Received: 3 December 2024 | Revised: 3 June 2025 | Accepted: 10 June 2025

DOI: 10.1111/sjos.70003

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Stein's method of moments

Bruno Ebner[1] | **Adrian Fischer**[2] | **Robert E. Gaunt**[3] | **Babette Picker**[1] | **Yvik Swan**[4]

[1]Institute of Stochastics, Karlsruher Institut für Technologie, Karlsruhe, Germany

[2]Department of Statistics, University of Oxford, Oxford, UK

[3]Department of Mathematics, The University of Manchester, Manchester, UK

[4]Département de Mathématique, Université Libre de Bruxelles, Brussels, Belgium

Correspondence
Adrian Fischer, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK.
Email: adrian.fischer@stats.ox.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/Y008650/1, UKRI068, EP/T018445/1; Fonds De La Recherche Scientifique - FNRS, Grant/Award Number: CDR/OLJ.0200.24

Abstract

Stein operators allow one to characterize probability distributions via differential operators. Based on these characterizations, we develop a new method of point estimation for marginal parameters of strictly stationary and ergodic processes, which we call *Stein's Method of Moments* (SMOM). These SMOM estimators satisfy the desirable classical properties such as consistency and asymptotic normality. As a consequence of the usually simple form of the operator, we obtain explicit estimators in cases where standard methods such as (pseudo-) maximum likelihood estimation require a numerical procedure to calculate the estimate. In addition, with our approach, one can choose from a large class of test functions, which typically allows for improvements over the moment estimator. Moreover, for i.i.d. observations, we retrieve data-dependent functions that result in asymptotically efficient estimators and give a sequence of explicit SMOM estimators that converge to the maximum likelihood estimator. Our simulation study demonstrates that for a number of important univariate continuous probability distributions, our SMOM estimators possess competitive small sample behavior, in comparison to the maximum likelihood estimator and other widely-used methods in terms of bias and mean squared

error. We also illustrate the pertinence of our approach on a real data set related to rainfall modelization.

KEYWORDS
Stein's method, point estimation, univariate distribution, method of moments

1 | INTRODUCTION

Point estimation in a parametric model is one of the most classical problems in statistics. In the case of independent and identically distributed (i.i.d.) data, maximum likelihood estimation (MLE) can count itself among the most sought-after, which is mostly due to its simple idea and asymptotic efficiency for regular target distributions. On the other hand, several difficulties can occur, including highly complex probability density functions (PDFs), failure of numerical procedures due to local extrema of the likelihood function or to censoring, and the complexity of extending the method to the non-i.i.d. case. Additionally, the efficient asymptotic behavior of the MLE does not necessarily guarantee a high performance for smaller sample sizes.

The method of moments provides a simple alternative to the MLE but requires that the moments of the target distribution can be calculated analytically. This is often the case for basic univariate probability distributions, resulting in an explicit estimator that can serve as an initial guess for the numerical procedure to calculate the MLE. However, if the moments are of a complicated form, the moment estimator itself can only be computed through a numerical algorithm and loses its simplicity. Moreover, it is well-known that moment estimation is in general, outplayed by the MLE regarding the asymptotic behavior in the i.i.d. case. The generalized method of moments was introduced in Hansen (1982) and is applicable for stationary and ergodic time series and does not require an i.i.d. setting. The generalized method of moments incorporates a wide class of estimation techniques, such as MLE and the classical method of moments. A difficulty that comes along with the method is the problem of finding a suitable target function. Moreover, estimation can get numerically tedious if the target function is complicated, and necessitates a first-step estimator if one wishes to minimise the asymptotic variance.

A vast number of alternative estimation techniques have been developed over the years. Amongst others, different kinds of minimum-distance approaches have been considered that compare characterizing functions of the target distributions, such as the Fourier or Laplace transform, to empirical approximations. We refer to (Adler et al., 1998, Chapter 3) (α-stable distributions), Koutrouvelis (1982) (Cauchy distribution), Meintanis (2016) (mixtures of normal distributions), Weber et al. (2006) (Gompertz and Power exponential distribution among others), to name just a few references.

However, the methods mentioned above can run into numerical hardships as soon as the characterizing object used for estimation becomes complicated. In this context, several approaches have been developed based on Stein's characterizations of probability distributions, which lie at the heart of the powerful probabilistic technique Stein's method (Stein (1972)). Through Stein characterizations it is possible to eliminate the normalizing constant; for example, Stein characterizations based on the *density approach* to Stein's method (Ley et al., 2017; Ley and Swan, 2013) involve the ratio p'/p, where p is the density of the target distribution.

Betsch et al. (2021) developed a new class of minimum-distance-type estimators based on Stein characterizations, in which new representations of the cumulative distribution function (CDF), which do not involve the normalizing constant, are obtained in terms of an expectation and compare the respective sample mean to the empirical CDF (see also Betsch and Ebner (2021)). Recently, Barp et al. (2019) (see Oates (2024) for a more recent reference) introduced a new class of estimators obtained through minimizing a Stein discrepancy, whereupon their method incorporates the score matching approach, a further technique to estimate the parameters of non-normalized model based on the score function (see Hyvärinen and Dayan (2005)). However, through these approaches, explicit estimators are only obtained in simple models, and estimation becomes computationally challenging as soon as a numerical procedure is required.

This is where we want to tie in. In this paper, we study a new class of point estimators, which we refer to as *Stein's Method of Moments (SMOM)* estimators, that are obtained through a Stein characterization based on the density approach by applying the corresponding Stein operator to selected test functions and solving the resulting empirical version of the Stein identity for the unknown parameter. This combines the benefits of independence from a possibly complicated normalizing constant and the simplicity of the estimator. A similar idea was already proposed in Arnold et al. (2001), in which the authors considered a generalized version of Hudson's identity to develop parameter estimators for exponential families. In a similar spirit, Wang and Weiß (2023) obtained a Stein-type characterization for the Lindley distribution and derived new estimators based on this characterization, whilst Nik and Weiß (2024) have also used this approach to obtain new parameter estimators for the exponential, inverse Gaussian and negative binomial distributions. However, our work can be seen as an extension in which we consider a larger class of probability distributions and Stein operators. We also develop an asymptotic theory for our Stein estimators, addressing measurability, existence, (strong) consistency, and asymptotic normality for marginal parameters of strictly stationary and ergodic processes, without even the need for an i.i.d. assumption. We make a further contribution by addressing the problem of how to choose "optimal" test functions that result in asymptotically efficient estimators, and we are able to obtain sequences of explicit Stein estimators that converge to the MLE.

Stein's method of moments is a rather universal approach to parameter estimation. Stein's density approach yields tractable Stein characterizations for many of the most important univariate distributions, and, with a suitable Stein characterization at hand, one can readily deduce estimators with the following desirable features: (i) simple, explicit moment estimators, which through suitable choices of test functions, typically offer improvements on the usual moment estimators in terms of efficiency or mean squared error (MSE); or (ii) asymptotically efficient estimators that remain fully explicit. As illustrated in the simulations presented in Section 3 and the Supporting Information, we observe that SMOM estimators often possess good small-sample behavior. Moreover, as discussed in Section 4, SMOM has recently been extended to multivariate continuous probability distributions (see Fischer et al. (2024, 2025)), and the performance of the estimators in this setting remains competitive, which demonstrates the versatility of SMOM for challenging estimation problems beyond the univariate setting of this article. We hope that this paper will inspire further research into this method, and hope to see it further extended to a discrete and multivariate setting beyond the recent work Fischer et al. (2024, 2025).

The rest of this paper is organized as follows. In Section 2.1, we provide the basic results concerning the density approach in the framework of Stein's method, which is used for the purpose of characterizing the target distribution. In Section 2.2, we present the notation, terminology, and setting that are employed in this paper. In Section 2.3, we introduce our new class of SMOM

estimators and deal with questions of existence, measurability, consistency, and asymptotic normality. In Section 2.4, we investigate the problem of how to choose the best possible estimator in the SMOM estimators in terms of asymptotic variance and develop a new type of two-step estimators that reach asymptotic efficiency. Moreover, we recover the MLE through an iteratively defined sequence of the aforementioned estimators. In Section 3, we provide applications of our new estimators to parameter estimation problems concerning the truncated normal, Cauchy, and exponential polynomial models, as well as the Nakagami distribution, in which our estimators are applied to a real data set related to rainfall modelization. We compare the new estimators to other available approaches by means of competitive simulation studies.

In the Section A of the Supporting Information, we provide further details for Example 2.5 that concerns SMOM for the gamma distribution, whilst in Section B we provide further details for the truncated normal, Cauchy, and exponential polynomial model applications of Section 3. Section C contains additional examples for the beta, Student's t, Lomax, Nakagami, one-sided truncated inverse-gamma distribution, and generalized logistic distributions, and an example for non-i.i.d. random variables with marginal Cauchy density. Finally, the proofs are given in Section D.

2 | STEIN'S METHOD OF MOMENTS

2.1 | Elements of Stein's method

We begin with a short introduction to the version of Stein's method employed in this paper.

Let $\mathbb{P}_\theta$ be a probability distribution on $(a,b) \subset \mathbb{R}$ with corresponding differentiable PDF $p_\theta(x)$ that depends on a parameter $\theta \in \Theta \subset \mathbb{R}^p$, where we assume that $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ implies $\theta_1 = \theta_2$ for $\theta_1, \theta_2 \in \Theta$. Throughout the paper, we assume that Θ is open and convex as well as that $-\infty \leq a < b \leq \infty$ and $p_\theta(x) > 0$ for all $\theta \in \Theta$ and $x \in (a,b)$. Let X be a real-valued random variable with values in (a,b), $\mathcal{F}_\theta$ a class of functions $f : (a,b) \to \mathbb{R}$, and $\mathcal{A}_\theta$ an operator defined on $\mathcal{F}_\theta$. We call $(\mathcal{A}_\theta, \mathcal{F}_\theta)$ a *Stein pair* for $\mathbb{P}_\theta$ if the following is satisfied:

$$\mathbb{E}[\mathcal{A}_\theta f(X)] = 0 \text{ for all } f \in \mathcal{F}_\theta \qquad \text{if and only if} \qquad X \sim \mathbb{P}_\theta; \tag{1}$$

operator $\mathcal{A}_\theta$ is called a *Stein operator* for $\mathbb{P}_\theta$, and $\mathcal{F}_\theta$ is the associated *Stein class*. There exist many ways to obtain Stein pairs for any given distribution, see, for example, Anastasiou et al. (2023) and Ley and Swan (2013). In this paper, we consider those obtained via the density approach as developed in Ley et al. (2017) and Ley and Swan (2013). First-order density approach, Stein operators are of the form

$$\mathcal{A}_\theta f(x) = \frac{(\tau_\theta(x) p_\theta(x) f(x))'}{p_\theta(x)}, \tag{2}$$

where τ_θ is some differentiable function $\tau_\theta : (a,b) \to \mathbb{R}$; they act on the function class

$$\mathcal{F}_\theta := \left\{ f : (a,b) \to \mathbb{R} \mid f \text{ is differentiable and } \int_a^b (f(x)\tau_\theta(x)p_\theta(x))' \, dx = 0 \right\}. \tag{3}$$

The next theorem states that $(\mathcal{A}_\theta, \mathcal{F}_\theta)$ is a Stein pair for $\mathbb{P}_\theta$. In the Supporting Information, we give a proof that goes along the lines of (Ley & Swan, 2013, Theorem 2.2).

Theorem 2.1. *Let $\mathcal{A}_\theta$ be the Stein operator defined in Equation (2) and $\mathcal{F}_\theta$ the corresponding class of functions introduced in Equation (3). Moreover, assume that $\tau_\theta(x) \neq 0$ almost everywhere on (a, b) and let X be a random variable with values in (a, b). Then the Stein characterization (1) holds.*

It is often convenient to use in Equation (2) the so-called *Stein kernel* $\tau_\theta(x) = (1/p_\theta(x))\int_x^b (\mathbb{E}[X] - y)p_\theta(y)\, dy, x \in (a, b)$, whose corresponding density approach Stein operator is

$$\mathcal{A}_\theta f(x) = \tau_\theta(x)f'(x) + (\mathbb{E}[X] - x)f(x). \tag{4}$$

This last operator takes a simple form in many cases. For instance τ_θ is polynomial for members of the Pearson family (see (Stein, 1986, Theorem 1, p. 65) and (Gaunt et al., 2019, Lemma 2.9)). We refer to (Ernst et al., 2020; Saumard, 2019) for an overview of Stein kernels and their properties. Other choices of functions τ_θ in Equation (2) are also sometimes better suited. We close this introductory section on Stein's method with two simple examples of how the density approach can be used to find suitable Stein operators, these being for the Gaussian and gamma distributions. In Sections 2.3 and 2.4, we will use the Gaussian and gamma distributions as running examples that demonstrate the application of Stein's method of moments in an uncomplicated manner; more involved applications are given in Section 3.

Example 2.1 (Gaussian distribution). Consider the Gaussian distribution $N(\mu, \sigma^2)$ with parameter $\theta = (\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$ and density $p_\theta(x) = 1/\sqrt{2\pi\sigma^2}\exp(-(x-\mu)^2/(2\sigma^2))$, $x \in \mathbb{R}$. A simple calculation gives that the Stein kernel is $\tau_\theta(x) = \sigma^2$. We have $\mathbb{E}[X] = \mu$ and retrieve from Equation (4) the well-known Stein operator of Stein (1972),

$$\mathcal{A}_\theta f(x) = \sigma^2 f'(x) + (\mu - x)f(x). \tag{5}$$

Example 2.2 (Gamma distribution). Consider the gamma distribution $\Gamma(\alpha, \beta)$ with parameter $\theta = (\alpha, \beta)$, $\alpha, \beta > 0$, and density $p_\theta(x) = \beta^\alpha x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$, $x > 0$. The Stein kernel is $\tau_\theta(x) = x$. Since $\mathbb{E}[X] = \alpha/\beta$, we recover the gamma Stein operator of Diaconis and Zabell (1991),

$$\mathcal{A}_\theta f(x) = xf'(x) + (\alpha - \beta x)f(x). \tag{6}$$

2.2 | Notation and setting

Let $\{X_n, n \in \mathbb{Z}\}$ be a real-valued strictly stationary and ergodic discrete-time process defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. To clarify this terminology, we elaborate on what we mean by strict stationarity and ergodicity. We say that $\{X_n, n \in \mathbb{Z}\}$ is strictly stationary if $\{X_n, n \in \mathbb{Z}\} =_D \{X_{n+k}, n \in \mathbb{Z}\}$ for each $k \in \mathbb{Z}$. Moreover, let $\zeta : \Omega \to \Omega$ be measurable such that $X_{n+1}(\omega) = X_n(\zeta(\omega))$ for each $\omega \in \Omega$ and $n \in \mathbb{N}$. Then we say that $\{X_n, n \in \mathbb{Z}\}$ is ergodic if ζ is measure-preserving ($\mathbb{P}(\zeta^{-1}(A)) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$) and the σ-algebra of invariant events $\mathcal{I} = \{A \in \mathcal{F} \mid \zeta^{-1}(A) = \zeta(A)\}$ is $\mathbb{P}$-trivial, that is, $\mathbb{P}(A) \in \{0, 1\}$ for all $A \in \mathcal{I}$. We assume that the marginal distribution of each X_n, $n \in \mathbb{Z}$, is $\mathbb{P}_{\theta_0}$ for some $\theta_0 \in \Theta$. Now suppose that the measures $\mathbb{P}_\theta$ are characterized through Stein pairs $(\mathcal{A}_\theta, \mathcal{F}_\theta)$ and let $\mathcal{F} = \cap_{\theta\in\Theta}\mathcal{F}_\theta$.

In this paper, we will also employ the following notation. For a real-valued function $(\theta, x) \mapsto g_\theta(x)$, where $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}$, we write $\frac{\partial}{\partial\theta}g_\theta(x)$ for its gradient with respect to $\theta = (\theta_1, \ldots, \theta_p)^\top$,

which is a column vector of size p. If the function $g_\theta(x)$ takes values in $\mathbb{R}^q$ with $q \geq 2$, $\frac{\partial}{\partial\theta} g_\theta(x)$ is its Jacobian with respect to θ, which is a $(q \times p)$-matrix. By $\frac{\partial}{\partial\theta_i} g_\theta(x)$ we mean the partial derivative with respect to θ_i. If we want to address the derivative with respect to the argument in the parentheses (with respect to x in $g_\theta(x)$) we simply write $g'_\theta(x)$ which remains real-valued or a column vector of size q. For a vector $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, we denote by $\|x\| = (x_1^2 + \cdots + x_n^2)^{1/2}$ the standard Euclidean norm. For a (possibly non-square) matrix $M \in \mathbb{R}^{p\times q}$, we let $\|M\|$ be the spectral norm, which is defined as the square-root of the largest eigenvalue of $X^\top X$. We also introduce the vectorization map that stacks the columns of a matrix $M = (m_{i,j}, 1 \leq i \leq p, 1 \leq j \leq q) \in \mathbb{R}^{p\times q}$, given by $\mathrm{vec} : \mathbb{R}^{p\times q} \to \mathbb{R}^{pq} : M \mapsto (m_{1,1}, \ldots, m_{p,1}, m_{1,2}, \ldots, m_{p,2}, \ldots, \ldots, m_{p,q})^\top$. Finally, we will write $\xrightarrow{a.s.}$ for convergence almost surely, $\xrightarrow{\mathbb{P}}$ for convergence in probability, and $\xrightarrow{D}$ for convergence in distribution.

2.3 | Stein's method of moments: Definition and properties

For the purpose of estimating the unknown parameter θ_0 from a sample $X_1, \ldots, X_n$ drawn from a real-valued strictly stationary and ergodic discrete-time process $\{X_n, n \in \mathbb{Z}\}$, we choose p measurable test functions $f_1, \ldots, f_p$ (belonging to $\mathcal{F}$) and, in light of (1), replace the expectations with their empirical counterparts. Therefore, we get the following system of equations:

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{A}_\theta f(X_i) = 0, \tag{7}$$

where we write $\mathcal{A}_\theta f : \Theta \times (a,b) \to \mathbb{R}^p$ for the function defined by $(\theta, x) \mapsto \mathcal{A}_\theta f(x) := (\mathcal{A}_\theta f_1(x), \ldots, \mathcal{A}_\theta f_p(x))^\top$.

In the following, we will refer to (7) as the *empirical* Stein identity. Moreover, we will call any solution to this system of equations with respect to θ a *Stein estimator*, which we denote by $\hat{\theta}_n$.

With this definition at hand, one observes that Stein estimators can be seen as moment estimators (resp. generalized moment estimators as proposed in Hansen (1982)), whereupon we suggest suitable target functions through Stein's method.

Remark 2.1. We give a few more details on the related approaches of score matching and minimum Stein discrepancy:

(1) Score matching: The score matching estimator (Hyvärinen, 2007; Hyvärinen and Dayan, 2005) aims to minimise the Fisher-Hyvärinen distance between the log-densities of the observed data and the model, and is defined through

$$\operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial x^2} \log p_\theta(x)\Big|_{x=X_i} + \frac{1}{2}\left(\frac{\partial}{\partial x} \log p_\theta(x)\Big|_{x=X_i}\right)^2.$$

Therefore, score matching can be seen as a minimum distance estimation technique and is also independent of the normalizing constant. If θ is the natural parameter of an exponential family and under some additional technical assumptions, the score matching estimator can be recovered with our approach

by the test function choice $f(x) = \frac{\partial^2}{\partial x \partial \theta} \log p_\theta(x)$. Meanwhile, the score matching method has been generalized to more complicated settings (see, e.g., (Liu et al., 2022; Lyu, 2009; Yu et al., 2022) and the references therein). We compare the Stein estimator to the score matching estimator in Section 3.3 for exponential polynomial models.

(2) Minimum Stein discrepancy: The Minimum Stein discrepancy estimator (Barp et al., 2019; Oates, 2024) is based on the distance

$$\sup_{f \in \mathcal{F}} |\mathbb{E}[\mathcal{A}_\theta f(X)]|^2,$$

where $\mathcal{A}_\theta$ is a density approach Stein operator for $\mathbb{P}_\theta$ and $X \sim \mathbb{Q}$, where $\mathbb{Q}$ is the true distribution of the data. Therefore, minimum Stein discrepancy estimators are independent of the normalizing constant and are also minimum distance estimators. For particular choices of $\mathcal{F}$ one can compute the supremum above in terms of an integral with respect to $\mathbb{Q}$ which is then estimated by means of the empirical distribution of $X_1, \ldots, X_n$. In Barp et al. (2019) the authors consider, for example, reproducing kernel Hilbert spaces for $\mathcal{F}$. The score matching estimator can be recovered by choosing the appropriate density Stein operator and function class. We compute a minimum Stein discrepancy estimator based on a reproducing kernel Hilbert space for the gamma distribution in Section A.2.

The necessary conditions on the test functions $f_1, \ldots, f_p$, the Stein operator $\mathcal{A}_\theta$ and the target distribution $\mathbb{P}_\theta$ to achieve existence, measurability and asymptotic normality of the Stein estimator $\hat{\theta}_n$ will be introduced below. Subsequently, we will impose the following assumptions.

Assumption 2.1.

(a) Let $X \sim \mathbb{P}_{\theta_0}$ and $\theta \in \Theta$. Then $f = (f_1, \ldots, f_p) \in \mathcal{F}$ is such that $\mathbb{E}[\mathcal{A}_\theta f(X)] = 0$ if and only if $\theta = \theta_0$.

(b) Let $q \geq p$ and $X \sim \mathbb{P}_{\theta_0}$. We can write $\mathcal{A}_\theta f(x) = M(x)g(\theta)$ for some measurable $p \times q$ matrix M with $\mathbb{E}[\|M(X)\|] < \infty$ and a continuously differentiable function $g = (g_1, \ldots, g_q)^\top : \Theta \to \mathbb{R}^q$ for all $\theta \in \Theta, x \in (a, b)$. We also assume that $\mathbb{E}[M(X)] \frac{\partial}{\partial \theta} g(\theta)|_{\theta=\theta_0}$ is invertible.

Assumption 2.1 (a) ensures that the true parameter θ_0 can be well identified by means of the Stein operator $\mathcal{A}_\theta$; this assumption can be easily verified (for a proper choice of test functions) for operators of the form (2) with Theorem 2.1. Assumption 2.1 (b) requires that the parameters can be well separated from the sample. Moreover, if the function g is fairly simple, we are likely to obtain explicit estimators; this turns out to be the case for all examples considered in this paper.

Theorem 2.2. *Suppose Assumption 2.1 (a), (b) is satisfied. The probability that a solution to (7), $\hat{\theta}_n$, exists and is measurable converges to 1 as $n \to \infty$. Furthermore, $\hat{\theta}_n$ is strongly consistent in the following sense: There is a set $A \subset \Omega$ with $\mathbb{P}(A) = 1$ such that for all $\omega \in A$ there exists $N = N(\omega) \in \mathbb{N}$ such that $\hat{\theta}_n$ exists for all $n \geq N$ and $\hat{\theta}_n(\omega) \to \theta_0$.*

As we will see in Section 3 and the further examples in the Supporting Information, the new estimators will mostly be solutions to systems of linear equations which exist and are measurable with probability 1 for any sample size if $\Theta = \mathbb{R}^p$. Nonetheless, it can happen that an estimator returns a value which lies outside of the truncation domain if the parameter space is a strict subset

of $\mathbb{R}^p$. These issues will be addressed separately for each example in Section 3 and the further examples in the Supporting Information.

Asymptotic normality can be obtained similarly to the classical moment estimators. We state the result in the next theorem. We slightly change the meaning of $\hat{\theta}_n$ as we need our estimator to be a random variable to establish weak convergence. To this end, let $X \sim \mathbb{P}_{\theta_0}$. Define the function $F(M,\theta) = Mg(\theta)$, where $M \in \mathbb{R}^{p\times q}$ and g is as in Assumption 2.1 (b). In the proof of Theorem 2.2, we see that there are neighborhoods $U \subset \mathbb{R}^{p\times q}$, $V \subset \mathbb{R}^p$ of $\mathbb{E}[M(X)]$ and θ_0 such that there exists a continuously differentiable function $h : U \to V$ with $F(M, h(M)) = 0$ for all $M \in U$. We now define $A_n = \left\{n^{-1}\sum_{i=1}^n M(X_i) \in U\right\}$, and note that it is shown in the proof of Theorem 2.2 that $\mathbb{P}(A_n) \to 1$ as $n \to \infty$. We now extend h to a differentiable function $\tilde{h}$ on $\mathbb{R}^{p\times q}$ and define $\hat{\theta}_n = \tilde{h}\left(n^{-1}\sum_{i=1}^n M(X_i)\right)$ (note that such an extension always exists by choosing U small enough). We will also need some further assumptions, which can be efficiently stated by recalling a version of a central limit theorem for strictly stationary and ergodic time series stated in Hannan (1973) and originally proved in Gordin (1969) (see also Hansen (1982)).

Theorem 2.3. *Let* $\{Y_n, n \in \mathbb{Z}\}$ *be vector-valued, strictly stationary and ergodic. Moreover, suppose that* $\mathbb{E}[Y_1] = 0$ *and* $\mathbb{E}[\|Y_1 Y_1^\top\|] < \infty$ *as well as* $\mathbb{E}[\mathbb{E}[Y_0 \mid Y_{-j},\ldots]\mathbb{E}[Y_0 \mid Y_{-j},\ldots]^\top] \to 0$, $j \to \infty$. *Furthermore, for* $Y_j' = \mathbb{E}[Y_0 \mid Y_{-j},\ldots] - \mathbb{E}[Y_0 \mid Y_{-j-1},\ldots]$, $j \geq 0$, *we suppose that* $\sum_{j=0}^\infty \mathbb{E}[(Y_j')^\top Y_j']^{1/2} < \infty$. *Then* $n^{-1/2}\sum_{i=1}^n Y_i \xrightarrow{D} N(0,\Xi)$, *where* $\Xi = \sum_{i\in\mathbb{Z}} \mathbb{E}[Y_0 Y_i^\top]$.

We can now state our asymptotic normality result.

Theorem 2.4. *Let* $X \sim \mathbb{P}_{\theta_0}$. *Suppose Assumption 2.1* (*a*), (*b*) *is fulfilled. Moreover, assume that the matrix* $\mathbb{E}[\mathrm{vec}(M(X))\mathrm{vec}(M(X))^\top]$ *exists and that the time series* $\{Y_n, n \in \mathbb{Z}\}$, *where* $Y_n = \mathrm{vec}(M(X_n)) - \mathbb{E}[\mathrm{vec}(M(X))]$, $n \in \mathbb{Z}$, *satisfies the assumptions of Theorem 2.3. Now let*

$$\Psi = \sum_{j\in\mathbb{Z}} \mathbb{E}\left[\mathcal{A}_{\theta_0} f(X_0)\mathcal{A}_{\theta_0} f(X_j)^\top\right] \quad \text{and} \quad G = \mathbb{E}\left[\frac{\partial}{\partial\theta}\mathcal{A}_\theta f(X)\Big|_{\theta=\theta_0}\right],$$

and let $\hat{\theta}_n$ *be defined as in the preceding paragraph. Then the sequence* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $G^{-1}\Psi G^{-\top}$.

Remark 2.2. Note that in the case where $\{X_n, n \in \mathbb{Z}\}$ is i.i.d., the assumptions of Theorem 2.3 are easily verified and the matrix Ψ appearing in the asymptotic covariance simplifies to

$$\Psi = \mathbb{E}\left[\mathcal{A}_{\theta_0} f(X)\mathcal{A}_{\theta_0} f(X)^\top\right], \quad X \sim \mathbb{P}_{\theta_0}.$$

Let us consider two simple examples to demonstrate our estimation method and its flexibility. For that purpose, we write $\overline{f(X)} = n^{-1}\sum_{i=1}^n f(X_i)$ for a measurable function $f : (a,b) \to \mathbb{R}$.

Example 2.3 (Gaussian distribution, continuation of Example 2.1). Since we have two unknown parameters, we choose two test functions f_1, f_2, and therefore from Equation (5) get

$$\begin{cases} \overline{f_1(X)}\mu + \overline{f_1'(X)}\sigma^2 = \overline{Xf_1(X)} \\ \overline{f_2(X)}\mu + \overline{f_2'(X)}\sigma^2 = \overline{Xf_2(X)}. \end{cases}$$

By solving this system of linear equations for μ and σ^2 we obtain the Stein estimators

$$\hat{\mu}_n = \frac{\overline{f_2'(X)}\,\overline{Xf_1(X)} - \overline{f_1'(X)}\,\overline{Xf_2(X)}}{\overline{f_1(X)}\,\overline{f_2'(X)} - \overline{f_1'(X)}\,\overline{f_2(X)}}, \quad \hat{\sigma}_n^2 = \frac{\overline{f_1(X)}\,\overline{Xf_2(X)} - \overline{f_2(X)}\,\overline{Xf_1(X)}}{\overline{f_1(X)}\,\overline{f_2'(X)} - \overline{f_1'(X)}\,\overline{f_2(X)}}. \tag{8}$$

Taking $f_1(x) = 1$, $f_2(x) = x$ yields the MLE or moment estimators $\hat{\mu}_n = \overline{X}$ and $\hat{\sigma}_n^2 = \overline{X^2} - \overline{X}^2$.

Here, we have $\mathrm{vec}(M(x)) = \begin{pmatrix} f_1'(x) & f_2'(x) & f_1(x) & f_2(x) & xf_1(x) & xf_2(x) \end{pmatrix}^\top$. Let $\{Z_n,\ n \in \mathbb{Z}\}$ be the Gaussian AR(1) process, that is, $Z_n = \rho Z_{n-1} + \epsilon_n$, where $|\rho| < 1$ and ϵ_n, $n \in \mathbb{N}$, are i.i.d. centered Gaussian with variance σ_ϵ^2. We therefore have that $(Z_n, \ldots, Z_{n+k})$ are multivariate Gaussian for every $k \in \mathbb{N}$ with zero mean and $\mathrm{Cov}(Z_{n_1}, Z_{n_2}) = \rho^{|n_1 - n_2|}\sigma_\epsilon^2/(1-\rho^2)$. Let then $X_n = Z_n + \mu$ for $n \in \mathbb{Z}$ and $\mu_0 \in \mathbb{R}$. In this setting, we wish to estimate the two marginal parameters μ_0 and $\sigma_0^2 = \sigma_\epsilon^2/(1-\rho^2)$ from a sample $X_1, \ldots, X_n$. For all test functions f_1, f_2 that satisfy the assumptions of Theorem 2.4, we have that the matrices for the asymptotic variance of the estimators in Equation (8) are given by

$$\Psi_{k,l} = \sum_{j \in \mathbb{Z}} \mathbb{E}\Big[\big(\sigma_0^2 f_k'(X_0) + (\mu_0 - X_0) f_k(X_0)\big)\big(\sigma_0^2 f_l'(X_j) + (\mu_0 - X_j) f_l(X_j)\big)\Big], \quad k, l = 1, 2,$$

$$G = \begin{pmatrix} \mathbb{E}[f_1(X_0)] & \mathbb{E}[f_1'(X_0)] \\ \mathbb{E}[f_2(X_0)] & \mathbb{E}[f_2'(X_0)] \end{pmatrix}.$$

For the test functions $f_1(x) = 1$ and $f_2(x) = x$ we get $\mathrm{vec}(M(x)) = \begin{pmatrix} 0 & 1 & 1 & x & x & x^2 \end{pmatrix}^\top$ and moreover

$$\mathbb{E}[X_0 - \mathbb{E}[X_0] \mid X_{-j}] = \rho^j (X_{-j} - \mu_0)$$

and

$$\mathbb{E}[X_0^2 - \mathbb{E}[X_0^2] \mid X_{-j}] = \rho^{2j}(X_{-j} - \mu_0) + \sigma_\epsilon^2 \left(\sum_{i=0}^{j} \rho^{2i} - \frac{1}{1-\rho^2} \right) + 2\mu_0 \rho^j (X_{-j} - \mu).$$

We also give the formula

$$\begin{aligned} &\mathbb{E}[X_0^2 - \mathbb{E}[X_0^2] \mid X_{-j}] - \mathbb{E}[X_0^2 - \mathbb{E}[X_0^2] \mid X_{-j-1}] \\ &\quad = (X_{-j-1} - \mu_0)(1-\rho)\rho^{2j+1} + \epsilon_{-j}(\rho^{2j} + 2\mu_0 \rho^j) - \sigma_\epsilon^2 \rho^{2(j+1)} \end{aligned}$$

and conclude that all assumptions from Theorem 2.4 are satisfied. Then we have asymptotic normality of the estimators $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ with

$$\Psi_{1,1} = \frac{\sigma_0^2}{1-\rho}, \quad \Psi_{1,2} = \Psi_{2,1} = \frac{\mu_0 \sigma_0^2}{1-\rho}, \quad \Psi_{2,2} = \frac{(2+\mu_0^2)\sigma_0^2}{1-\rho}, \quad G = \begin{pmatrix} 1 & 0 \\ \mu_0 & 1 \end{pmatrix},$$

from which we obtain that the asymptotic covariance matrix is given by

$$G^{-1}\Psi G^{-T} = \operatorname{diag}\left(\frac{\sigma_0^2}{1-\rho}, \frac{2\sigma_0^2}{1-\rho}\right).$$

Example 2.4 (Gamma distribution, continuation of Example 2.2). We choose two different test functions f_1, f_2, and from Equation (6) we readily obtain the estimators

$$\hat{\alpha}_n = \frac{\overline{Xf_2(X)}\;\overline{Xf_1'(X)} - \overline{Xf_1(X)}\;\overline{Xf_2'(X)}}{\overline{Xf_1(X)}\;\overline{f_2(X)} - \overline{f_1(X)}\;\overline{Xf_2(X)}}, \quad \hat{\beta}_n = \frac{\overline{f_2(X)}\;\overline{Xf_1'(X)} - \overline{f_1(X)}\;\overline{Xf_2'(X)}}{\overline{Xf_1(X)}\;\overline{f_2(X)} - \overline{f_1(X)}\;\overline{Xf_2(X)}}.$$

By choosing $f_1(x) = 1$ and $f_2(x) = x$ we retrieve the moment estimators

$$\hat{\alpha}_n^{\mathrm{MO}} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2} \quad \text{and} \quad \hat{\beta}_n^{\mathrm{MO}} = \frac{\overline{X}}{\overline{X^2} - \overline{X}^2}.$$

Moreover, by choosing $f_1(x) = 1$ and $f_2(x) = \log x$ we obtain the logarithmic estimators

$$\hat{\alpha}_n^{LOG} = \frac{\overline{X}}{\overline{X}\,\overline{\log X} - \overline{X}\;\overline{\log X}} \quad \text{and} \quad \hat{\beta}_n^{LOG} = \frac{1}{\overline{X}\,\overline{\log X} - \overline{X}\;\overline{\log X}}, \tag{9}$$

which show a behavior close to asymptotic efficiency and were obtained through the generalized gamma distribution in Ye and Chen (2017) (see also Wiens et al. (2003) for an earlier reference).

2.4 | Optimal functions

We show that it is possible to achieve asymptotic efficiency under certain regularity conditions using Stein estimators by using specific parameter-dependent test functions. To this end, we suppose in this section without further notice that the sequence of random variables $\{X_n, n \in \mathbb{Z}\}$ is i.i.d. (for possible extensions to non-i.i.d. data see Remark 2.3). In addition, we assume that the Stein operator $\mathcal{A}_\theta$ can be written in the form (2). Within this framework, we compare our estimators to the MLE, which we will denote by $\hat{\theta}_n^{\mathrm{ML}}$, and which, under certain regularity conditions on the likelihood function, is defined through the equation

$$\frac{\partial}{\partial\theta}\overline{\log p_\theta(X)}\Big|_{\theta=\hat{\theta}_n^{\mathrm{ML}}} = 0. \tag{10}$$

It is well-known that for regular probability distributions, the expectation of the latter expression is equal to zero. It is a standard result that, under certain regularity conditions, a suitable standardization of the MLE $\hat{\theta}_n^{\mathrm{ML}}$ is asymptotically efficient with covariance matrix $I_{\mathrm{ML}}^{-1}(\theta_0)$, the inverse of the Fisher-information matrix $I_{\mathrm{ML}}(\theta)$.

Motivated by the definition of the MLE, we consider the score function as the right-hand side of the Stein identity

$$\mathcal{A}_\theta f(x) = \frac{\partial}{\partial\theta}\log p_\theta(x). \tag{11}$$

This is an ordinary differential equation whose solution f_θ clearly depends on the unknown parameter θ. If the Stein operator is of the form (2), then the solution of (11) is given by

$$f_\theta(x) = \left(f_\theta^{(1)}(x),\ldots,f_\theta^{(p)}(x)\right)^\top = \frac{\frac{\partial}{\partial\theta}P_\theta(x) + c}{\tau_\theta(x)p_\theta(x)}, \quad x \in (a,b), \tag{12}$$

where $c \in \mathbb{R}$ and P_θ is the CDF corresponding to p_θ, with the convention that $f_\theta(x) = 0$ at all $x \in (a,b)$ such that $\tau_\theta(x) = 0$. We will refer to the functions (12) as the *optimal functions*. Thus, $\overline{\mathcal{A}_\theta f_\theta(X)} = 0$ is the maximum likelihood equation rewritten in terms of Stein operators. One can now use a consistent first-step estimator $\tilde{\theta}_n$ for the unknown parameter θ in $f_\theta = \left(f_\theta^{(1)},\ldots,f_\theta^{(p)}\right)^\top$ and resolve the system of equations (7) with respect to these test functions. This holds the advantage that estimators may remain explicit if the Stein operator is simple. Mathematically speaking, given a first-step estimator $\tilde{\theta}_n$, we define $\hat{\theta}_n^\star$ through the equation

$$\overline{\mathcal{A}_{\hat{\theta}_n^\star} f_{\tilde{\theta}_n}(X)} = 0 \tag{13}$$

if such a solution exists. In this setting, the matrix M from Assumption 2.1 (b) depends on the parameter θ through data-dependent test functions. Hence, we introduce a new set of assumptions.

Assumption 2.2.

(a) $\tilde{\theta}_n$ is a consistent estimator, that is, $\tilde{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

(b) Let $X \sim \mathbb{P}_{\theta_0}$ and $\theta_1, \theta_2 \in \Theta$. Then $f_{\theta_2} \in \mathscr{F}$, and $\mathbb{E}[\mathcal{A}_{\theta_1} f_{\theta_2}(X)] = 0$ if and only if $\theta_1 = \theta_0$.

(c) For $q \geq p$, we can write $\mathcal{A}_{\theta_1} f_{\theta_2}(x) = M_{\theta_2}(x)g(\theta_1)$ for some measurable $p \times q$ matrix M_{θ_2} and a continuously differentiable function $g = (g_1,\ldots,g_q)^\top : \Theta \to \mathbb{R}^q$ for all $\theta_1, \theta_2 \in \Theta$, $x \in (a,b)$. Moreover, we assume that $\mathbb{E}[M_{\theta_0}(X)]\frac{\partial}{\partial\theta}g(\theta)|_{\theta=\theta_0}$, where $X \sim \mathbb{P}_{\theta_0}$, is invertible and the function $\theta \mapsto \mathrm{vec}(M_\theta(x))$ is continuously differentiable on Θ for all $x \in (a,b)$.

(d) For $X \sim \mathbb{P}_{\theta_0}$ there exist two functions F_1, F_2 on (a,b) with $\mathbb{E}[F_i(X)] < \infty$, $i = 1,2$, and compact neighborhoods Θ', Θ'' of θ_0 such that $\|M_\theta(x)\| \leq F_1(x)$ for all $\theta \in \Theta'$ and $\|\frac{\partial}{\partial\theta}\mathrm{vec}(M_\theta(x))\| \leq F_2(x)$ for all $\theta \in \Theta''$, $x \in (a,b)$.

Assumption 2.2 (b), (c) are adapted versions of Assumption 2.1 (a), (b) with the supplement that the optimal function f_θ needs to be an element of the Stein class $\mathscr{F}$ for each $\theta \in \Theta$. The invertibility of $\mathbb{E}[M_{\theta_0}(X)]\frac{\partial}{\partial\theta}g(\theta)|_{\theta=\theta_0}$, $X \sim \mathbb{P}_{\theta_0}$, in (c) is easily verified for Stein operators that are linear in θ. However, we have the additional Assumption 2.2 (d), which can be tedious to verify if f_θ is complicated. Nevertheless, the latter assumption is satisfied for all our applications in Section 3 and the Supporting Information.

Theorem 2.5. *Suppose Assumption 2.2 (a) to (d) are fulfilled. The probability that a solution to (13), $\hat{\theta}_n^\star$, exists and is measurable converges to 1 as $n \to \infty$.*

In the following theorem, we show that, under some additional technical assumptions, the two-step Stein estimators are asymptotically normal and reach asymptotic efficiency. Again, to manoeuvre around existence and measurability issues, we define $\hat{\theta}_n^\star = \tilde{h}\left(n^{-1}\sum_{i=1}^n M_{\tilde{\theta}_n}(X_i)\right)$, where $\tilde{h}$ is the differentiable extension to $\mathbb{R}^{p\times q}$ of the function h from the proof of Theorem 2.5. In view

of the proof of Theorem 2.5, it follows immediately from the continuous mapping theorem that we have $\hat{\theta}_n^\star \xrightarrow{\mathbb{P}} \theta_0$.

Theorem 2.6. *Suppose that Assumption 2.2 (a) to (d) are satisfied. Moreover, assume that p_θ is differentiable with respect to θ and*

(i) the sequence of random vectors $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is uniformly tight;
(ii) $\mathcal{A}_\theta$ is of the form (2) with τ_θ differentiable with respect to θ;
(iii) we have $\lim_{x\to a,b} \frac{\partial}{\partial\theta}(p_\theta(x)\tau_\theta(x))\Big|_{\theta=\theta_0} f_{\theta_0}(x) = 0$;
(iv) and $I_{\mathrm{ML}}(\theta_0)$ exists and is finite.

Then, for $\hat{\theta}_n^\star$ as defined in the preceding paragraph, $\sqrt{n}(\hat{\theta}_n^\star - \theta_0) \xrightarrow{D} N(0, I_{\mathrm{ML}}^{-1}(\theta_0))$, as $n \to \infty$.

We refer the reader as well to (Newey & McFadden, 1994, section 6), in which the asymptotic theory of two-step estimators is studied—although under slightly different assumptions and with the additional restriction that the first-step estimate needs to be obtained through the generalized method of moments.

Remark 2.3. It is possible to extend the results from Theorems 2.5 and 2.6 to strictly stationary and ergodic time series as introduced in Section 2.3. For Theorem 2.5, it suffices to apply an adapted uniform strong law of large numbers as stated in (Hansen, 2012, Theorem 2.1) (note that with Assumption 2.2 (d) the random function $\theta \to M_\theta(X), X \sim \mathbb{P}_{\theta_0}$, is automatically first-moment-continuous, compare (DeGroot, 2005, p. 206)). With the latter result together with Theorem 2.3 we can also generalize Theorem 2.6, although we need the additional assumption that the sequence $\{\mathcal{A}_{\theta_0} f_{\theta_0}(X_n), n \in \mathbb{Z}\}$ satisfies the assumptions of Theorem 2.3. We then get

$$\sqrt{n}(\hat{\theta}_n^\star - \theta_0) \xrightarrow{D} N\left(0, I_{\mathrm{ML}}^{-1}(\theta_0)\left(\sum_{j\in\mathbb{Z}} \mathbb{E}\left[\mathcal{A}_{\theta_0} f_{\theta_0}(X_0)\mathcal{A}_{\theta_0} f_{\theta_0}(X_j)^\top\right]\right) I_{\mathrm{ML}}^{-1}(\theta_0)\right), \quad n \to \infty.$$

Remark 2.4. There is another possibility to achieve the asymptotic efficiency of point estimators. Carrasco and Florens (2000) proposed a generalized method-of-moments-type estimator with a continuum of moment conditions. The idea is based on using an uncountably infinite number of moment conditions, that is, a class of functions $h^t : \Theta \times (a,b) \to \mathbb{R}, t \in \Pi \subset \mathbb{R}$ such that $\mathbb{E}[h^t(\theta_0, X)] = 0$ for all $t \in \Pi$, where $X \sim \mathbb{P}_{\theta_0}$. Under some conditions, the sequence of functions $n^{-1/2}\sum_{i=1}^n h^t(\theta, X_i)$, $t \in \Pi$, converges to some zero-mean Gaussian process with covariance operator Υ by the functional central limit theorem as $n \to \infty$. Let Υ^{α_n} be its Tikhonov regularization with smoothing term α_n. Then, under additional assumptions, it can be shown that the estimator

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \left\| (\Upsilon_n^{\alpha_n})^{-1/2} h_n^t(\theta, X) \right\|_{\mathscr{L}^2},$$

where $(\Upsilon_n^{\alpha_n})^{-1/2}$ is an estimate of $(\Upsilon^{\alpha_n})^{-1/2}$, $h_n^t(\theta, X) = n^{-1}\sum_{i=1}^n h^t(\theta, X_i)$ and $\|\cdot\|_{\mathscr{L}^2}$ is the standard $\mathscr{L}^2$-norm with respect to some positive measure, is asymptotically efficient (see Carrasco and Florens (2014)). Note that this procedure requires an estimation of a covariance operator and is computationally ambitious. For more information

and some applications, see also (Carrasco et al., 2007; Carrasco & Florens, 2002a; Carrasco & Florens, 2002b).

Example 2.5 (Gamma distribution, continuation of Example 2.4). Let us now introduce a two-step Stein estimator for the gamma distribution. We first recall that the CDF of the gamma distribution is given by $P_\theta(x) = \gamma(\alpha, \beta x)/\Gamma(\alpha)$, where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function. With this formula at hand, we can calculate the optimal functions, which are given by

$$f_\theta^{(1)}(x) = e^{\beta x}\left(\frac{\gamma(\alpha, \beta x)}{(\beta x)^\alpha}\big(\log(\beta x) - \psi(\alpha)\big) - \frac{1}{\alpha^2}\,{}_2F_2(\alpha, \alpha; 1+\alpha, 1+\alpha; -\beta x)\right), \quad f_\theta^{(2)}(x) = \frac{1}{\beta},$$

where ${}_2F_2$ denotes the generalized hypergeometric function. Taking $\hat{\theta}_n^{LOG}$ as a first-step estimate results in a two-step estimator, which we denote by $\hat{\theta}_n^{\mathrm{ST}}$. This estimator takes a rather complicated form, but can be expressed in closed-form in terms of the generalized hypergeometric function. However, in practice, it is computationally more efficient to estimate the derivative in Equation (12) numerically. In the Supporting Information, we show that the assumptions of Theorems 2.5 and 2.6 hold, which implies (strong) consistency and asymptotic efficiency of $\hat{\theta}_n^{\mathrm{ST}}$. Simulation results are reported in the Supporting Information, which show that the Stein estimator $\hat{\theta}_n^{\mathrm{ST}}$ has a marginally improved performance in terms of lower bias and mean square error over the (non-explicit) MLE in small sample sizes across a range of parameter values.

As the CDF of the gamma distribution is expressed in terms of special functions, the optimal functions take a rather complicated form. For distributions with simpler CDFs, simpler optimal functions can be obtained; see, for example, the Cauchy distribution in Section 3.2.

In the remainder of this section, we study the sequence of Stein estimators which is obtained as follows: Choose some $\theta^0 \in \Theta$ as a value for θ in f_θ and solve for the two-step Stein estimator $\hat{\theta}_n^\star$. Take then the obtained estimate as a new value for θ in f_θ to update the Stein estimator $\hat{\theta}_n^\star$. Formally speaking, we consider the sequence of Stein estimators $\hat{\theta}_n^{(m)}$ defined by

$$0 = \overline{\mathcal{A}_{\hat{\theta}_n^{(m+1)}} f_{\hat{\theta}_n^{(m)}}(X)}, \tag{14}$$

where $\hat{\theta}_n^{(0)} = \theta^0 \in \Theta$ is the starting value of the iterating process. Moreover, let $\Theta_0 \subset \Theta$ be compact and convex with $\theta_0, \theta^0 \in \Theta_0$. We briefly discuss the existence of such a sequence. It is clear from Theorem 2.5 that, for fixed $m \in \mathbb{N}$, the probability that $\hat{\theta}_n^{(m)}$ exists converges to 1. However, this does not guarantee the existence of the sequence. Therefore, when we study the asymptotic behavior of the sequence $\hat{\theta}_n^{(m)}$, $m \in \mathbb{N}$, we have to assume that such a sequence of solutions of (14) exist. Before stating the theorem, we introduce a new set of assumptions.

Assumption 2.3.

(a) The MLE exists and is unique with probability converging to 1. Moreover, we assume that the MLE is consistent (in the sense of Theorem 2.5) and that if the MLE exists, it is characterized by (10).

(b) Let $X \sim \mathbb{P}_{\theta_0}$ and $\theta_1, \theta_2 \in \Theta_0$. Then $f_{\theta_2} \in \mathscr{F}$, and $\mathbb{E}[\mathcal{A}_{\theta_1} f_{\theta_2}(X)] = 0$ if and only if $\theta_1 = \theta_0$.

(c) For $q \geq p$, we can write $\mathcal{A}_{\theta_1} f_{\theta_2}(x) = M_{\theta_2}(x) g(\theta_1)$ for some measurable $p \times q$ matrix M_{θ_2} and $g = (g_1, \ldots, g_q)^\top : \Theta \to \mathbb{R}^q$ continuously differentiable for all $\theta_1, \theta_2 \in \Theta_0$, $x \in (a, b)$. Moreover, we assume that $\mathbb{E}[M_\theta(X)] \frac{\partial}{\partial \theta} g(\theta)$ (where $X \sim \mathbb{P}_{\theta_0}$) is invertible for all $\theta \in \Theta_0$ and that the function $\theta \mapsto \mathrm{vec}(M_\theta(x))$ is continuously differentiable on Θ_0 for all $x \in (a, b)$.

(d) For $X \sim \mathbb{P}_{\theta_0}$, there exist two functions F_1, F_2 on (a, b) with $\mathbb{E}[F_i(X)] < \infty$, $i = 1, 2$, such that $\|M_\theta(x)\| \leq F_1(x)$ and $\|\frac{\partial}{\partial \theta} \mathrm{vec}(M_\theta(x))\| \leq F_2(x)$ for all $\theta \in \Theta_0$, $x \in (a, b)$.

Assumptions 2.3(b)–(d) introduced above are mostly equivalent to Assumptions 2.2(b)–(d), although we have a slight modification in (c). Here, we require the matrix $\mathbb{E}[M_\theta(X)] \frac{\partial}{\partial \theta} g(\theta)$, $X \sim \mathbb{P}_{\theta_0}$, to be invertible for all $\theta \in \Theta_0$, in contrast to 2.2(c) in which this needs to be the case only for $\theta = \theta_0$, which can be difficult to verify, especially if f_θ is complicated.

Theorem 2.7. *Suppose that Assumptions 2.3(a) to (d) hold. Then, for each sequence $\hat{\theta}_n^{(m)}$, $m \in \mathbb{N}$, satisfying (14), there exists a sequence of sets $A_n \subset \Omega$, $n \in \mathbb{N}$, with $\mathbb{P}(A_n) \to 1$ as $n \to \infty$ such that for each n we have that on A_n, $\hat{\theta}_n^{(m)} \to \hat{\theta}_n^{\mathrm{ML}}$ as $m \to \infty$.*

3 | APPLICATIONS

In this section, we apply Stein's method of moments to three challenging estimation problems for univariate distributions that have received interest in the literature. Here we establish small sample performance of our asymptotically efficient estimators obtained in Section 2.4, and, by choosing suitable test functions, we propose alternatives to moment estimation that are as simple and improve significantly in terms of asymptotic variance. We conclude the section with an application to a less challenging setting for which censoring can break down the performance of ML and other MOM estimates, thereby demonstrating the usefulness of our approach even for regular models. For all examples, we suppose that $\{X_n, n \in \mathbb{Z}\}$ is i.i.d. However, we stress that SMOM can be applied to dependent data, and we give such an application in Example C.7 of the Supporting Information.

In each example, we compare to the MLE and, where appropriate, the classical moment estimators, as well as more specialist estimators that have been found to perform well for the particular distribution under consideration. It would seem that the minimum Stein discrepancy estimators developed in Barp et al. (2019) would be natural competitors as the discrepancy is based on the density approach Stein identity. However, we have excluded them from our simulation studies, as we found that they are outperformed for almost all parameter values in terms of bias and MSE, involve more computational effort and for certain distribution require a numerical procedure even when our Stein estimators are completely explicit; a more detailed justification is given in Section A of the Supporting Information.

Further examples for the beta, Student's t, Lomax, Nakagami, one-sided truncated inverse-gamma distribution, and generalized logistic distributions, as well as a non-i.i.d. example, are given in the Supporting Information. Some of these estimators also have competitive performance.

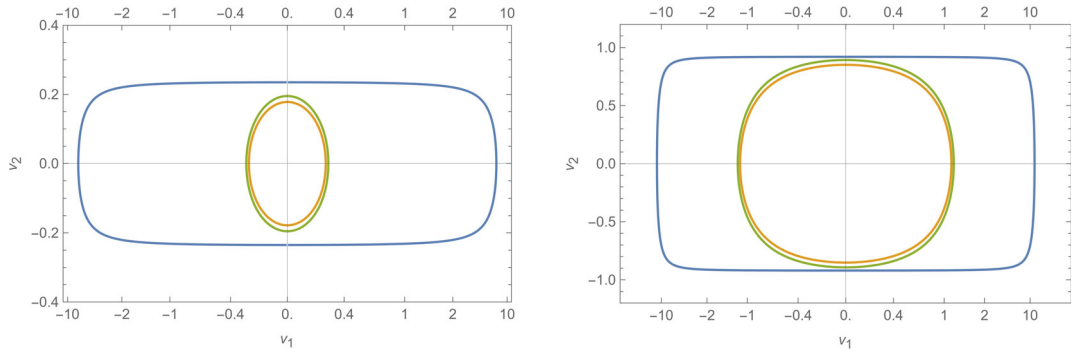

FIGURE 1 Asymptotic confidence regions for the estimators of the $TN(\mu, \sigma^2)$ *distribution for* $q = 0.95$, $a = 0$, *and* $b = 1$ *in the directions of the eigenvectors* v_1 *and* v_2. Plotted are the MLE ●, the moment estimator ●, and the Stein estimator ●. The x-axis scale is transformed via $x \mapsto \arctan x$. (a) $\mu = 0.5$, $\sigma = 0.2$, (b) $\mu = 0.5$, $\sigma = 0.3$.

3.1 | Truncated normal distribution

The density of the two-sided truncated Gaussian distribution on (a, b) with $a, b \in \mathbb{R}$, denoted by $TN(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, is given by $p_\theta(x) = C_\theta^{-1}\phi((x-\mu)/\sigma)$, where $C_\theta = \sigma[\Phi(b-\mu)/\sigma) - \Phi(a-\mu)/\sigma)]$ and ϕ and Φ are the standard Gaussian PDF and CDF, respectively. With $\tau_\theta(x) = \sigma^2$ we obtain the same Stein operator as in Example 2.1. Note that the function class $\mathcal{F}_\theta$ differs from the one in the untruncated case. As we have the same Stein operator as in Example 2.1, we obtain for two test functions f_1, f_2 the same expressions for the Stein estimators as in the untrucated case, as given by (8). Note that the normalizing constant drops out, and therefore, completely explicit and easily computable estimators are retrieved. A natural choice seems to be the polynomials

$$f_1(x) = -x^2 + (a+b)x - ab, \quad f_2(x) = x^3 - \frac{3}{2}(a+b)x^2 + \frac{1}{2}(a^2 + 4ab + b^2)x - \frac{1}{2}(a^2 b + ab^2).$$

We denote the Stein estimator based on the latter test functions by $\hat{\theta}_n^{\mathrm{ST}} = (\hat{\mu}_n^{\mathrm{ST}}, \hat{\sigma}_n^{\mathrm{ST}})$.

The first and second moments of the truncated normal distribution take a rather complicated form involving the functions ϕ and Φ (see the Supporting Information), and consequently the classical moment estimators, which we denote by $\hat{\mu}_n^{\mathrm{MO}}$ and $\hat{\sigma}_n^{\mathrm{MO}}$, must be obtained numerically. The MLE $\hat{\theta}_n^{\mathrm{ML}} = (\hat{\mu}_n^{\mathrm{ML}}, \hat{\sigma}_n^{\mathrm{ML}})$ is also not explicit, and, as for the classical moment estimator, the numerical calculation can be tedious.

A q-confidence region of the corresponding asymptotic normal distribution for the above estimation techniques is reported in Figure 1 for two parameter constellations. The ellipses are plotted with respect to the two eigenvectors v_1 and v_2 of the covariance matrix and are therefore parallel to the x- and y-axes, respectively. One can see that the performance of the proposed Stein estimator essentially coincides with that of the MLE, indicating a behavior close to efficiency. The moment estimator performs poorly and was hence excluded from our finite sample simulation study.

Hegde and Dahiya (1989) showed that the MLE exists if and only if $\overline{Y}^2 < \overline{Y^2} < 1 - 2\overline{Y}/x^\star$ with $\coth(x^\star) - 1/x^\star = \overline{Y}$, where $Y_i = 2(X_i - a)/(b-a)$. The conditions for the existence of the moment estimator seem to be difficult to work out. It is a known issue for any explicit estimator that it is possible for the estimate to lie outside of the parameter space if the latter is restricted to

TABLE 1 Simulation results for the $TN(\mu, \sigma)$ distribution with $a = 0, b = 1$ for $n = 20$ and 10,000 repetitions.

		Bias		MSE		NE	
θ_0		$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$
(0.5, 0.05)	μ	−4.56e−5	−4.64e−5	1.22e−4	1.22e−4	0	0
	σ	−1.27e−4	−1.21e−4	6.07e−7	6.14e−7		
(0.5, 0.1)	μ	2.58e−4	2.97e−4	4.95e−4	5e−4	0	0
	σ	−4.84e−4	−3.59e−4	9.94e−6	1.05e−5		
(0.5, 0.2)	μ	−4.48e−5	−4.76e−5	2.87e−3	3.46e−3	0	0
	σ	1.37e−3	1.88e−3	1.45e−3	1.44e−3		
(0.5, 0.3)	μ	−3.18e−3	−1.7e−3	0.038	0.044	3	3
	σ	0.044	0.044	0.061	0.06		
(0.6, 0.05)	μ	−1.93e−4	−1.86e−4	1.25e−4	1.25e−4	0	0
	σ	−1.31e−4	−1.24e−4	6.09e−7	6.19e−7		
(0.6, 0.1)	μ	−2.65e−4	−1.6e−4	5e−4	5.08e−4	0	0
	σ	−5.49e−4	−4.24e−4	9.76e−6	1.06e−5		
(0.6, 0.2)	μ	4.64e−3	4.31e−3	3.29e−3	3.43e−3	0	0
	σ	1.67e−3	2.28e−3	6.87e−4	7.64e−4		
(0.7, 0.05)	μ	1.12e−4	1.3e−4	1.26e−4	1.27e−4	0	0
	σ	−1.39e−4	−1.3e−4	6.1e−7	6.34e−7		
(0.7, 0.1)	μ	2.44e−4	1.37e−4	5.28e−4	5.42e−4	0	0
	σ	−4.01e−4	−3.68e−4	1.11e−5	1.23e−5		
(0.7, 0.2)	μ	0.015	0.016	0.013	0.015	0	0
	σ	3.92e−3	4.83e−3	3.06e−3	4.16e−3		

a certain subset of Euclidean space. This problem also applies to the Stein estimator. We added a column *NE* to the tables to report the estimated relative frequency of cases in which the estimator does not exist (the relative frequency is given as a number between 0 and 100). These estimates are based on the same Monte Carlo samples as the estimates for bias and MSE. However, for the considered parameter constellations, the existence of the estimator seems to be hardly an issue. Nevertheless, we noticed that in cases where parameter estimation for the $TN(\mu, \sigma^2)$-distribution becomes in general more difficult (e.g., when μ lies outside of the truncation domain and σ^2 is large), the number of Monte Carlo samples for which the MLE and the Stein estimator does not exist grows rapidly. As can be seen in Tables 1 and 2, the Stein estimator performs well compared to the MLE throughout all parameter constellations and both sample sizes (the best performances are highlighted in orange). However, we remark that we used the truncated mean to estimate the bias and MSE, as we considered large estimates as outliers. We set the truncation threshold equal to 5 for the bias and equal to 25 for the MSE. Estimates not taken into account for bias or MSE were then considered as non-existent.

TABLE 2 Simulation results for the $TN(\mu, \sigma)$ distribution with $a = 0, b = 1$ for $n = 50$ and 10,000 repetitions.

		Bias		MSE		NE	
θ_0		$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST}$
(0.5, 0.05)	μ	3.8e−6	4.67e−6	4.87e−5	4.87e−5	0	0
	σ	−5.43e−5	−5.14e−5	2.45e−7	2.46e−7		
(0.5, 0.1)	μ	−8.44e−5	−1.32e−4	2e−4	2.02e−4	0	0
	σ	−1.91e−4	−1.45e−4	3.97e−6	4.14e−6		
(0.5, 0.2)	μ	−1.1e−4	−5.72e−5	9.02e−4	9.53e−4	0	0
	σ	2.53e−4	4.83e−4	1.19e−4	1.3e−4		
(0.5, 0.3)	μ	1.03e−4	−4.65e−5	9.33e−3	9.97e−3	0	0
	σ	0.017	0.017	0.013	0.014		
(0.6, 0.05)	μ	6.62e−6	6.46e−6	4.84e−5	4.85e−5	0	0
	σ	−4.97e−5	−4.69e−5	2.46e−7	2.49e−7		
(0.6, 0.1)	μ	3.19e−5	4.72e−5	2.02e−4	2.05e−4	0	0
	σ	−1.9e−4	−1.43e−4	3.94e−6	4.18e−6		
(0.6, 0.2)	μ	1.71e−3	1.41e−3	1.02e−3	1.08e−3	0	0
	σ	2.48e−4	4.89e−4	1.25e−4	1.37e−4		
(0.7, 0.05)	μ	2.06e−5	2.77e−5	4.9e−5	4.9e−5	0	0
	σ	−4.46e−5	−4.1e−5	2.5e−7	2.59e−7		
(0.7, 0.1)	μ	2.85e−4	2.58e−4	2.02e−4	2.09e−4	0	0
	σ	−1.44e−4	−1.27e−4	4.48e−6	4.96e−6		
(0.7, 0.2)	μ	4.21e−3	4.13e−3	1.72e−3	1.84e−3	0	0
	σ	6.31e−4	9.94e−4	1.78e−4	2e−4		

3.2 | Cauchy distribution

For $\theta = (\mu, \gamma) \in \mathbb{R} \times (0, \infty)$, the density of the Cauchy distribution is given by $p_\theta(x) = (\pi\gamma)^{-1}(1 + ((x-\mu)/\gamma)^2)^{-1}$, $x \in \mathbb{R}$. We fix $\tau_\theta = (x-\mu)^2 + \gamma^2$, and obtain $\mathcal{A}_\theta f(x) = \left((x-\mu)^2 + \gamma^2\right) f'(x)$ (see also Schoutens (2001)). For test functions f_1, f_2 we obtain the estimators

$$\hat{\mu}_n = \frac{\overline{f_2'(X)}\ \overline{X^2 f_1'(X)} - \overline{f_1'(X)}\ \overline{X^2 f_2'(X)}}{2\left[\overline{f_2'(X)}\ \overline{X f_1'(X)} - \overline{f_1'(X)}\ \overline{X f_2'(X)}\right]}, \quad \hat{\gamma}_n^2 = \frac{\overline{X^2 f_2'(X)}\ \overline{X f_1'(X)} - \overline{X^2 f_1'(X)}\ \overline{X f_2'(X)}}{\overline{f_1'(X)}\ \overline{X f_2'(X)} - \overline{f_2'(X)}\ \overline{X f_1'(X)}} - \hat{\mu}_n^2.$$

The CDF is $P_\theta(x) = \pi^{-1} \arctan((x-\mu)/\gamma) + 1/2$, and we thus obtain simple optimal functions:

$$f_\theta^{(1)}(x) = -\frac{1}{\gamma^2 + (x-\mu)^2} \quad \text{and} \quad f_\theta^{(2)}(x) = \frac{\mu - x}{\gamma\left(\gamma^2 + (x-\mu)^2\right)}. \tag{15}$$

With a suitable first step estimate, we have an efficient estimator which is considerably simpler to compute than the MLE, which involves solving polynomial equations of degree $2n-1$ given by

$$\sum_{i=1}^{n} \frac{2(X_i-\mu)}{\gamma^2+(X_i-\mu)^2} = 0 \quad \text{and} \quad \frac{n}{\gamma} - \sum_{i=1}^{n} \frac{2\gamma}{\gamma^2+(X_i-\mu)^2} = 0.$$

In (Copas, 1975; Gabrielsen, 1982) it is shown that, in the case where both parameters μ and γ are unknown, the likelihood function is unimodal under some regularity assumptions. Clearly, moment estimation is not tractable due to the non-existence of all moments.

Interestingly, parameter estimation for the Cauchy distribution can be difficult in the case where γ is known and one is left with estimation of the location parameter μ. We therefore now focus on the case that γ is known. The parameter space thus reduces to $\Theta = \mathbb{R}$ with $\theta = \mu$. This estimation problem has received great attention in the literature; see Zhang (2010) for an overview of available estimation techniques. The MLE of μ with known γ is often cited as an example of computational failure (Bai and Fu, 1987; Zhang, 2010, summarize the challenges) although Bai and Fu (1987) show that the MLE remains the asymptotically optimal estimator in the Bahadur sense. One reason for this is a multimodal likelihood function (in fact, the number of local maxima is asymptotically Poisson distributed with mean $1/\pi$; see Reeds (1985)). However, closed-form expressions for the MLE exist for sample sizes 3 and 4; see Ferguson (1978). Due to the difficulties concerning the MLE, other methods have been developed. In our simulation study we consider the L-estimator methods of Rothenberg et al. (1964), Bloch (1966), Chernoff et al. (1967) and Zhang (2010), which we denote by $\hat{\mu}_n^{\mathrm{L1}}$, $\hat{\mu}_n^{\mathrm{L2}}$, $\hat{\mu}_n^{\mathrm{L3}}$ and $\hat{\mu}_n^{\mathrm{L4}}$, respectively. We also consider the Pitman estimator of Freue (2007), which we denote by $\hat{\mu}_n^{\mathrm{PI}}$. The explicit forms of these estimators are given in the Supporting Information. Zhang (2010) modified the estimator $\hat{\mu}_n^{\mathrm{L3}}$ to also achieve high efficiency for finite sample sizes, and so we do not include the estimator $\hat{\mu}_n^{\mathrm{L3}}$ in our simulation study.

Let us now describe a procedure based on the Stein operator. Note that if we choose one test function (since we only have to estimate μ), the corresponding equation is quadratic and has, in general, two solutions. This is why we choose two test functions and consider the estimator $\hat{\mu}_n$ with test functions $f_\theta^{(1)}(x)$ and $f_\theta^{(2)}(x)$ as defined in Equation (15), whereby γ is now considered known. We take $\hat{\mu}_n^{\mathrm{L4}}$ as a first-step estimate, and denote the resulting estimator by $\hat{\mu}_n^{\mathrm{ST1}}$. Note that $\hat{\mu}_n^{\mathrm{ST1}}$ is not translation-invariant, and we therefore consider different values of μ in our simulation.

This slight modification of the estimation procedure still results in an asymptotically efficient estimator. We apply Theorem 2.5 in the setting where both parameters are estimated (where we take $\tilde{\gamma}_n = \gamma_0$ as the first-step estimator for γ). Then the asymptotic variance of $\hat{\mu}_n^{\mathrm{ST1}}$ is the top-left element of the inverse Fisher information matrix in the case where γ is unknown. The latter is given by the diagonal matrix $I_{\mathrm{ML}}^{-1}(\mu,\gamma) = \mathrm{diag}(2\gamma^2, 2\gamma^2)$, and we conclude that the asymptotic variance for both estimators $\hat{\mu}_n^{\mathrm{ST1}}$ and the (one-dimensional) MLE equals $2\gamma^2$.

However, when performing the simulations we noticed a very large variance for $\hat{\mu}_n^{\mathrm{ST1}}$ for small sample sizes, which is consistent with the trade-off between small sample size and asymptotic efficiency noticed by Zhang (2010) for $\hat{\mu}_n^{\mathrm{L3}}$ and $\hat{\mu}_n^{\mathrm{L4}}$. This is why we propose a modified version of $\hat{\mu}_n^{\mathrm{ST1}}$, denoted by $\hat{\mu}_n^{\mathrm{ST2}}$. In a similar manner to the estimator $\hat{\mu}_n^{\mathrm{L1}}$, we cut off the bottom and top p-quantile of the sample at hand and calculate the sample means in $\hat{\mu}_n$ and $\hat{\gamma}_n^2$ by the means of the remaining observations. Pursuant to $\hat{\mu}_n^{\mathrm{L1}}$, we choose $p = 0.38$ and disregard the first and last $\lfloor np \rfloor$ observations of the sorted sample. Simulation results can be found in Tables 3 and 4. For the sample size $n = 20$, the Pitman estimator $\hat{\mu}_n^{\mathrm{PI}}$ seems to be globally the best, with the modified L-estimator $\hat{\mu}_n^{\mathrm{L4}}$

TABLE 3 Simulation results for the $C(\mu, \gamma)$ distribution for $n = 20$ and 10,000 repetitions.

		Bias					MSE				
θ_0		$\hat{\mu}_n^{L1}$	$\hat{\mu}_n^{L2}$	$\hat{\mu}_n^{L4}$	$\hat{\mu}_n^{PI}$	$\hat{\mu}_n^{ST2}$	$\hat{\mu}_n^{L1}$	$\hat{\mu}_n^{L2}$	$\hat{\mu}_n^{L4}$	$\hat{\mu}_n^{PI}$	$\hat{\mu}_n^{ST2}$
$(-5, 1)$	μ	−0.725	0.131	−2.54e−3	1.72e−4	0.085	0.721	0.944	0.136	0.116	0.158
$(-4, 1.5)$	μ	−0.506	0.193	−4.22e−3	−3.13e−5	0.122	0.695	1.04	0.307	0.261	0.351
$(-2, 2)$	μ	−0.099	0.261	8.84e−3	3.88e−3	0.182	0.782	2.26	0.537	0.46	0.627
$(0, 1)$	μ	0.118	0.117	5.1e−3	3.95e−3	0.091	0.206	0.296	0.134	0.113	0.161
$(0, 3)$	μ	0.343	0.394	5.14e−3	3.29e−3	0.273	1.87	4.81	1.22	1.02	1.42
$(2, 0.1)$	μ	0.345	0.013	3.05e−4	2.73e−4	9.01e−3	0.121	5.89e−3	1.36e−3	1.15e−3	1.59e−3
$(2, 0.5)$	μ	0.388	0.065	−1.29e−3	−1.22e−3	0.043	0.2	0.24	0.034	0.028	0.039
$(4, 0.8)$	μ	0.751	0.092	−4.79e−3	−4.02e−3	0.066	0.69	0.325	0.087	0.074	0.101
$(6, 2.3)$	μ	1.26	0.29	7.79e−3	3.64e−3	0.217	2.59	2.02	0.693	0.59	2.92
$(10, 0.2)$	μ	1.69	0.026	5.34e−4	3.58e−4	0.018	2.86	0.03	5.62e−3	4.49e−3	6.78e−3

close behind. The Stein estimator $\hat{\mu}_n^{ST2}$ delivers good results as well, outperforming $\hat{\mu}_n^{L1}$ and $\hat{\mu}_n^{L2}$ for most parameter constellations. For the sample size $n = 50$, $\hat{\mu}_n^{L4}$, $\hat{\mu}_n^{PI}$ and $\hat{\mu}_n^{ST1}$ show the best performance regarding the bias and the Stein estimator $\hat{\mu}_n^{ST1}$ has the lowest MSE for most parameter constellations. Further simulation results for the sample size $n = 100$ are given in the Supporting Information. For this sample size, the Stein estimator $\hat{\mu}_n^{ST1}$ has the lowest MSE for all parameter constellations, and the Pitman estimator performs very poorly. Indeed, our simulations suggest that $\hat{\mu}^{PI}$ is not a consistent estimator.

3.3 | Exponential polynomial models

For $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^{p-1} \times (-\infty, 0)$, the density of an exponential polynomial model is given by $p_\theta(x) = C_\theta^{-1} \exp(\theta_1 x + \cdots + \theta_p x^p)$, $x > 0$, where $C_\theta = \int_0^\infty \exp(\theta_1 x + \cdots + \theta_p x^p)\, dx$ is the normalizing constant which cannot be calculated analytically. We choose $\tau_\theta(x) = 1$ and obtain the Stein operator

$$\mathcal{A}_\theta f(x) = (\theta_1 + 2\theta_2 x + \cdots + p\theta_p x^{p-1}) f(x) + f'(x).$$

Here, we need p test functions $f_1, \ldots, f_p$ and the Stein estimator is then given by

$$\hat{\theta}_n = A^{-1} b,$$

where A is a $p \times p$ matrix with (i, j)-th entry $j\overline{X^{j-1} f_i(X)}$ and $b = (-\overline{f_1'(X)}, \ldots, -\overline{f_p'(X)})^\top$. We propose the test functions $f_1(x) = x^i$, for $i = 1, \ldots, p$, and denote the corresponding estimator by $\hat{\theta}_n^{ST1}$. We also consider $f_i(x) = x^i e^{-ix}$, for $i = 0, \ldots, p-1$, and call the respective Stein estimator $\hat{\theta}_n^{ST2}$. Further, we study the two-step Stein estimator, which we denote by $\hat{\theta}_n^{ST3}$, whereby we take $\hat{\theta}_n^{ST2}$ as a first-step estimate. This estimator is consistent and asymptotically efficient.

Let us walk through the estimation methods in the literature. Hayakawa and Takemura (2016) and Nakayama et al. (2011) used the holomorphic gradient method to compute the MLE. For

TABLE 4 Simulation results for the $C(\mu, \gamma)$ distribution for $n = 50$ and 10,000 repetitions.

		Bias					MSE				
θ_0		$\hat{\mu}_n^{L1}$	$\hat{\mu}_n^{L2}$	$\hat{\mu}_n^{L4}$	$\hat{\mu}_n^{PI}$	$\hat{\mu}_n^{ST1}$	$\hat{\mu}_n^{L1}$	$\hat{\mu}_n^{L2}$	$\hat{\mu}_n^{L4}$	$\hat{\mu}_n^{PI}$	$\hat{\mu}_n^{ST1}$
$(-5, 1)$	μ	−0.378	0.012	1.63e−3	3.7e−3	1.71e−3	0.201	0.048	0.05	0.082	**0.044**
$(-4, 1.5)$	μ	−0.276	0.015	1.38e−3	1.5e−3	1.17e−3	0.206	0.105	0.11	**0.099**	**0.099**
$(-2, 2)$	μ	−0.092	0.025	−7.46e−4	7.45e−3	1.33e−3	0.248	0.195	0.202	0.554	**0.182**
$(0, 1)$	μ	0.034	7.29e−3	−2.69e−3	3.68e−3	−3.39e−3	0.058	0.047	0.048	0.517	**0.044**
$(0, 3)$	μ	0.094	0.024	−0.016	−9.89e−3	−0.01	0.527	0.42	0.437	**0.384**	0.39
$(2, 0.1)$	μ	0.171	1.29e−3	1.9e−4	4.96e−4	1.93e−4	0.03	4.79e−4	4.99e−4	1.08e−3	4.46e−4
$(2, 0.5)$	μ	0.186	6.37e−3	2.87e−4	1.14e−3	1.12e−3	0.049	0.012	0.012	0.019	**0.011**
$(4, 0.8)$	μ	0.364	8.99e−3	4.44e−4	−2.36e−3	−2.75e−4	0.169	0.03	0.031	0.037	**0.028**
$(6, 2.3)$	μ	0.582	0.024	−3.22e−3	−9.5e−4	−2.31e−3	0.639	0.243	0.253	0.24	**0.226**
$(10, 0.2)$	μ	0.84	1.84e−3	−4.34e−4	−2.94e−3	−2.73e−4	0.708	1.83e−3	1.87e−3	0.154	1.7e−3

exponential polynomial models, the MLE coincides with the moment estimator. Gutmann and Hyvärinen (2012) (a refined version of one from Gutmann and Hyvärinen (2010)) proposed the noise-contrastive estimator, which we denote by $\hat{\theta}_n^{\mathrm{NC}}$. We consider the score matching approach from Hyvärinen (2007) (a refined version of Hyvärinen and Dayan (2005)) and denote the score matching estimator by $\hat{\theta}_n^{\mathrm{SM}}$. The estimators $\hat{\theta}_n^{\mathrm{NC}}$ and $\hat{\theta}_n^{\mathrm{SM}}$ are computed via numerical optimization, and the procedure for implementing them is given in the Supporting Information. It is also natural to consider the minimum $\mathscr{L}^q$-estimator obtained from Betsch et al. (2021), which is also motivated by a Stein characterization; more precisely, by an expectation-based representation of the CDF. The minimum distance estimator is only explicit for a parameter space of dimension less than or equal to 2, and we thus exclude this estimator from our simulation study, since the numerical calculation turns out to be too heavy for a parameter space dimension of 3 or higher.

In our simulation study, for the MLE, C_θ is calculated through numerical integration, and optimizing the log-likelihood function is performed with the Nelder-Mead algorithm. The vector $(-1,\dots,-1)^\top \in \mathbb{R}^p$ is used as an initial guess for the optimization procedure. This implementation seems, at least for the parameter constellations we consider, to be computationally manageable and numerically stable. For the noise-contrastive estimator $\hat{\theta}_n^{\mathrm{NC}}$ and the score matching approach $\hat{\theta}_n^{\mathrm{SM}}$, we also used the Nelder-Mead algorithm with initial guess $(-1,\dots,-1)^\top \in \mathbb{R}^p$. For the two-step Stein estimator $\hat{\theta}_n^{\mathrm{ST3}}$, the normalizing constant C_θ needs to be calculated to evaluate the optimal function. This is done through numerical differentiation.

The results are reported in Tables 5 and 6. The column *NE* is interpreted as follows. First, the last element of the parameter vector θ_p has to be negative. Thus, if any estimator returns a positive value for this parameter, we count the estimator as non-existent. Secondly, we restrict the computation time for each estimator to 20 s, meaning that an estimator counts equally as non-existent if it requires more time to be calculated or if the numerical procedure fails completely. Concerning $\hat{\theta}_n^{\mathrm{ST3}}$, we also used the parameter vector $(-1,\dots,-1)^\top \in \mathbb{R}^p$ as a first-step estimate if $\hat{\theta}_n^{\mathrm{ST2}}$ was not available for a Monte Carlo sample. The sample size for this simulation was chosen to be larger than in the other simulation studies, since we are concerned with an estimation problem in which the variance of the estimator typically increases as the dimension of the parameter space grows. This makes it difficult to compare estimators for small sample sizes in the case of parameter dimensions of 3 and 4. Additionally, the Stein estimators $\hat{\theta}_n^{\mathrm{ST1}}$ and $\hat{\theta}_n^{\mathrm{ST2}}$ often return positive values for θ_p, which makes a comparison even more difficult since the number of samples on which the bias and MSE are based is in truth lower than the number of Monte Carlo repetitions. However, for rather small sample sizes of $n = 20$ or $n = 50$, we found that our simulation results are reliable for a parameter space dimension of 2 with similar results as described below, which is why we did not include a separate table for these results. Therefore, we chose the sample size $n = 1000$, where we feel comfortable in drawing conclusions from the study. Overall, we observe a solid performance of the Stein estimators. For example, the explicit Stein estimators $\hat{\theta}_n^{\mathrm{ST1}}$ and $\hat{\theta}_n^{\mathrm{ST2}}$ outperform all other methods for the parameter vector $(-2, 0.1, 3, -2)^\top$ in terms of bias and MSE. The two-step Stein estimator $\hat{\theta}_n^{\mathrm{ST3}}$ together with $\hat{\theta}_n^{\mathrm{ML}}$ and $\hat{\theta}_n^{\mathrm{NC}}$ seem to be globally the best. Moreover, one observes that $\hat{\theta}_n^{\mathrm{ST3}}$ can often improve in terms of bias and MSE with respect to the first-step estimator $\hat{\theta}_n^{\mathrm{ST2}}$ (although there are some exceptions). In the end, we advise to use $\hat{\theta}_n^{\mathrm{ST3}}$, the MLE or the noise-contrastive estimator $\hat{\theta}_n^{\mathrm{NC}}$, while the explicit Stein estimators can serve as a reliable initial guess, if they exist.

TABLE 5 Simulation results regarding the bias for the exponential polynomial models for $n = 1000$ and 10,000 repetitions.

		Bias					
θ_0		$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{NC}$	$\hat{\theta}_n^{SM}$	$\hat{\theta}_n^{ST1}$	$\hat{\theta}_n^{ST2}$	$\hat{\theta}_n^{ST3}$
$(1, -2)$	θ_1	0.02	0.019	0.085	0.03	0.03	0.017
	θ_2	−0.017	−0.017	−0.055	−0.024	−0.024	−0.015
$(-2, -1)$	θ_1	0.025	0.026	0.13	0.035	0.035	0.023
	θ_2	−0.031	−0.032	−0.106	−0.04	−0.04	−0.029
$(1, 2, -3)$	θ_1	−0.025	−0.027	−0.477	−0.122	−0.122	−0.011
	θ_2	0.073	0.079	0.737	0.244	0.244	0.042
	θ_3	−0.051	−0.054	−0.34	−0.134	−0.134	−0.033
$(-3, 5, -1)$	θ_1	2.6	2.59	−4.83	−0.078	−0.078	2.82
	θ_2	−0.92	−0.923	1.62	0.048	0.048	−0.986
	θ_3	0.107	0.107	−0.18	−7.71e−3	−7.71e−3	0.113
$(0.2, -0.8, -2)$	θ_1	−0.048	−0.054	−0.923	−0.208	−0.189	−0.093
	θ_2	0.156	0.176	1.83	0.532	0.491	0.252
	θ_3	−0.127	−0.142	−1.04	−0.359	−0.336	−0.183
$(3, 0.5, -0.5)$	θ_1	−0.018	0.019	−0.332	−0.101	−0.101	0.052
	θ_2	0.026	−5.63e−4	0.205	0.075	0.075	−0.026
	θ_3	−7.56e−3	−1.87e−3	−0.04	−0.016	−0.016	3.63e−3
$(0.1, 2, -3)$	θ_1	−0.036	−0.04	−0.544	−0.136	−0.136	−0.022
	θ_2	0.102	0.112	0.909	0.296	0.295	0.067
	θ_3	−0.07	−0.077	−0.445	−0.172	−0.171	−0.049
$(3, 0, -4)$	θ_1	−0.018	−0.021	−0.745	−0.164	−0.16	−5.05e−3
	θ_2	0.088	0.093	1.38	0.391	0.382	0.049
	θ_3	−0.082	−0.085	−0.763	−0.258	−0.253	−0.054
$(1, 2, 0.5, -2)$	θ_1	−0.881	−0.94	3.42	1.16	0.888	0.58
	θ_2	2.19	2.34	−7.61	−3.14	−2.43	−1.65
	θ_3	−2.01	−2.16	6.73	3.15	2.48	1.71
	θ_4	0.614	0.66	−2.06	−1.06	−0.842	−0.591
$(-2, 0.1, 3, -2)$	θ_1	1.2	1.57	1.9	0.382	0.344	0.139
	θ_2	−3.21	−4.2	−4.51	−1.19	−1.08	−0.43
	θ_3	3.04	3.97	4.01	1.26	1.14	0.458
	θ_4	−0.939	−1.22	−1.19	−0.421	−0.386	−0.159

3.4 | Nakagami distribution for censored data: A real data application

The Nakagami distribution $NG(m, O)$, also known as the m-distribution Nakagami (1960), is a continuous probability distribution on the positive real numbers. Its probability density function is given by

$$p_\theta(x) = \frac{2m^m}{\Gamma(m)O^m} x^{2m-1} \exp\left(-\frac{m}{O}x^2\right), \quad x > 0$$

TABLE 6 Simulation results regarding the MSE and existence for the exponential polynomial models for $n = 1000$ and 10,000 repetitions.

		MSE						NE					
θ_0		$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{NC}$	$\hat{\theta}_n^{SM}$	$\hat{\theta}_n^{ST1}$	$\hat{\theta}_n^{ST2}$	$\hat{\theta}_n^{ST3}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{NC}$	$\hat{\theta}_n^{SM}$	$\hat{\theta}_n^{ST1}$	$\hat{\theta}_n^{ST2}$	$\hat{\theta}_n^{ST3}$
$(1, -2)$	θ_1	0.073	0.079	0.28	0.09	0.09	0.073	0	0	0	0	0	0
	θ_2	0.039	0.042	0.109	0.047	0.047	0.039						
$(-2, -1)$	θ_1	0.084	0.092	0.289	0.09	0.09	0.084	0	0	0	0	0	0
	θ_2	0.069	0.074	0.171	0.074	0.074	0.068						
$(1, 2, -3)$	θ_1	0.687	0.743	4.83	1.15	1.15	0.687	0	0	0	0	0	0
	θ_2	2.08	2.26	10.7	3.4	3.4	2.08						
	θ_3	0.533	0.576	2.12	0.834	0.834	0.532						
$(-3, 5, -1)$	θ_1	6.74	6.73	44.8	39.6	39.6	29.4	0	0	0	0	0	0
	θ_2	0.873	0.882	5.29	4.99	4.99	3.7						
	θ_3	0.013	0.013	0.069	0.068	0.068	0.051						
$(0.2, -0.8, -2)$	θ_1	0.675	0.725	3.8	0.839	0.871	0.62	0	0	9	3	3	1
	θ_2	3.46	3.72	14.8	4.43	4.57	3.16						
	θ_3	1.42	1.51	4.85	1.82	1.86	1.31						
$(3, 0.5, -0.5)$	θ_1	0.873	0.769	3.82	1.38	1.38	0.67	0	0	0	0	0	0
	θ_2	0.402	0.345	1.35	0.565	0.565	0.292						
	θ_3	0.018	0.015	0.047	0.023	0.023	0.013						
$(0.1, 2, -3)$	θ_1	0.636	0.681	4.23	0.967	0.968	0.636	0	0	1	0	0	0
	θ_2	2.31	2.48	11.1	3.48	3.49	2.31						
	θ_3	0.684	0.73	2.53	0.996	0.997	0.681						
$(3, 0, -4)$	θ_1	1.08	1.17	6.92	1.77	1.78	1.08	0	0	2	0	0	0
	θ_2	4.75	5.18	22.4	7.55	7.6	4.71						
	θ_3	1.78	1.94	6.59	2.71	2.73	1.77						
$(1, 2, 0.5, -2)$	θ_1	1.34	1.27	33.2	6.71	7.23	3.31	0	0	16	13	9	1
	θ_2	9.3	8.96	161	45.2	48.7	21.6						
	θ_3	10.2	10	128	44.2	47.2	21.6						
	θ_4	1.29	1.28	12.3	4.93	5.22	2.51						
$(-2, 0.1, 3, -2)$	θ_1	2.61	3.2	13.8	1.94	2.02	2.54	0	0	9	2	1	0
	θ_2	19.9	24.2	78	16.4	17.1	21.4						
	θ_3	18.9	22.9	62	17	17.7	22						
	θ_4	1.9	2.31	5.54	1.83	1.9	2.34						

where the parameter $\theta = (m, O)$ has strictly positive components: m (the shape parameter) and O (a scale parameter). This distribution, which is the distribution of the square root of a gamma variable, is designed to model phenomena characterized by fading and variability and its applications span various fields, including wireless communications, hydrology, mining, medical imaging (see, e.g., Bağci, 2024; Kolar et al., 2004; Kumar et al., 2024; Miyoshi and Shirai, 2015; Reyes et al., 2020; Tegos et al., 2022, among many others).

Although this is a regular family with straightforward theoretical properties (and thus the MLE is the best all-round estimator), the problem of estimating the parameters of the Nakagami distribution has attracted much attention because of the importance of this distribution for modeling purposes. The MLE of the scale O is

$$\hat{O}_n^{\mathrm{ML}} = \overline{X^2},$$

and this cannot be improved upon. The MLE of the shape m requires solving the likelihood equations

$$\log \hat{m}_n^{\mathrm{ML}} - \psi(\hat{m}_n^{\mathrm{ML}}) - \log \hat{O}_n^{\mathrm{ML}} + 2\overline{\log X} = 0$$

with $\psi(\cdot)$ the digamma function. See, for example, Kolar et al. (2004) for a comparative study of various algorithms computing roots of digamma functions numerically; see Schwartz et al. (2013) for a bias-corrected version. Starting points for the optimization are given by the MOM estimators; the MOM estimate for O is the same as above, but for m can only be computed through a numerical solver and has high asymptotic variance. In Artyushenko and Volovach (2019) the modified MOM estimator

$$\hat{m}_n^{\mathrm{MO2}} = \frac{(\overline{X^2})^2}{\overline{X^4} - (\overline{X^2})^2}, \qquad \hat{O}_n^{\mathrm{MO2}} = \overline{X^2} \tag{16}$$

is proposed and it is argued through simulations (the asymptotic properties are not studied) that this estimator is a more suitable first-step estimator. More recently, Zhao et al. (2021) uses a generalized Nakagami distribution to derive another closed-form moment-type estimator

$$\hat{m}_n^{\mathrm{MO3}} = \frac{1}{2}\frac{\overline{X^2}}{\overline{X^2 \log(X)} - \overline{X^2}\,\overline{\log X}}, \qquad \hat{O}_n^{\mathrm{MO3}} = \overline{X^2} \tag{17}$$

whose asymptotic variance is obtained and is shown to be very close to that of the MLE.

Setting up our SMOM estimators for Nakagami distributions is simple. A Stein operator is known for the Nakagami distribution (and is also easy to obtain, for example, through the density approach) and is given by

$$\mathcal{A}_\theta f(x) = 2m(O - x^2)f(x) + xOf'(x).$$

As the approach conceals no surprise, we postpone the details of the computations to the Supporting Information (Section C.4). In particular, one immediately sees that both previous modified MOM estimators fall directly within our general SMOM estimation procedure, with (16) obtained through $f_1(x) = 1$ and $f_2(x) = x$, and (17) through $f_1(x) = 1$ and $f_2(x) = \log(x)$. Our Theorem 2.4 immediately yields the asymptotic variance of these estimators, confirming Zhao et al. (2021) in that particular case and also indicating that $\hat{m}_n^{\mathrm{MO2}}$ is not competitive in terms of variance. All expected behaviors are illustrated in our simulation study, detailed in Section C.4 from the Supporting Information.

Despite their excellent performances, the above $\hat{\theta}_n^{\mathrm{ML}}$ and $\hat{\theta}_n^{\mathrm{MO3}}$ estimators are nevertheless plagued by numerical instability whenever the data contains 0's, as can happen for instance when the data is rounded to the first decimal. This is a classical issue, see, for example, Wilks (1990), and such data does occur in many real-life scenarios, as we shall illustrate below. Here, the flexibility of our SMOM estimators can be exploited to design explicit estimators with low asymptotic variance,

which are not sensitive to the presence of 0's in the data. Some explorations lead us to propose the compromise $f_1(x) = 1$ and $f_2(x) = x$ which yields the new estimator

$$\hat{m}_n^{\mathrm{ST1}} = \frac{1}{2} \frac{\overline{X^2}\,\overline{X}}{\overline{X^3} - \overline{X}\,\overline{X^2}}, \qquad \hat{O}_n^{\mathrm{ST1}} = \overline{X^2} \tag{18}$$

which is explicit, immediate to compute, does not suffer from the numerical instability of (17) and has a better variance (since it involves lower moments) than (16). We illustrate the various behaviors in Tables 7 and 8 where all the estimators are applied to the same samples rounded to the first decimal over a variety of problematic parameter ranges (the MLE is started at initial estimates provided by the estimator $\hat{\theta}_n^{\mathrm{MO2}}$). For the rather small sample size of $n = 50$, the MLE and $\hat{\theta}_n^{\mathrm{MO3}}$ estimator do not exist for a significant number of Monte Carlo samples; increasing the sample size increases the volume of zeros and MLE and $\hat{\theta}_n^{\mathrm{MO3}}$ break down completely for the more problematic parameter constellations, hence requiring more sophisticated approaches for these estimators. $\hat{\theta}_n^{\mathrm{ST}}$ remains good irrespective of the parameter values, range, or sample size.

To illustrate the use of our estimators on real-world data, we consider rainfall measurements for which it is well-established that Nakagami distributions provide an excellent fit (see, e.g., Bağci (2024)). Specifically, we use the 20,820 daily rainfall entries recorded at the Rhymney at Bargoed station (grid reference ST1559698381) between January 01, 1961, and December 31, 2017, available from the National River Flow Archive[1].

Aggregating this data by summing daily values over each month across the entire timespan, we construct twelve datasets (one per calendar month), each containing 57 yearly totals. All monthly datasets exhibit characteristics consistent with i.i.d. samples. Goodness-of-fit (GOF) tests indicate that the Nakagami distribution models the monthly data reasonably well. Since none of the monthly totals are zero, all parameter estimation methods yield consistent results across months. The corresponding shape estimates are plotted in Figure 2a.

TABLE 7 Bias and MSE for $NG(m, O)$ over 10,000 simulations of samples of size $n = 50$, with the prescribed parameters; each time the samples are rounded to the first decimal.

		Bias				**MSE**				**NE**			
θ_0		$\hat{\theta}_n^{\mathrm{MO2}}$	$\hat{\theta}_n^{\mathrm{MO3}}$	$\hat{\theta}_n^{\mathrm{ML}}$	$\hat{\theta}_n^{\mathrm{ST1}}$	$\hat{\theta}_n^{\mathrm{MO2}}$	$\hat{\theta}_n^{\mathrm{MO3}}$	$\hat{\theta}_n^{\mathrm{ML}}$	$\hat{\theta}_n^{\mathrm{ST1}}$	$\hat{\theta}_n^{\mathrm{MO2}}$	$\hat{\theta}_n^{\mathrm{MO3}}$	$\hat{\theta}_n^{\mathrm{ML}}$	$\hat{\theta}_n^{\mathrm{ST1}}$
(1, 1)	m	0.111	0.068	0.069	0.073	0.092	0.044	0.041	0.06	0	11	11	0
(0.8, 1)	m	0.096	0.073	0.078	0.058	0.064	0.03	0.029	0.04	0	32	32	0
(1.4, 0.8)	m	0.14	0.083	0.079	0.1	0.164	0.093	0.089	0.118	0	2	2	0
(3, 5)	m	0.227	0.164	0.159	0.185	0.583	0.422	0.412	0.48	0	0	0	0
(3, 1)	m	0.227	0.157	0.151	0.183	0.587	0.431	0.423	0.486	0	0	0	0
(2, 5)	m	0.177	0.118	0.115	0.138	0.291	0.188	0.183	0.225	0	0	0	0
(4, 0.5)	m	0.222	0.141	0.131	0.172	0.95	0.738	0.723	0.814	0	0	0	0
(8, 4)	m	0.476	0.416	0.414	0.435	3.57	3.19	3.18	3.32	0	0	0	0
(3, 3)	m	0.247	0.184	0.179	0.205	0.618	0.446	0.435	0.508	0	0	0	0
(0.8, 0.2)	m	0.095	0.147	0.167	0.055	0.065	0.048	0.05	0.04	0	74	74	0

TABLE 8 Bias and MSE for $NG(m, O)$ over 10,000 simulations of samples of size $n = 500$, with the prescribed parameters; each time the samples are rounded to the first decimal.

		Bias				MSE				NE			
θ_0		$\hat{\theta}_n^{MO2}$	$\hat{\theta}_n^{MO3}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST1}$	$\hat{\theta}_n^{MO2}$	$\hat{\theta}_n^{MO3}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST1}$	$\hat{\theta}_n^{MO2}$	$\hat{\theta}_n^{MO3}$	$\hat{\theta}_n^{ML}$	$\hat{\theta}_n^{ST1}$
$(1, 1)$	m	9.87e−3	0.019	0.023	4.51e−3	8.1e−3	3.73e−3	3.58e−3	5.06e−3	0	71	71	0
$(0.8, 1)$	m	0.01	0.035	0.044	5.06e−3	5.84e−3	3.44e−3	3.78e−3	3.4e−3	0	98	98	0
$(1.4, 0.8)$	m	8.95e−3	3.53e−3	3.07e−3	3.29e−3	0.013	6.62e−3	6.32e−3	8.83e−3	0	18	18	0
$(3, 5)$	m	0.024	0.015	0.014	0.019	0.05	0.034	0.033	0.04	0	0	0	0
$(3, 1)$	m	−1.37e−3	−0.015	−0.018	−8.89e−3	0.048	0.034	0.034	0.039	0	0	0	0
$(2, 5)$	m	0.016	8.62e−3	8.03e−3	0.011	0.024	0.014	0.014	0.018	0	0	0	0
$(4, 0.5)$	m	−0.062	−0.084	−0.09	−0.073	0.082	0.066	0.066	0.071	0	0	0	0
$(8, 4)$	m	1.65e−3	−8.39e−3	−0.01	−4.54e−3	0.289	0.252	0.25	0.265	0	0	0	0
$(3, 3)$	m	0.016	7.8e−3	6.71e−3	0.011	0.048	0.034	0.033	0.039	0	0	0	0
$(0.8, 0.2)$	m	7.2e−3	—	—	−1.57e−3	5.69e−3	—	—	3.35e−3	0	100	100	0

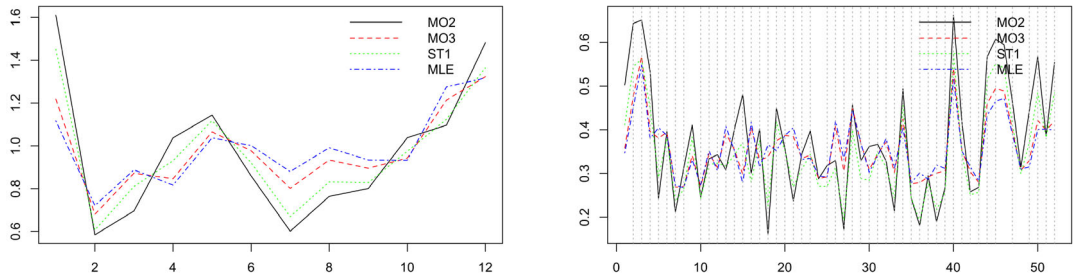

FIGURE 2 Estimates of the shape parameter m obtained via the four methods, using monthly data (a) and weekly data (b) from January 01, 1961 to December 31, 2017. Vertical lines indicate datasets containing zero values; in these cases, the estimators $\hat{\theta}_n^{\mathrm{ML}}$ and $\hat{\theta}_n^{\mathrm{MO3}}$ were adjusted by excluding zeros. Vertical lines are only visible in Figure 2b since this issue does not arise in the data plotted in Figure 2a. (a) Monthly rainfall data, (b) weekly rainfall data.

We also aggregate the data into weekly totals, producing 52 values per year across 57 years. In this higher-resolution dataset, the presence of zeros (the weeks corresponding to summer months contain as many as 9 zero-entries) causes the estimators $\hat{\theta}_n^{\mathrm{ML}}$ and $\hat{\theta}_n^{\mathrm{MO3}}$ to break down, leaving $\hat{\theta}_n^{\mathrm{ST1}}$ as the only reliable estimator for the full data set. GOF tests based on the Stein estimator confirm an excellent fit of the Nakagami distribution to the weekly data. We modified estimators $\hat{\theta}_n^{\mathrm{MO3}}$ and $\hat{\theta}_n^{\mathrm{ML}}$ to handle zero entries. This comes at the cost of the loss of information. The resulting weekly estimates are shown in Figure 2b. Using $\hat{\theta}_n^{\mathrm{MO3}}$ or $\hat{\theta}_n^{\mathrm{ML}}$ estimators on corrected data does not confirm a good fit. Simulation results confirm that for such parameter constellations on rounded data with zeros only, our Stein estimator is reliable.

It may also be interesting to investigate GOF or CP analysis with our approach. A more detailed study of such data sets will be the topic of a future publication.

4 | DISCUSSION

In this paper, we have developed Stein's method of moments in the context of univariate continuous probability distributions, which can be characterized in a tractable manner via the density approach. Restricting ourselves to this setting has allowed us to develop a detailed asymptotic theory and analysis of "optimal functions" and to carefully assess performance via simulations; however, many directions for research remain. Whilst we have treated a number of important univariate continuous distributions, our treatment is not comprehensive. We refer the reader to the recent references Wang and Weiß (2023) and Nik and Weiß (2024) for applications to the Lindley, exponential, and inverse Gaussian distributions, as well as the discrete negative binomial distribution. In this direction, it would be interesting to develop a general theory for univariate discrete distributions akin to our detailed treatment of the continuous case. The density approach generalizes in a natural manner to multivariate continuous distributions (see Mijoule et al. (2023)), and SMOM has recently been applied in a multivariate setting to truncated multivariate distributions Fischer et al. (2025) and to the notoriously difficult problem of parameter estimation on the sphere by Fischer et al. (2024). Finally, a number of important univariate continuous distributions

do not have simple characterizations via the density approach, and characterizations are instead based on higher-order differential operators (e.g., the variance-gamma distribution Gaunt (2014)) or fractional operators (e.g., stable distributions Xu (2019)). It would therefore be interesting to extend Stein's method of moments beyond the current density method setting.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments and suggestions. The second and fifth authors are funded in part by an ARC Consolidator grant from ULB and FNRS Grant CDR/OL J.0200.24. The second author is, in addition, in part funded by EPSRC Grant EP/T018445/1. The third author is funded in part by EPSRC grant EP/Y008650/1 and EPSRC grant UKRI068. This research was funded in whole, or in part, by the UKRI EP/T018445/1. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

ENDNOTE

1https://nrfa.ceh.ac.uk/data/.

ORCID

Robert E. Gaunt https://orcid.org/0000-0001-6187-0657

REFERENCES

Adler, R., Feldman, R., & Taqqu, M. (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Springer Science & Business Media.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., & Swan, Y. (2023). Stein's Method Meets Computational Statistics: A Review of Some Recent Developments. *Statistical Science*, *38*(1), 120–139.

Arnold, B. C., Castillo, E., & Sarabia, J. M. (2001). A multivariate version of Stein's identity with applications to moment calculations and estimation of conditionally specified distributions. *Communications in Statistics - Theory and Methods*, *30*(12), 2517–2542.

Artyushenko, V., & Volovach, V. (2019). Nakagami Distribution Parameters Comparatively Estimated by the Moment and Maximum Likelihood Methods. *Optoelectronics, Instrumentation and Data Processing*, *55*, 237–242.

Bağci, K. (2024). Nakagami distribution for modeling monthly precipitations in van, türkiye. *International Journal of Environment and Geoinformatics*, *11*(3), 19–23.

Bai, Z., & Fu, J. (1987). On the maximum-likelihood estimator for the location parameter of a Cauchy distribution. *Canadian Journal of Statistics*, *15*(2), 137–146.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., & Mackey, L. (2019). Minimum Stein Discrepancy Estimators. *Advances in Neural Information Processing Systems*, *32*, 12964–12976.

Betsch, S., & Ebner, B. (2021). Fixed point characterizations of continuous univariate probability distributions and their applications. *Annals of the Institute of Statistical Mathematics*, *73*, 31–59.

Betsch, S., Ebner, B., & Klar, B. (2021). Minimum L^q-distance estimators for non-normalized parametric models. *Canadian Journal of Statistics*, *49*(2), 514–548.

Bloch, D. (1966). A Note on the Estimation of the Location Parameter of the Cauchy Distribution. *Journal of the American Statistical Association*, *61*(315), 852–855.

Carrasco, M., Chernov, M., Florens, J.-P., & Ghysels, E. (2007). Efficient estimation of general dynamic models with a continuum of moment conditions. *Journal of Econometrics*, *140*(2), 529–573.

Carrasco, M., & Florens, J.-P. (2000). Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*, *16*(6), 797–834.

Carrasco, M., & Florens, J.-P. (2002a). Efficient GMM Estimation Using the Empirical Characteristic Function. IDEI working paper.

Carrasco, M., & Florens, J.-P. (2002b). Simulation-Based Method of Moments and Efficiency. *Journal of Business & Economic Statistics*, *20*(4), 482–492.

Carrasco, M., & Florens, J.-P. (2014). On the asymptotic efficiency of GMM. *Econometric Theory*, *30*(2), 372–406.

Chernoff, H., Gastwirth, J. L., & Johns, M. V. (1967). Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation. *Annals of Mathematical Statistics*, *38*(1), 52–72.

Copas, J. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika*, *62*(3), 701–704.

DeGroot, M. H. (2005). *Optimal Statistical Decisions*. John Wiley & Sons.

Diaconis, P., & Zabell, S. (1991). Closed Form Summation for Classical Distributions: Variations on a Theme of de Moivre. *Statistical Science*, *6*(3), 284–302.

Ernst, M., Reinert, G., & Swan, Y. (2020). First-order covariance inequalities via Stein's method. *Bernoulli*, *26*(3), 2051–2081.

Ferguson, T. S. (1978). Maximum Likelihood Estimates of the Parameters of the Cauchy Distribution for Samples of Size 3 and 4. *Journal of the American Statistical Association*, *73*(361), 211–213.

Fischer, A., Gaunt, R. E., & Swan, Y. (2024). Stein's Method of Moments on the Sphere. arXiv:2407.02299.

Fischer, A., Gaunt, R. E., & Swan, Y. (2025). Stein's method of moments for truncated multivariate distributions. *Electronic Journal of Statistics*, *19*(1), 1784–1808.

Freue, G. V. C. (2007). The Pitman estimator of the Cauchy location parameter. *Journal of Statistical Planning and Inference*, *137*(6), 1900–1913.

Gabrielsen, G. (1982). On the unimodality of the likelihood for the Cauchy distribution: Some comments. *Biometrika*, *69*(3), 677–678.

Gaunt, R. E. (2014). Variance-gamma approximation via Stein's method. *Electronic Journal of Probability*, *19*, 1–33.

Gaunt, R. E., Mijoule, G., & Swan, Y. (2019). An algebra of Stein operators. *Journal of Mathematical Analysis and Applications*, *496*, 260–279.

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Doklady Akademii Nauk SSSR*, *188*, 6.

Gutmann, M., & Hyvärinen, A. (2010). *Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models*. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 297–304). JMLR Workshop and Conference Proceedings.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of Machine Learning Research*, *13*(2), 307–361.

Hannan, E. J. (1973). Central Limit Theorems for Time Series Regression. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *26*(2), 157–170.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, *50*(4), 1029–1054.

Hansen, L. P. (2012). Proofs for large sample properties of generalized method of moments estimators. *Journal of Econometrics*, *170*(2), 325–330.

Hayakawa, J., & Takemura, A. (2016). Estimation of exponential-polynomial distribution by holonomic gradient descent. *Communications in Statistics - Theory and Methods*, *45*(23), 6860–6882.

Hegde, L. M., & Dahiya, R. C. (1989). Estimation of the parameters in a truncated normal distribution. *Communications in Statistics - Theory and Methods*, *18*(11), 4177–4195.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics & Data Analysis*, *51*(5), 2499–2512.

Hyvärinen, A., & Dayan, P. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, *6*(4), 695–709.

Kolar, R., Jirik, R., & Jan, J. (2004). Estimator Comparison of the Nakagami-*m* Parameter and its Application in Echocardiography. *Radioengineering*, *13*(1), 8–12.

Koutrouvelis, I. A. (1982). Estimation of location and scale in Cauchy distributions using the empirical characteristic function. *Biometrika*, *69*(1), 205–213.

Kumar, N., Dixit, A., & Vijay, V. (2024). q-generalization of nakagami distribution with applications. *Japanese Journal of Statistics and Data Science*, *8*, 69–92.

Ley, C., Reinert, G., & Swan, Y. (2017). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability*, *27*, 216–241.

Ley, C., & Swan, Y. (2013). Stein's density approach and information inequalities. *Electronic Communications in Probability*, *18*(7), 1–14.

Liu, S., Kanamori, T., & Williams, D. J. (2022). Estimating density models with truncation boundaries using score matching. *Journal of Machine Learning Research*, *23*(186), 1–38.

Lyu, S. (2009). *Interpretation and generalization of score matching*. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 359–366). UAI Press.

Meintanis, S. G. (2016). A review of testing procedures based on the empirical characteristic function. *South African Statistical Journal*, *50*(1), 1–14.

Mijoule, G., Raic, M., Reinert, G., & Swan, Y. (2023). Stein's density method for multivariate continuous distributions. *Electronic Journal of Probability*, *28*, 1–40.

Miyoshi, N., & Shirai, T. (2015). *Downlink coverage probability in a cellular network with ginibre deployed base stations and nakagami-m fading channels*. In *13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (pp. 483–489). IEEE.

Nakagami, M. (1960). *The m-distribution—a general formula of intensity distribution of rapid fading*. In *Statistical methods in radio wave propagation* (pp. 3–36). Elsevier.

Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N., & Takemura, A. (2011). Holonomic Gradient Descent and its Application to the Fisher–Bingham Integral. *Advances in Applied Mathematics*, *47*(3), 639–658.

Newey, W., & McFadden, D. (1994). *Large sample estimation and hypothesis testing*. In *Handbook of Econometrics* (Vol. *4*, pp. 2113–2245). Elsevier Science.

Nik, S., & Weiß, C. H. (2024). Generalized Moment Estimators based on Stein Identities. *Journal of Statistical Theory and Applications*, *23*(3), 240–274.

Oates, C. (2024). *Minimum Kernel Discrepancy Estimators*. In A. Hinrichs, P. Kritzer, & F. Pillichshammer (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2022*. Springer Verlag.

Reeds, J. A. (1985). Asymptotic Number of Roots of Cauchy Location Likelihood Equations. *The Annals of Statistics*, *13*, 775–784.

Reyes, J., Rojas, M. A., Venegas, O., & Gómez, H. W. (2020). Nakagami distribution with heavy tails and applications to mining engineering data. *Journal of Statistical Theory and Practice*, *14*, 1–20.

Rothenberg, T. J., Fisher, F. M., & Tilanus, C. B. (1964). A Note on Estimation from a Cauchy Sample. *Journal of the American Statistical Association*, *59*(306), 460–463.

Saumard, A. (2019). Weighted Poincaré inequalities, concentration inequalities and tail bounds related to Stein kernels in dimension one. *Bernoulli*, *25*(4b), 3978–4006.

Schoutens, W. (2001). Orthogonal Polynomials in Stein's Method. *Journal of Mathematical Analysis and Applications*, *253*(2), 515–531.

Schwartz, J., Godwin, R. T., & Giles, D. E. (2013). Improved maximum-likelihood estimation of the shape parameter in the nakagami distribution. *Journal of Statistical Computation and Simulation*, *83*(3), 434–445.

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* (Vol. *6*, pp. 583–603). University of California Press.

Stein, C. (1986). *Approximate Computation of Expectations*. IMS.

Tegos, S. A., Tyrovolas, D., Diamantoulakis, P. D., Liaskos, C. K., & Karagiannidis, G. K. (2022). On the distribution of the sum of double-nakagami-m random vectors and application in randomly reconfigurable surfaces. *IEEE Transactions on Vehicular Technology*, *71*(7), 7297–7307.

Wang, S., & Weiß, C. H. (2023). New characterizations of the (discrete) Lindley distribution and their applications. *Mathematics and Computers in Simulation*, *212*, 310–322.

Weber, M. D., Leemis, L. M., & Kincaid, R. K. (2006). Minimum Kolmogorov–Smirnov test statistic parameter estimates. *Journal of Statistical Computation and Simulation*, *76*(3), 195–206.

Wiens, D. P., Cheng, J., & Beaulieu, N. C. (2003). A Class of Method of Moments Estimators for the Two-Parameter Gamma Family. *Pakistan Journal of Statistics*, *19*(1), 129–141.

Wilks, D. S. (1990). Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of Climate*, *3*, 1495–1501.

Xu, L. (2019). Approximation of stable law in Wasserstein-1 distance by Stein's method. *The Annals of Applied Probability*, *29*, 458–504.

Ye, Z.-S., & Chen, N. (2017). Closed-Form Estimators for the Gamma Distribution Derived from Likelihood Equations. *The American Statistician*, *71*(2), 177–181.

Yu, S., Drton, M., & Shojaie, A. (2022). Generalized score matching for general domains. *Information and Inference: A Journal of the IMA*, *11*(2), 739–780.

Zhang, J. (2010). A highly efficient L-estimator for the location parameter of the Cauchy distribution. *Computational Statistics*, *25*(1), 97–105.

Zhao, J., Kim, S., & Kim, H.-M. (2021). Closed-form estimators and bias-corrected estimators for the nakagami distribution. *Mathematics and Computers in Simulation*, *185*, 308–324.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ebner, B., Fischer, A., Gaunt, R. E., Picker, B., & Swan, Y. (2025). Stein's method of moments. *Scandinavian Journal of Statistics*, *52*(4), 1594–1624. https://doi.org/10.1111/sjos.70003