mu \in \mathbb{R}$, and assuming differentiability of g, the starting point of the proof by Azzalini and Genton (2007) consists in considering the score equation

$$\sum_{i=1}^n \varphi_g(x_i - \bar{x}) = 0, \tag{1}$$

with $\varphi_g(x) = (\log g(x))'$ for all samples $x_1, \ldots, x_n$ of a fixed sample size $n \geq 3$. The particular choices $x_1 = x_2 = \cdots = x_n = 0$, $x_1 = a = -x_2, x_3 = x_4 = \cdots = x_n = 0$ and $x_1 = a, x_2 = b, x_3 = -a-b, x_4 = \cdots = x_n = 0$ with $a, b \in \mathbb{R}$ lead to the functional Cauchy equation $\varphi_g(a+b) = \varphi_g(a) + \varphi_g(b)$ for all $a, b \in \mathbb{R}$. Its unique solution is $\varphi_g(x) = cx$ for some real constant c, which is precisely the identity function characterization of the normal distribution that we mentioned in the Introduction, up to the constant c.

Now, how can we extend this result to a general density p? To this end, let p denote an arbitrary but for the rest of the proof fixed density and g any density up to regularity conditions. The key idea, provided in Duerinckx et al. (2014) and

slightly simplified here for exposition purposes, lies in the fact that the score Eq. (1) is a system of two equations, which in terms of our general density p reads

$$\sum_{i=1}^{n} \varphi_g(x_i-\mu)=0 \quad \text{subject to} \quad \sum_{i=1}^{n} \varphi_p(x_i-\mu)=0. \tag{2}$$

The second equation was somewhat hidden in (1) under the form $\sum_{i=1}^n \varphi_g(x_i-\mu)=0 \quad \text{subject to} \quad \sum_{i=1}^n (x_i-\mu)=0$. Writing $\alpha_i=\varphi_p(x_i-\mu)$ and assuming φ_p to be monotone with image $\mathbb{R}$, Eq. (2) can be rewritten as

$$\sum_{i=1}^{n} \varphi_g\circ\varphi_p^{-1}(\alpha_i)=0 \quad \text{subject to} \quad \sum_{i=1}^{n} \alpha_i=0. \tag{3}$$

Comparing (3) with (1), we notice that both are the same set of equations thanks to the monotonicity assumption together with the fact that the α_i span over $\mathbb{R}$. Consequently, we find that $\varphi_g\circ\varphi_p^{-1}(x)=cx$, leading to g equal to p^c with c necessarily positive and hence a location MLE characterization for p. From the similarity with the normal proof, it inherits validity for all samples from a fixed sample size $n\geq 3$. For the sake of illustration, this for instance allows characterizing the power-Gumbel distribution with density $c\exp(-dx-d\exp(-x)), c,d\in\mathbb{R}_0^+$, via its location MLE $\log[(n^{-1}\sum_{i=1}^n \exp(-x_i))^{-1}]$.

We conclude this section with a few comments. The monotonicity assumption and the fact that φ_p maps $\mathbb{R}$ onto all $\mathbb{R}$ are natural extensions from the normal log-density being the identity. Actually, the second requirement may be weakened by only asking that φ_p crosses the x-axis (otherwise the equation $\sum_{i=1}^n \varphi_p(x_i-\mu)=0$ would have no solution), as is the case for the power-Gumbel for example. We refer the reader to Duerinckx et al. (2014) for a formal proof of the latter statement. Strict monotonicity and crossing the x-axis are two requirements that define the class of *strong unimodal* or *log-concave* densities. The location MLE characterization of the normal distribution thus can more or less readily be extended to this broader family of distributions thanks to the understanding of the role played by the "normal" identity function.

2.2. Singularity of the Fisher information matrix in skew-symmetric distributions

Nice as it is, the symmetric shape of the normal distribution also has its drawbacks, as it does not allow modelling data exhibiting skewness. Consequently, several proposals for modifying the normal distribution have been brought forward in the literature, such as transformations, mixture models or symmetry modulation, see Ley (2015) for details. The most famous instance of symmetry modulation is the *skew-normal* of Azzalini (1985) with density

$$x\mapsto 2\phi_{\mu,\sigma}(x)\Phi_{0,1}\left(\delta\frac{(x-\mu)}{\sigma}\right), \quad x\in\mathbb{R}, \tag{4}$$

where $\Phi_{\mu,\sigma}$ stands for the cumulative distribution function (cdf) associated with the normal density $\phi_{\mu,\sigma}$ and $\delta\in\mathbb{R}$ plays the role of a skewness parameter. At $\delta=0$ we retrieve the normal distribution, and all non-zero values of δ lead to a skewed distribution. Many further papers have studied various aspects of the skew-normal, and generalizations to other so-called skew-symmetric densities have been proposed both in the univariate and multivariate settings. Scalar skew-symmetric densities are of the form

$$x\mapsto \frac{2}{\sigma}p\left(\frac{x-\mu}{\sigma}\right)\Pi\left(\frac{x-\mu}{\sigma},\delta\right), \quad x\in\mathbb{R},$$

where p is a symmetric density to be skewed and the skewing function Π satisfies $\Pi(z,\delta)+\Pi(-z,\delta)=1\,\forall z,\delta\in\mathbb{R}$ and $\Pi(z,0)=\frac{1}{2}\,\forall z\in\mathbb{R}$. The most typical choice of skewing function is $F(\delta\frac{(x-\mu)}{\sigma})$ for some symmetric univariate cdf F, see (4) where $F=\Phi_{0,1}$. We refer the interested reader to Azzalini and Capitanio (2014) for a recent overview on skew-symmetric distributions.

Besides its nice stochastic properties, the skew-normal is also "infamous" for an inferential peculiarity. When $\delta=0$, the Fisher information matrix associated with the model (4) is singular with rank 2 instead of 3, due to a collinearity between the scores for location and skewness. Straightforward manipulations show that both these scores are proportional to (the identity function) $\frac{x-\mu}{\sigma}$. This singularity is for instance not compatible with the assumptions needed for the standard asymptotic behavior of the maximum likelihood estimators. Azzalini (1985) proposed a reparameterization that avoids this issue in the scalar case, Arellano-Valle and Azzalini (2008) extended this idea to the multivariate setting, and Hallin and Ley (2014) have suggested an alternative reparameterization in the scalar case. The unpleasant Fisher information singularity issue, and the ensuing difficulty of building efficient tests for normality, has received a lot of attention in the literature. Mentioned, from the very beginning, by Azzalini (1985), it is discussed, in the univariate and multivariate context, by Azzalini and Capitanio (1999), Pewsey (2000), Chiogna (2005), Ley and Paindaveine (2010), Hallin and Ley (2012) and Ho and Nguyen (2019) among others.

A long time open question in the literature was which other skew-symmetric distributions would suffer from this type of singularity. For instance, Azzalini and Capitanio (2003) have proposed a skew-t distribution (we cautiously write "a skew-t" since others exist in the literature, see, e.g., Branco and Dey (2001) or Jones and Faddy (2003)) by using the skewing function $T_{\nu+1}(\delta\frac{(x-\mu)}{\sigma}\frac{\nu+1}{\nu+\sigma^{-2}(x-\mu)^2})$ with $T_{\nu+1}$ the cdf of the Student distribution with $\nu+1>1$ degrees of freedom, and noticed "*It was a pleasant surprise to find that in the present setting the behaviour of the log-likelihood function was to be much more*

regular, at least for those numerical cases which we have explored". The alerted reader will by now have discovered what conveys this seemingly special "property" to the skew-normal (and hence the normal): the presence of the identity function inside $\Phi_{0,1}(\delta\frac{(x-\mu)}{\sigma})$. Indeed, location and skewness scores at $\delta=0$ in the skew-normal case respectively are given by $\frac{x-\mu}{\sigma^2}$ and $\sqrt{2/\pi}\frac{(x-\mu)}{\sigma}$. Now, starting from a symmetric density p, its location score function is $-\sigma^{-1}\varphi_p(\frac{x-\mu}{\sigma})$ and, consequently, a skew-p density will be singular at $\delta=0$ if it is of the form

$$x\mapsto \frac{2}{\sigma}p\Big(\frac{x-\mu}{\sigma}\Big)F\Big(\delta\varphi_p\Big(\frac{x-\mu}{\sigma}\Big)\Big),\quad x\in\mathbb{R}, \tag{5}$$

where the choice of F does not matter. When presented under the form (5), the information singularity does not look surprising, and one would expect that, for instance, $\frac{2}{\sigma}q\big(\frac{x-\mu}{\sigma}\big)F\big(\delta\varphi_p\big(\frac{x-\mu}{\sigma}\big)\big)$ for some symmetric density q (not proportional to p^c for some $c>0$ such that p^c is integrable) will not lead to singularity issues. In the skew-normal case, this fact was hidden behind the seemingly inessential identity function. The latter indeed makes it possible to characterize the skew-normal as the only skew-symmetric distribution suffering from a Fisher information singularity when using the popular skewing function $F\big(\delta\big(\frac{x-\mu}{\sigma}\big)\big)$ irrespective of the choice of F, while such a characterization readily extends to density p if we were to use $F\big(\delta\varphi_p\big(\frac{x-\mu}{\sigma}\big)\big)$. The awareness of the link between Gaussian density and identity function, which we wish to underline in the present paper, allows a direct and complete understanding of the problem. For the multivariate case, which is more complex but follows the same reasoning based on identity functions, we refer the interested reader to Hallin and Ley (2012) for a complete solution.

2.3. Stein characterizations and Stein's density approach

Stein characterizations are an important building block of a famous probabilistic tool, namely Stein's Method. The goal of this method, initiated by Charles Stein in 1972 (Stein, 1972), is to provide quantitative assessments in distributional comparison statements of the form "W is close to Z" where Z follows a known and well-understood probability distribution (typically normal as in the Central Limit Theorem) and W is the object of interest. In a nutshell, Stein's method consists of two distinct components, namely

- Part A: a framework allowing transforming the problem of bounding the error in the approximation of W by Z into a problem of bounding the expectation of a certain functional of W.
- Part B: a bunch of techniques to bound the expectation appearing in Part A; the details of these techniques heavily depend on the form of the functional and on the properties of W.

Part B is not of interest for the purpose of this paper, hence we shall not further discuss it and rather refer the interested reader to Ley et al. (2017b) and Ross (2011). Part A directly involves the Stein characterization. For a suitable operator $\mathcal{A}_Z$ and for a wide class of functions $\mathcal{F}(\mathcal{A}_Z)$, the equivalence

$$W\stackrel{d}{=}Z \text{ if and only if } \mathbb{E}[\mathcal{A}_Z f(W)]=0 \text{ for all } f\in\mathcal{F}(\mathcal{A}_Z),$$

where $\stackrel{d}{=}$ means equality in distribution, represents the Stein characterization of Z. The usefulness of such a characterization can readily be seen by noticing that $|\mathbb{E}[\mathcal{A}_Z f(W)]|$ can be used as measure of distance between Z and W, and the operator $\mathcal{A}_Z f(\cdot)$ is the abovementioned functional of W. Such distance measures have recently been put to use in various statistical problems. For instance, Liu et al. (2016) build goodness-of-fit tests, Ley et al. (2017a) propose a new measure of the impact of the prior in Bayesian statistics, Barp et al. (2019) construct estimators that can deal with distributions with intractable normalizing constant, and Betsch and Ebner (2020) propose normality tests.

Stein (1972) tackled the normal approximation problem, meaning that Z follows a standard normal distribution, and proposed as operator $\mathcal{A}_Z f(x)=f'(x)-xf(x)$. Taking the class $\mathcal{F}(\mathcal{A}_Z)$ so as to ensure the integrability conditions, one readily sees that

$$\mathbb{E}[f'(Z)-Zf(Z)]=\int_{-\infty}^{\infty}f'(z)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz-\int_{-\infty}^{\infty}zf(z)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz=0,$$

by integration by parts of the second term. This establishes the sufficiency part of the characterization, and we spare the reader the details of the necessity part. The key behind this integration by parts lies in the fact that $-z\exp(-\frac{z^2}{2})$ integrates to $\exp(-\frac{z^2}{2})$ which, of course, is due to the fact that $-z$ is the derivative of the log-density of the standard normal.

The latter observation allows us directly to deduce the form of an operator that should lead to a Stein characterization for a given target density p. Replacing $-z$ with $\varphi_p(z)=\frac{p'(z)}{p(z)}$ and letting Z follow the distribution determined by p, we readily see that

$$\mathbb{E}[f'(Z)+\varphi_p(Z)f(Z)]=\int_{-\infty}^{\infty}f'(z)p(z)dz+\int_{-\infty}^{\infty}\frac{p'(z)}{p(z)}f(z)p(z)dz=0,$$

by integration by parts (assuming the required minimal integrability conditions). Thus a simple observation and manipulation leads us to postulate that

$$W\stackrel{d}{=}Z\sim p \text{ if and only if } \mathbb{E}[f'(W)+\varphi_p(W)f(W)]=0 \text{ for all } f\in\mathcal{F}(\mathcal{A}_Z)$$

is a Stein characterization for p, where the class $\mathcal{F}(\mathcal{A}_Z)$ is determined by the regularity conditions needed for the proof. And indeed: this equivalence happens to be characterizing for any differentiable density p (there exist infinitely many other Stein characterizations for any density p, e.g. involving higher order differential operators) and has come to knowledge in the literature under the name *Stein's density approach* as proposed in Stein et al. (2004) and further studied in Ley and Swan (2013).

2.4. On a result by Cacoullos (1982) regarding variance bounds

A famous result of Chernoff (1981) states that, if $X \sim \mathcal{N}(0,1)$, then the inequality

$$\text{Var}[g(X)] \leq \text{E}[(g'(X))^2] \tag{6}$$

holds for an absolutely continuous real-valued function g for which $g(X)$ has finite variance, and the inequality becomes an equality if and only if $g(x) = ax + b$ for some real constants a and b. This type of inequality falls under the umbrella of variance bounds and is useful for solving variations of the classical isoperimetric problem. Within statistics, variance bounds are related to Cramér-Rao bounds, efficiency and asymptotic relative efficiency computations, or maximum correlation coefficients, see Ernst et al. (2020).

The inequality (6) has stimulated the search for general variance bounds, see for instance Cacoullos (1982), Klaassen (1985), Afendras and Papadatos (2014), Ley and Swan (2016b) and Ernst et al. (2020). In particular, Cacoullos (1982) presented the following lemma as basis for upper variance bounds for several densities.

Lemma 1 (Cacoullos (1982))**.** *Let X be a continuous random variable with density function $p(x)$. Let g and g' be real-valued functions on $\mathbb{R}$ such that g is an indefinite integral of g', and $\text{Var}[g(X)] < \infty$. Then*

$$\text{Var}[g(X)] \leq \int_0^\infty \int_t^\infty xp(x)[g'(t)]^2 dxdt - \int_{-\infty}^0 \int_{-\infty}^t xp(x)[g'(t)]^2 dxdt. \tag{7}$$

This result is a direct consequence of

$$\text{Var}[g(X)] = \text{Var}\left[\int_0^X g'(t)dt\right] \leq \text{E}\left[\left(\int_0^X g'(t)dt\right)^2\right] \leq \text{E}\left[X\int_0^X (g'(t))^2 dt\right], \tag{8}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Expression (7) is then readily obtained by writing out explicitly the integrals and switching the integration order. An equality in Lemma 1 corresponds to $\text{E}[g(X)] = g(0)$ (first inequality) and $g'(t) \propto 1$ (second, Cauchy-Schwarz, inequality), in other words it occurs if and only is g is linear and $\text{E}[X] = 0$ under p. When p is the standard normal density in (7), it readily follows that $\int_t^\infty xp(x)dx = p(t)$ and $-\int_{-\infty}^t xp(x)dx = p(t)$ and consequently Lemma 1 yields that $\text{Var}[g(X)] \leq \text{E}[(g'(X))^2]$ with equality if and only if g is linear. Cacoullos (1982) also applies this result to the exponential distribution but, in all generality, the lemma is designed for any continuous density p. The upper bound he gets for the exponential distribution is however far from optimal (see Section 3.4), and the reason lies in the *a priori* hidden fact that Lemma 1 is designed for the normal distribution. This can be recognized by the presence of the identity function x in the right-hand side integrals in (7) and by the fact that equality holds when g is a linear (hence nearly identity) function. Only when p is the normal density do we have that $\int_t^\infty xp(x)dx = p(t)$ which is the key element for obtaining a sharp upper bound. Intuitively, this can be seen as the most direct way to get to the density p and hence the expectation in the upper bound, and any superfluous terms would yield worse bounds.

Again, the alerted reader will now have noted how to improve on Cacoullos' approach for general densities p, namely by replacing x with $-\varphi_p(x)$ in (7), which however requires a clever prior replacement in (8). This is achieved as follows:

$$\begin{aligned}\text{Var}[g(X)] = \text{Var}\left[\int_0^{-\varphi_p(X)} (g\circ(-\varphi_p)^{-1})'(t)dt\right] &\leq \text{E}\left[-\varphi_p(X)\int_0^{-\varphi_p(X)} ((g\circ(-\varphi_p)^{-1})'(t))^2 dt\right] \\ &= \text{E}\left[-\varphi_p(X)\int_0^{-\varphi_p(X)} \frac{(g'((-\varphi_p)^{-1}(t)))^2}{((-\varphi_p)'((-\varphi_p)^{-1}(t)))^2} dt\right],\end{aligned}$$

where the monotonicity of $\varphi_p(X)$ is crucial. An equivalent to Lemma 1 is then readily written down. The key element in obtaining sharp upper bounds for any density p is that, here, $\int_t^\infty -\varphi_p(x)p(x)dx = p(t)$, which (after some manipulations) leads to the sharp upper variance bound

$$\text{Var}[g(X)] \leq \text{E}\left[\frac{(g'(X))^2}{(-\varphi_p)'(X)}\right]$$

with equality if and only if $g \propto \varphi_p$. This result (inequality plus equality statement) perfectly extends the famous variance bound for the Gaussian distribution and provides insights as to why this approach by Cacoullos worked out for the Gaussian, but not for other distributions. The steps showcased here are a simplified version of the proof provided in Ley and Swan (2016b); we refer the reader to that paper for rigorous conditions and for a discussion on the sharpness of such bounds as compared to competitors from the literature.

3. Further extensions by means of the scale score function

So far our extensions of characterization theorems of the normal distribution to any other continuous distribution with density p have been based on the score function $\varphi_p(x) = p'(x)/p(x)$ as natural extension of $-x$, the normal score function. It is important to keep in mind that these are location score functions, as described in the Introduction, and consequently all described characterizations are location-based. For certain distributions this may lead to somewhat artificial results if, for instance, they do not contain a location parameter under their natural form. Think for example of the negative exponential distribution with density $p(x) = \lambda \exp(-\lambda x)$ over $\mathbb{R}^+$, where $\lambda > 0$ is a scale parameter. Its score function $\varphi_p(x) = -\lambda$ (or -1 if we consider the standardized form) is constant and hence not very attractive, unlike its scale score function $\frac{d}{d\lambda} \log(\lambda \exp(-\lambda x)) = \frac{1}{\lambda} - x$. The same reasoning applies to many distributions over $\mathbb{R}^+$ where no location parameter is naturally present. For a scale family $\sigma\, p(\sigma x)$, the scale score (after setting $\sigma = 1$) is given by $\psi_p(x) = 1 + x\varphi_p(x)$ and substituting this quantity for φ_p in the examples considered in the previous sections allows one to obtain further interesting extensions. In the rest of this section we shall briefly discuss the four previous topics under the light of scale-based characterizations and general parametric characterizations.

3.1. MLE characterizations

Various papers have provided MLE characterizations with respect to the scale parameter. Teicher (1961) shows that, in a scale family $\sigma\, p(\sigma x)$ and under some regularity assumptions, the MLE for the scale parameter σ is the sample mean $\bar{x}$ for all samples $x_1, \ldots, x_n$ over $\mathbb{R}^+$ of all sample sizes n if and only if p is the exponential density, while if it corresponds to the square root of the sample arithmetic mean of squares $\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{1/2}$, then this characterizes the normal distribution over $\mathbb{R}$. Marshall and Olkin (1993) extend the characterization of Teicher (1961) from the negative exponential to the Gamma distribution. Duerinckx et al. (2014) provide a general characterization result for scale families that incorporates all those from the literature. These authors also provide a general MLE characterization for one-parameter group families of the form $H'_\theta(x)\, p(H_\theta(x))$ where the parameter of interest θ can take on diverse roles and H_θ is a differentiable transformation.

3.2. Fisher information singularity issue

The Fisher information singularity within skew-symmetric distributions has only been studied from what we call a location-based view. This is due to the fact that the notorious singularity in the skew-normal case is due to a collinearity between the scores for location and skewness, and consequently papers such as Hallin and Ley (2012, 2014) studied the singularity from this viewpoint. We shall therefore consider here for the first time a skewness-scale induced singularity. It is easy to see that skew-symmetric densities of the form

$$\frac{2}{\sigma} q\left(\frac{x-\mu}{\sigma}\right) F\left(\delta \psi_p\left(\frac{x-\mu}{\sigma}\right)\right), \quad x \in \mathbb{R}, \tag{9}$$

where p, q are symmetric densities and F is some univariate symmetric cdf, suffer from a singular Fisher information when $\delta = 0$ if and only if the scores for scale and skewness are collinear almost everywhere, i.e., $\psi_q(x) = c_1 \psi_p(x) + c_2$ a.e. in $x \in \mathbb{R}$ and for some real constants c_1, c_2 (since the location score $\varphi_q(x)$ is an odd function, contrary to the even scale score and, here, the also even skewness score at $\delta = 0$). The latter equation can be re-expressed under the form

$$\varphi_q(x) = \frac{(c_1 + c_2 - 1)}{x} + c_1 \varphi_p(x) \quad \text{a.e.}$$

The solution to this first-order differential equation is $q(x) = d x^{c_1+c_2-1} p^{c_1}(x)$ a.e. for some normalizing constant $d > 0$. Hence, for all values $c_1, c_2 \in \mathbb{R}$ for which $x^{c_1+c_2-1} p^{c_1}(x)$ is integrable and symmetric, the density $d x^{c_1+c_2-1} p^{c_1}(x)$ leads to a singular Fisher information in the model (9) when $\delta = 0$. The symmetry requirement reduces the possible values of $c_1 + c_2$ to odd integers. For the sake of illustration, when p is the normal density, $x^{c_1+c_2-1} \exp\left(-c_1 \frac{x^2}{2}\right)$ is integrable for all $c_1 > 0$ and c_2 such that $c_1 + c_2 \geq 1$ and odd.

3.3. Stein characterizations

For exponential approximation problems, Chatterjee et al. (2011) use and combine two different Stein characterizations of the negative exponential density $p(x) = \exp(-x)$ over $\mathbb{R}^+$. The first involves the operator $f'(x) - f(x)$ and is applied, in Part B of Stein's Method, for $x \in [0, 1]$ while the second considers $xf'(x) - (x-1)f(x)$ and is applied for $x > 1$. For the sake of readability we do not specify the regularity assumptions on f and refer to Chatterjee et al. (2011) for that purpose. The reader will have noticed that the -1 appearing in the first operator corresponds to the location score function $\varphi_p(x) = -1$ while the second operator is based on $\psi_p(x) = 1 - x$. Thus, without a proper mention in that paper, the authors obtained improved upper bounds for exponential approximation by combining location- and scale-based operators. We will not delve here into deeper structural reasons for the x appearing in $xf'(x)$ in the second characterization (for a quick and simple observation, the reader may notice that replacing $f(x)$ with $xg(x)$ in the operator $f'(x) - f(x)$ yields this second operator), and refer the interested reader to Ley et al. (2017b) for more information about setting up useful operators in Stein's Method.

General parametric Stein characterizations, based on parameters of interest of other natures than location and scale, have been studied in Ley and Swan (2016b) and Ley and Swan (2016a). The latter paper also develops a link between typical operators from the literature, such as those from Chatterjee et al. (2011), and the operators obtained by adopting the (till then not considered) parametric viewpoint.

3.4. Variance bounds

General parametric variance bounds have been studied in detail in Ley and Swan (2016b), where the scale case is given particular attention. Since previously no mention on scale-based variance bounds has been made in the literature, this paper compares the resulting bounds to those of Cacoullos (1982) and Klaassen (1985), noting in particular that the scale-based bounds clearly improve on the Cacoullos bounds and in many situations on the Klaassen bounds. This underlines the strength of the parametric approach, whose wealth is further undermined in Ley and Swan (2016b) via novel skewness-based variance bounds. It is notable that Cacoullos (1982) noticed that his variance bounds for the exponential distribution were not very sharp by having recourse to (7) (in fact, they reached equality if and only if the function g is constant since $\mathrm{E}[X]$ is not zero in that case) and proposed a way to lower the bound (yielding equality for g linear). However, this was no structural improvement grounded on the nature of the exponential distribution as a scale-based distribution, and hence cannot match the upper bounds from Ley and Swan (2016b).

4. Final comments

We hope to have conveyed through the previous examples from very different topics the important message that many characterizations of the normal distribution and, consequently, the seemingly special role of the normal distribution, are (at least to a large degree) to be attributed to the fact that its score function is the identity function which happens to appear in many circumstances. While a general score function of the form $\frac{p'(x)}{p(x)}$ would immediately hint at a special role played by the density p, the same does not hold true for $-x$ unless one is aware that $-x = \frac{\phi'_{0,1}(x)}{\phi_{0,1}(x)}$. Keeping this in mind, many results can be better understood and the theory can move forward more quickly. There exist several further situations where this observation turns out to be useful, for instance in the definition of a generalized Fisher information distance and ensuing information inequalities (see Ley and Swan (2013)) or in the extension to any target p of the normal characterization provided in Nourdin and Viens (2009), see Theorem 2 of Kusuoka and Tudor (2012).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to thank the Editor, Associate Editor and three anonymous reviewers for insightful comments that helped to improve the present paper.

References

Afendras, G., Papadatos, N., 2014. Strengthened Chernoff-type variance bounds. Bernoulli 20 (1), 245–264.
Arellano-Valle, R.B., Azzalini, A., 2008. The centred parametrization for the multivariate skew-normal distribution. Journal of Multivariate Analysis 99 (7), 1362–1382.
Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12 (2), 171–178.
Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (3), 579–602.
Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (2), 367–389.
Azzalini, A., Capitanio, A., 2014. The skew-normal and related families.. Cambridge University Press, Cambridge.
Azzalini, A., Genton, M.G., 2007. On Gauss's characterization of the normal distribution. Bernoulli 13 (1), 169–174.
Barp, A., Briol, F.-X., D.A.G.M., Mackey, L., 2019. Minimum Stein discrepancy estimators. In: Neural Information Processing Systems, pp. 12964–12976.
Betsch, S., Ebner, B., 2020. Testing normality via a distributional fixed point property in the Stein characterization. TEST 29, 105–138.
Branco, M.D., Dey, D.K., 2001. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis 79, 99–113.
Cacoullos, T., 1982. On upper and lower bounds for the variance of a function of a random variable. The Annals of Probability 10 (3), 799–809.
Chatterjee, S., Fulman, J., Röllin, A., 2011. Exponential approximation by Stein's method and spectral graph theory. ALEA Latin American Journal of Probability and Mathematical Statistics 8, 197–223.
Chernoff, H., 1981. A note on an inequality involving the normal distribution. The Annals of Probability 9 (3), 533–535.
Chiogna, M., 2005. A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. Statistical Methods and Applications 14 (3), 331–341.
Duerinckx, M., Ley, C., Swan, Y., 2014. Maximum likelihood characterization of distributions. Bernoulli 20 (2), 775–802.
Ernst, M., Reinert, G., Swan, Y., 2020. First order covariance inequalities via Stein's method. Bernoulli to appear.
Gauss, C.F., 1809. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Cambridge Library Collection. Cambridge University Press, Cambridge. Reprint of the 1809 original..
Hallin, M., Ley, C., 2012. Skew-symmetric distributions and Fisher information–a tale of two densities. Bernoulli 18 (3), 747–763.

Hallin, M., Ley, C., 2014. Skew-symmetric distributions and Fisher information: the double sin of the skew-normal. Bernoulli 20 (3), 1432–1453.
Ho, N., Nguyen, X., 2019. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. SIAM Journal on Mathematics of Data Science 1, 730–758.
Jones, M.C., Faddy, M.J., 2003. A skew extension of the t-distribution, with applications. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 159–174.
Klaassen, C.A., 1985. On an inequality of Chernoff. The Annals of Probability 13 (3), 966–974.
Kusuoka, S., Tudor, C.A., 2012. Stein's method for invariant measures of diffusions via Malliavin calculus. Stochastic Processes and their Applications 122 (4), 1627–1651.
Le Cam, L., 1986. The Central Limit Theorem around 1935. Statistical Science 1, 78–96.
Ley, C., 2015. Flexible modelling in statistics: past, present and future. Journal de la Société Française de Statistique 156, 76–96.
Ley, C., Paindaveine, D., 2010. On the singularity of multivariate skew-symmetric models. Journal of Multivariate Analysis 101 (6), 1434–1444.
Ley, C., Reinert, G., Swan, Y., 2017. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. Annals of Applied Probability 27, 216–241.
Ley, C., Reinert, G., Swan, Y., 2017. Stein's method for comparison of univariate distributions. Probability Surveys 14, 1–52.
Ley, C., Swan, Y., 2013. Stein's density approach and information inequalities. Electronic Communications in Probability 18 (7), 1–14.
Ley, C., Swan, Y., 2016. A general parametric Stein characterization. Statistics & Probability Letters 111, 67–71.
Ley, C., Swan, Y., 2016. Parametric Stein operators and variance bounds. Brazilian Journal of Probability and Statistics 30 (2), 171–195.
Liu, Q., Lee, J.D., Jordan, M.I., 2016. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In: International Conference on Machine Learning, pp. 276–284.
Marshall, A.W., Olkin, I., 1993. Maximum likelihood characterizations of distributions. Statistica Sinica 3 (1), 157–171.
Nourdin, I., Viens, F., 2009. Density formula and concentration inequalities with Malliavin calculus. Electronic Journal of Probability 14 (78), 2287–2309.
Pewsey, A., 2000. Problems of inference for Azzalini's skew-normal distribution. Journal of Applied Statistics 27 (7), 859–870.
Ross, N., 2011. Fundamentals of Stein's method. Probability Surveys 8, 210–293.
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, Volume 2: Probability Theory. University of California Press.
Stein, C., Diaconis, P., Holmes, S., Reinert, G., 2004. Use of exchangeable pairs in the analysis of simulations. In: Diaconis, P., Holmes, S. (Eds.), Stein's method: expository lectures and applications. IMS Lecture Notes Monogr. Ser, vol. 46, Beachwood, Ohio, USA: Institute of Mathematical Statistics, pp. 1–26.
Stigler, S., 1980. Stigler's law of eponymy. Transactions of the New York Academy of Sciences 39, 147–158.
Teicher, H., 1961. Maximum likelihood characterization of distributions. The Annals of Mathematical Statistics 32 (4), 1214–1222.