Maximum Likelihood and the Maximum Product of Spacings from the Viewpoint of the Method of Weighted Residuals

Computational and Applied Mathematics manuscript No.
(will be inserted by the editor)

Maximum Likelihood and the Maximum Product of Spacings from the Viewpoint of the Method of Weighted Residuals

Takuya Kawanishi

Received: date / Accepted: date

Abstract In parameter estimation, the maximum-likelihood method (ML) does not work for some distributions. In such cases, the maximum product of spacings method (MPS) is used as an alternative. However, the advantages and disadvantages of the MPS, its variants, and the ML are still unclear. These methods are based on the Kullback–Leibler divergence (KLD), and we consider applying the method of weighted residuals (MWR) to it. We prove that, after transforming the KLD to the integral over $[0, 1]$, the application of the collocation method yields the ML, and that of the Galerkin method yields the MPS and Jiang's modified MPS (JMMPS); and the application of zero boundary conditions yields the ML and JMMPS, and that of non-zero boundary conditions yields the MPS. Additionally, we establish formulas for the approximate difference among the ML, MPS, and JMMPS estimators. Our simulation for seven distributions demonstrates that, for zero boundary condition parameters, for the bias convergence rate, ML and JMMPS are better than the MPS; however, regarding the MSE for small samples, the relative performance of the methods differs according to the distributions and parameters. For non-zero boundary condition parameters, the MPS outperforms the other methods: the MPS yields an unbiased estimator and the smallest MSE among the methods. We demonstrate that from the viewpoint of the MWR, the counterpart of the ML is JMMPS not the MPS. Using this KLD-MWR approach, we introduce a unified view for comparing estimators, and provide a new tool for analyzing and selecting estimators.

Keywords Parameter estimation · Kullback–Leibler divergence · Bias · Mean squared error · Point collocation method · Galerkin method

We thank Maxine Garcia, PhD, from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

T. Kawanishi
Institute of Science and Engineering, Kanazawa University, Kakuma-machi, Kanazawa, 9201192 Japan
Tel.: +81-76-2344809
Fax: +123-76-2344829
E-mail: kawanishi@se.kanazawa-u.ac.jp

1 Introduction

The maximum-likelihood method (ML) is one of the most popular methods of parameter estimation in statistics. However, the ML does not work for some distributions. For example, Smith (1985) showed that for the three-parameter Weibull distribution, $F(x) = 1 - \exp\left[-\{(x-\mu)/\sigma\}^{\gamma}\right]$, if $\gamma < 2$, then the ML estimator (MLE) is not consistent, and if $\gamma < 1$, then there is no MLE. An alternative to the MLE for such cases is the maximum product of spacings method (MPS, Cheng and Amin, 1983), also known as the maximum spacing method (Ranneby, 1984), in which we minimize the product of spacings instead of the likelihood. The MPS can be applied to a wider class of distributions than the ML, and it shares several good properties with the ML, for example, the MPS estimators (MPSEs) are consistent and asymptotically normal. For details of the MPS, refer to Ekström (2008).

Several variants of the MPS have been proposed. For example, Ranneby and Ekström extended the MPS by replacing the logarithmic function with generic convex functions (e.g., Ekström, 2001). Huang and Lin (2014) added a fitting parameter to optimize the selection of the convex function. Jiang (2013) proposed another version of the MPS in which the lowermost and uppermost spacings were square rooted to obtain the product of n effective spacings; the original MPS has $n+1$. Jiang (2013) showed that their method provided a more accurate estimation for the three-parameter Weibull distribution than the MPS. However, little is known about the advantages and disadvantages of the MPS, its variants, and the ML, and which method should be used in which cases.

The ML, MPS, and their variants are derived from the Kullback–Leibler divergence (KLD). If we change the variables and move the interval of integration to $[0,1]$, then minimizing the KLD can be considered as a boundary value problem. Then, we can apply the method of weighted residuals (MWR) to it. The MWR is the name of a family of methods for solving differential and integral equations (Finlayson, 1972). The popular finite element method (FEM) is an MWR. In the MWR, we can choose the various trial and weighting functions; thus, we have a variety of methods, such as the point collocation method and Galerkin method. In this paper, we apply the MWR to minimizing the KLD. The transformation of the KLD followed by the application of the point collocation method leads to the ML, and the application of the Galerkin method leads to Jiang's modified version of the MPS (JMMPS). Additionally, we prove that the original MPS is good for distributions that have non-zero boundary values for the derivative of the distribution function with respect to the parameters.

In the following section, we explain the MPS and JMMPS, and provide necessary definitions. In Section 3, we present the main results of the application of the MWR to KLD. In Section 4, we demonstrate the approximate difference of score functions and estimators among the ML, MPS, and JMMPS. In Section 5, we present the simulation results for the relative performance of the ML, MPS, and JMMPS, and in Section 6, we discuss the results.

2 Maximum Product of Spacings and Jiang's Modified Version

Let $\{X_i\}_{i=1}^n$ be an independent and identically distributed (iid) sample from distribution F and $\{X_{(i)}\}$ be its order statistics. For brevity, we use the notation $F_i := F(X_{(i)};\theta)$. Additionally, we define the endpoints of x as follows:

$$ {}_*x = \inf\{x : F(x) > 0\}, \quad x_* = \sup\{x : F(x) < 1\}. $$

The spacings of sample $\{X_i\}_{i=1}^n$ are defined as

$$ D_i(\theta) := F_i - F_{i-1} \quad (i = 1,\ldots,n+1), $$

where F_0 and F_{n+1} are

$$ F_0 := \lim_{x \downarrow {}_*x} F(x;\theta) = 0 \quad \text{and} \quad F_{n+1} := \lim_{x \uparrow x_*} F(x;\theta) = 1, $$

respectively. Additionally, we express the partial derivatives as follows:

$$ \begin{aligned} F_{\theta,i} &:= \frac{\partial F(X_i;\theta)}{\partial \theta}, \\ D_{\theta,i} &:= F_{\theta,i} - F_{\theta,i-1}. \end{aligned} $$

Then $F_{\theta,0}$ and $F_{\theta,n+1}$ are defined as

$$ F_{\theta,0} := \lim_{x \downarrow {}_*x} F_\theta(x;\theta) \quad \text{and} \quad F_{\theta,n+1} := \lim_{x \uparrow x_*} F_\theta(x;\theta), $$

respectively.

In the MPS, we calculate the maximizer of the following product of spacings:

$$ \mathscr{P}_{\mathrm{MPS}}(\theta) = \prod_{i=1}^{n+1} D_i. $$

By contrast, in JMMPS, we maximize the following:

$$ \mathscr{P}_{\mathrm{JMMPS}}(\theta) = D_1^{1/2} \left(\prod_{i=2}^{n} D_i \right) D_{n+1}^{1/2}. $$

The log product of spacings, which is the counterpart of the log likelihood, is defined as follows:

$$ \log \mathscr{P}_{\mathrm{MPS}}(\theta) = \sum_{i=1}^{n+1} \log D_i, \tag{1} $$

$$ \log \mathscr{P}_{\mathrm{JMPPS}}(\theta) = \frac{1}{2}\log D_1 + \sum_{i=2}^{n} \log D_i + \frac{1}{2}\log D_{n+1} \tag{2} $$

for the MPS and JMMPS, respectively.

When we apply the ML to estimation, in many cases, we solve the likelihood equation (LE). The MPS and JMMPS equivalents of the LE are as follows; we call them the product of spacings equations (PSEs) of the MPS:

$$\sum_{i=1}^{n+1} \frac{D_{\theta,i}}{D_i} = 0, \tag{3}$$

and JMMPS:

$$\frac{1}{2}\frac{D_{\theta,1}}{D_1} + \sum_{i=2}^{n} \frac{D_{\theta,i}}{D_i} + \frac{1}{2}\frac{D_{\theta,n+1}}{D_{n+1}} = 0. \tag{4}$$

3 Application of the Method of Weighted Residuals to the Kullback–Leibler Divergence

3.1 Kullback–Leibler Divergence over $[0,1]$

The KLD is defined as follows:

$$D_{\mathrm{KL}} = \int_{*x}^{x_*} \log\left(\frac{f(x;\theta_0)}{f(x;\theta)}\right) f(x;\theta_0)\,dx.$$

To apply the MWR to the KLD, we move the interval of integration of the KLD from $({}_*x, x_*)$ to $[0,1]$, where the latter is the range of the distribution function and is a closed interval. We introduce the following notation:

$$\begin{aligned} Q(p,\theta) &:= F^{-1}(p,\theta), \\ Q_0(p) &:= Q(p,\theta_0), \\ \tilde{F}(p,\theta) &:= F(Q_0(p),\theta). \end{aligned}$$

Then, by applying the change of variables, for the negative KLD, we have

$$-D_{\mathrm{KL}} = \int_0^1 \left[\log\frac{\partial \tilde{F}}{\partial p}(p,\theta) - \log\frac{\partial \tilde{F}}{\partial p}(p,\theta_0)\right] dp.$$

We assume that there exists the global maximum of $-D_{\mathrm{KL}}$, and that the maximizer is the solution of the following equation:

$$-\frac{\partial D_{\mathrm{KL}}}{\partial \theta} = \int_0^1 \frac{\partial p}{\partial \tilde{F}}\frac{\partial \tilde{F}_\theta}{\partial p}(p,\theta)\,dp = 0. \tag{5}$$

3.2 Method of Weighted Residuals

The MWR, including the FEM, is a popular method for solving differential equations. The procedure is as follows: (i) approximate the equation to be solved using trial functions, (ii) define residual R as the difference between the approximate and exact

equations, (iii) define appropriate weighting function w, and (iv) force the weighted residual, that is, the inner product of the residual and weight, to be zero:

$$(R,w) := \int_0^1 R(p)w(p)\,dp = 0. \tag{6}$$

In our setting, we determine the estimator $\hat{\theta}$ that approximately satisfies (5) given sample $\{X_i\}_{1\leq i\leq n}$ by introducing the empirical quantile function:

$$Q^{\mathrm{E}} : [0,1] \to \mathbb{R}.$$

We require Q^{E} to be continuous and to satisfy

$$Q^{\mathrm{E}}(p_i) = X_{(i)}.$$

Additionally, we define the empirical version of $\tilde{F}$:

$$\tilde{F}^{\mathrm{E}}(p,\theta) := F(Q^{\mathrm{E}}(p),\theta).$$

Then our problem is to solve the following under Q^{E} of the sample:

$$\int_0^1 \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \frac{\partial \tilde{F}^{\mathrm{E}}_{\theta}(p,\theta)}{\partial p}\,dp = 0. \tag{7}$$

To solve this, let R be

$$R(p,\theta) := \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \frac{\partial \tilde{F}^{\mathrm{E}}_{\theta}}{\partial p}(p,\theta),$$

and choose an appropriate weighting function w and make $(R,w) = 0$. When solving differential equations, the unknown values are the values of the function at nodes p_i (or x_i); however, our unknown values are the parameters θ. The weighted residual (6) results in the same number of equations as the dimensions of the parameters, and we solve them to obtain the estimators.

3.3 Point Collocation Method Leads to Maximum Likelihood

In the point collocation method, the weighting function is chosen to be Dirac's delta functions at selected points; in our case, a natural choice is n equally spaced points in $[0,1]$, $i/(n+1)$ $(i = 1,\ldots,n)$:

$$w(p) = \sum_{i=1}^{n} \delta\left(p - \frac{i}{n+1}\right). \tag{8}$$

Then the weighted residual of the integral equation (7) is

$$\begin{aligned}(R,w) &= \int_0^1 \left[\frac{\partial F_{\theta}}{\partial F}(Q^{\mathrm{E}}(p),\theta)\right] \sum_{i=1}^{n} \delta\left(p - \frac{i}{n+1}\right) dp \\ &= \sum_{i=1}^{n} \frac{f_{\theta}(x_i,\theta)}{f(x_i,\theta)}.\end{aligned}$$

Setting (R,w) to zero yields the LE.

3.4 Galerkin Method Leads to Jiang's Modified Maximum Product of Spacings

The Galerkin method makes the residual orthogonal to the basis of the trial function. To achieve this, the typical approach is to make the weighting function have the same form as the trial function. We use the piecewise linear function for the trial function, which is the simplest function applicable to our case, with grid points $p_i = i/(n+1)$, $(i = 1, \ldots, n)$. For trial function ϕ, we have

$$\phi_i = \begin{cases} \dfrac{y - p_{i-1}}{p_i - p_{i-1}} & p_{i-1} \le y \le p_i, \\ \dfrac{p_{i+1} - y}{p_{i+1} - p_i} & p_i \le y \le p_{i+1}, \end{cases}$$

$$\left(\frac{d\phi}{dp}\right)_i = \begin{cases} \dfrac{1}{p_i - p_{i-1}} & p_{i-1} \le y \le p_i, \\ \dfrac{-1}{p_{i+1} - p_i} & p_i \le y \le p_{i+1}. \end{cases}$$

Using these trial functions, we have the following approximations:

$$\left(\frac{\partial \tilde{F}^{\mathrm{E}}}{\partial p}\right)^h = \sum_{i=1}^{n} \tilde{F}_i^{\mathrm{E}} \left(\frac{d\phi}{dp}\right)_i,$$
$$\left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p}\right)^h = \sum_{i=1}^{n} \tilde{F}_{\theta,i}^{\mathrm{E}} \left(\frac{d\phi}{dp}\right)_i,$$

where

$$\tilde{F}_i^{\mathrm{E}} = F(Q^{\mathrm{E}}(p_i), \theta) = F(X_{(i)}, \theta) = F_i,$$
$$\tilde{F}_{\theta,i}^{\mathrm{E}} = \frac{\partial}{\partial \theta} F(Q^{\mathrm{E}}(p_i), \theta) = F_\theta(X_{(i)}, \theta) = F_{\theta,i}.$$

Additionally, let the weighting function have the same form as the trial function:

$$w = \sum_{i=1}^{n} \phi_i. \tag{9}$$

Our problem is then written as

$$\left(\left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h, w \right) = 0. \tag{10}$$

For each interval $[p_{i-1}, p_i]$, $2 \le i \le n$, the integral is

$$\begin{aligned}
&\int_{p_{i-1}}^{p_i} \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\, d\xi \\
&= \int_{p_{i-1}}^{p_i} \frac{\tilde{F}_{\theta,i}^{\mathrm{E}} - \tilde{F}_{\theta,i-1}^{\mathrm{E}}}{p_i - p_{i-1}} \frac{p_i - p_{i-1}}{\tilde{F}_i^{\mathrm{E}} - \tilde{F}_{i-1}^{\mathrm{E}}} \left\{ \frac{\xi - p_{i-1}}{p_i - p_{i-1}} + \frac{p_i - \xi}{p_i - p_{i-1}} \right\} d\xi \\
&= (p_i - p_{i-1}) \frac{\tilde{F}_{\theta,i}^{\mathrm{E}} - \tilde{F}_{\theta,i-1}^{\mathrm{E}}}{\tilde{F}_i^{\mathrm{E}} - \tilde{F}_{i-1}^{\mathrm{E}}} = \frac{1}{n+1} \frac{D_{\theta,i}}{D_i}.
\end{aligned}$$

For the lowermost and uppermost spacings,

$$\begin{aligned}
\int_0^{p_1} \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\, d\xi &= \int_0^{p_1} \frac{\tilde{F}_{\theta,1}^{\mathrm{E}}}{\tilde{F}_1^{\mathrm{E}}} \left(\frac{\xi}{p_1} \right) d\xi = \frac{1}{2(n+1)} \frac{D_{\theta,1}}{D_1} \\
\int_{p_n}^1 \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\, d\xi &= \int_{p_n}^1 \frac{-\tilde{F}_{\theta,n}^{\mathrm{E}}}{1 - \tilde{F}_n^{\mathrm{E}}} \frac{1-\xi}{1-p_n} d\xi = \frac{1}{2(n+1)} \frac{D_{\theta,n+1}}{D_{n+1}}.
\end{aligned}$$

The summation yields

$$0 = (R, w) = \frac{1}{n+1} \left(\frac{1}{2} \frac{D_{\theta,1}}{D_1} + \sum_{i=2}^{n} \frac{D_{\theta,i}}{D_i} + \frac{1}{2} \frac{D_{\theta,n+1}}{D_{n+1}} \right),$$

which is the PSE of JMMPS (Eq. (4)).

3.5 Non-Zero Boundary Values and the Maximum Product of Spacings

Thus far, we have used the weighting function of (8) and (9). Hence, we implicitly assume that the boundary values of $\tilde{F}_\theta$ are

$$\tilde{F}_\theta(0) = \tilde{F}_\theta(1) = 0. \tag{11}$$

However, we can adapt the above methods to distributions that have non-zero, and possibly unknown boundary values of $\tilde{F}_\theta(0)$ or $\tilde{F}_\theta(1)$. To achieve this, we add the boundary basis,

$$\phi_0 = \frac{y}{p_1} \quad \phi_{n+1} = \frac{y - p_n}{1 - p_n}, \tag{12}$$

and let the weighting function, for example, for the Galerkin method, be

$$w = \sum_{i=0}^{n+1} \phi_i,$$

and let the weighted residual Eq. (10) incorporate the boundary conditions

$$\int_0^1 \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\, dp + w(0)\tilde{F}_\theta(0) - w(1)\tilde{F}_\theta(1) = 0.$$

We combine the terms for the lowermost spacing and lower boundary condition:

$$\begin{aligned} r_0 &:= \int_0^{p_1} \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\,dp + w(0)\tilde{F}_\theta^{\mathrm{E}}(0) \\ &= \frac{1}{n+1} \frac{\tilde{F}_{\theta,1}^{\mathrm{E}} - \tilde{F}_\theta^{\mathrm{E}}(0)}{p_1} \frac{p_1}{\tilde{F}_1^{\mathrm{E}}} + w(0)\tilde{F}_\theta^{\mathrm{E}}(0). \end{aligned}$$

At the boundary of $p = 0$, we have $\partial p / \partial \tilde{F}^{\mathrm{E}} = 1$; thus, it is reasonable to set $p_1 / \tilde{F}_1^{\mathrm{E}} = 1$. Additionally, we have $w(0) = 1$ and $p_1 = 1/(n+1)$. Hence, the $\tilde{F}_\theta^{\mathrm{E}}(0)$ terms cancel, and we have

$$r_0 = \frac{1}{n+1} \frac{\tilde{F}_{\theta,1}^{\mathrm{E}}}{\tilde{F}_1^{\mathrm{E}}}.$$

For the uppermost spacing, the same procedure applies and

$$\begin{aligned} r_{n+1} &:= \int_{p_n}^1 \left(\frac{\partial \tilde{F}_\theta^{\mathrm{E}}}{\partial p} \frac{\partial p}{\partial \tilde{F}^{\mathrm{E}}} \right)^h w\,dp - w(1)\tilde{F}_\theta^{\mathrm{E}}(1) \\ &= \frac{1}{n+1} \frac{-\tilde{F}_{\theta,n}^{\mathrm{E}}}{1 - \tilde{F}_n^{\mathrm{E}}}. \end{aligned}$$

Then, for the Galerkin method, the weighted residual finally becomes

$$\begin{aligned} 0 = (R, w) &= r_0 + \sum_{i=2}^{n} \frac{D_{\theta,i}}{D_i} + r_{n+1} \\ &= \sum_{i=1}^{n+1} \frac{1}{n+1} \frac{D_{\theta,i}}{D_i}, \end{aligned}$$

which is the PSE of the MPS.

For convenience, we call the condition expressed in Eq. (11) the zero-BC, and the parameter that satisfies the condition the zero-BC parameter. We call the parameter that does not satisfy Eq. (11) the non-zero-BC parameter.

4 Difference among the methods

4.1 Difference of score functions

We start by transforming each term of the partial derivative of the log product of spacings using the trapezoidal rule of integration. For $2 \le i \le n$,

$$\begin{aligned} \frac{F_{\theta,i} - F_{\theta,i-1}}{F_i - F_{i-1}} &= \frac{1}{F_i - F_{i-1}} \int_{F_i}^{F_i} \frac{\partial F_\theta}{\partial F}\,dF = \frac{1}{F_i - F_{i-1}} \int_{F_i}^{F_i} \frac{\partial F_\theta}{\partial x} \frac{\partial x}{\partial F}\,dF \\ &= \frac{1}{F_i - F_{i-1}} \int_{F_{i-1}}^{F_i} \frac{f_\theta}{f}\,dF = \frac{1}{2} \left(\left.\frac{f_\theta}{f}\right|_i + \left.\frac{f_\theta}{f}\right|_{i-1} \right) + O((F_i - F_{i-1})^2). \end{aligned}$$

With this relationship, noting that $O((F_i - F_{i-1})) = O(1/n)$, we can calculate the approximate difference of score functions between the MPS and ML, and between JMMPS and the ML:

$$\begin{aligned}\varepsilon_{\mathrm{MPS/ML}}(\theta) &:= \frac{\partial}{\partial\theta} l_{\mathrm{MPS}}(\theta) - \frac{\partial}{\partial\theta} l_{\mathrm{ML}}(\theta) \\ &= \frac{F_{\theta,1}}{F_1} - \frac{1}{2}\left.\frac{f_\theta}{f}\right|_{X_{(1)}} + \frac{-F_{\theta,n}}{1-F_n} - \frac{1}{2}\left.\frac{f_\theta}{f}\right|_{X_{(n)}} + O\left(\frac{1}{n}\right); \end{aligned} \tag{13}$$

and

$$\begin{aligned}\varepsilon_{\mathrm{JMMPS/ML}}(\theta) &:= \frac{\partial}{\partial\theta} l_{\mathrm{JMMPS}}(\theta) - \frac{\partial}{\partial\theta} l_{\mathrm{ML}}(\theta) \\ &= \frac{F_{\theta,1}}{2F_1} - \frac{1}{2}\left.\frac{f_\theta}{f}\right|_{X_{(1)}} + \frac{-F_{\theta,n}}{2(1-F_n)} - \frac{1}{2}\left.\frac{f_\theta}{f}\right|_{X_{(n)}} + O\left(\frac{1}{n}\right).\end{aligned}$$

4.2 Difference of estimators

Substituting the MPSE $\hat{\theta}_{\mathrm{MPS}}$ for θ in (13) and noting that $\partial l_{\mathrm{MPS}}(\hat{\theta}_{\mathrm{MPS}})/\partial\theta = 0$, we have

$$\varepsilon_{\mathrm{MPS/ML}}(\hat{\theta}_{\mathrm{MPS}}) = \frac{\partial l_{\mathrm{MPS}}}{\partial\theta}(\hat{\theta}_{\mathrm{MPS}}) - \frac{\partial l_{\mathrm{ML}}}{\partial\theta}(\hat{\theta}_{\mathrm{MPS}}) = -\frac{\partial l_{\mathrm{ML}}}{\partial\theta}(\hat{\theta}_{\mathrm{MPS}}).$$

From this and the fact that $\partial l_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}})/\partial\theta = 0$,

$$\begin{aligned}-\varepsilon_{\mathrm{MPS/ML}}(\hat{\theta}_{\mathrm{MPS}}) &= \frac{\partial l_{\mathrm{ML}}}{\partial\theta}(\hat{\theta}_{\mathrm{MPS}}) - \frac{\partial l_{\mathrm{ML}}}{\partial\theta}(\hat{\theta}_{\mathrm{ML}}) \\ &\approx \frac{\partial^2 l_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}})}{\partial\theta^2}(\hat{\theta}_{\mathrm{MPS}} - \hat{\theta}_{\mathrm{ML}}),\end{aligned}$$

where the last approximation is from Taylor's theorem. By multiplying both sides by the inverse of information, we finally obtain

$$\hat{\theta}_{\mathrm{MPS}} - \hat{\theta}_{\mathrm{ML}} \approx -\left\{\frac{\partial^2 l_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}})}{\partial\theta^2}\right\}^{-1} \varepsilon_{\mathrm{MPS/ML}}(\hat{\theta}_{\mathrm{MPS}}).$$

We apply the same procedure to JMMPS, which yields

$$\hat{\theta}_{\mathrm{JMMPS}} - \hat{\theta}_{\mathrm{ML}} \approx -\left\{\frac{\partial^2 l_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}})}{\partial\theta^2}\right\}^{-1} \varepsilon_{\mathrm{JMMPS/ML}}(\hat{\theta}_{\mathrm{JMMPS}}).$$

4.3 Examples

Exponential distribution

The cumulative distribution function (cdf) of the exponential distribution with the rate parameter λ is

$$F(x;\lambda) = 1 - e^{-\lambda x}.$$

Note that we have $\tilde{F}_\lambda(0) = \tilde{F}_\lambda(1) = 0$, that is, the rate parameter is zero-BC. With some calculus, for the MPS we have

$$\varepsilon_{\mathrm{MPS/ML}}(\lambda) = -\frac{X_{(1)}}{2} - \frac{X_{(n)}}{2} + O\left(\frac{1}{n}\right).$$

On the right-hand side, the term $X_{(n)}/2$ dominates, and from the quantile function $x = Q(F) = (1/\lambda)\log(1-F)$ and the fact that $1 - F_n \approx 1/n$, we have

$$X_{(n)} = \frac{1}{\lambda}\log(1-F_n) \approx \frac{1}{\lambda}\log(1/n) = -\frac{1}{\lambda}\log n;$$

hence,

$$\varepsilon_{\mathrm{MPS/ML}}(\lambda) = O(\log n).$$

For JMMPS,

$$\varepsilon_{\mathrm{JMMPS/ML}}(\lambda) = -\frac{1}{2\lambda} + O\left(\frac{1}{n}\right).$$

The information is $\partial^2 l_{\mathrm{ML}}(\hat{\theta}_{\mathrm{ML}})/\partial\theta^2 = n/\hat{\lambda}_{\mathrm{ML}}^2$; thus,

$$\hat{\lambda}_{\mathrm{MPS}} - \hat{\lambda}_{\mathrm{ML}} = \frac{\hat{\lambda}_{\mathrm{ML}}}{n} O(\log n) = O\left(\frac{\log n}{n}\right) \tag{14}$$

$$\hat{\lambda}_{\mathrm{JMMPS}} - \hat{\lambda}_{\mathrm{ML}} = O\left(\frac{1}{n}\right). \tag{15}$$

The bias of $\hat{\lambda}_{\mathrm{ML}}$ is λ_0/n, which is of $O(1/n)$. From this, and from (14) and (15), we prove that, for the rate parameter of the exponential distribution, the bias of the MLE and JMMPS estimator (JMMPSE) are of $O(1/n)$, and by contrast, the bias of the MPSE is of $O((\log n)/n)$.

Uniform distribution

For the uniform distribution, the LE is not relevant. However, its counterparts for the MPS and JMMPS, PSEs (Eqs. (3), (4)), have valid solutions. We consider the simplest example. Suppose that

$$F(x;b) = x/b.$$

Then its log product of spacings of the MPS is

$$l_{\mathrm{MPS}}(\theta) = \sum_{i=1}^{n} \log(X_{(i)} - X_{(i-1)}) - n\log b + \log\left(1 - \frac{X_{(n)}}{b}\right),$$

where we interpret $X_{(0)} = 0$. Then the PSE for the MPS is

$$\begin{aligned} 0 = \frac{\partial}{\partial b} l_{\mathrm{MPS}}(\theta) &= -\frac{n}{\hat{b}} + \frac{1}{\left(1 - x_n/\hat{b}\right)} \frac{x_n}{\hat{b}^2} \\ &= -\frac{n}{\hat{b}} + \frac{X_{(n)}}{\hat{b}(\hat{b} - X_{(n)})}. \end{aligned}$$

The solution of the PSE is

$$\hat{b}_{\mathrm{MPS}} = \frac{n+1}{n} X_{(n)}.$$

We apply the same method to JMMPS, which yields

$$\hat{b}_{\mathrm{JMMPS}} = \frac{2n}{2n-1} X_{(n)}.$$

From this and from the fact that $\hat{b}_{\mathrm{ML}} = X_{(n)}$ and $\mathbb{E}X_{(n)} = bn/(n+1)$, we conclude that for the uniform distribution, the MPS yields an unbiased estimator, whereas the MLE and JMMPSE have biases of $O(1/n)$.

5 Simulation

5.1 Methods

We numerically generated $N = 10000$ iid samples of size n from various probability distributions. Then, the MLE, MPSE, and JMMPSE were calculated for the same set of samples. Then the bias and MSE were evaluated from N estimators. All the estimators were calculated by directly minimizing the log likelihood or log product of spacings (Eqs. (1), (2)) rather than solving the LE or PSEs. The list of the investigated distributions is presented in Table 1.

In the implementation of the MPS and JMMPS, we need to manage tied values. The MPS with the simple implementation of minimizing Eq. (1) or (2) breaks down when there are ties (Cheng and Amin, 1983). For a discussion of ties, see Ekström (2008). Suppose we have ρ distinct values in a sample of size n. Let $\{x_{(i)}\}_{i=1}^{\rho}$ be the order statistics of these distinct values and let λ_i be the occurrences of $x_{(i)}$, and let F_i represent $F(x_{(i)})$. The original definition of the modified product of spacings by Jiang (2013) is

$$\mathscr{P}_{\mathrm{JMMPS}} = \prod_{i=1}^{\rho} \left[\{F_i - F_{i-1}\}\{F_{i+1} - F_i\}\right]^{\lambda_i/2}.$$

Note that this formulation is applicable to cases in which there are ties. An equivalent formulation can be obtained by introducing weights w_i as follows:

$$l_{\mathrm{PS}}(\theta|X) = \sum_{i=1}^{\rho+1} w_i \log D_i(\theta|X).$$

Weights w_i for JMMPS are

$$w_{\mathrm{JMMPS},i} = \begin{cases} \frac{1}{2}\lambda_1 & i = 1, \\ \dfrac{\lambda_{i-1}+\lambda_i}{2} & i = 2,\ldots,\rho, \\ \frac{1}{2}\lambda_\rho & i = \rho+1, \end{cases}$$

and those for the MPS are

$$w_{\mathrm{MPS},i} = \begin{cases} \lambda_i & i = 1,\ldots,n, \\ 1 & i = \rho+1. \end{cases}$$

For the calculation of the cdf and probability density function, generation of random samples, and minimization of the log likelihood and log product of spacings, we used the Scipy library for Python, particularly the packages `scipy.stats` and `scipy.optimize`. The Python snippets used in this paper are available at Github `https://github.com/takuyakawanishi/kldmwr`.

5.2 Results

Exponential distribution: Figure 1 shows the bias and MSE of the estimators of the exponential distribution using two different parameterizations: one using the rate parameter λ (left panels) and the other using the scale parameter η (right panels). The lower panels show the relationship between n and $n$bias. For both parameterizations, $n$bias of the MPSE increases or decreases linearly with respect to $\log n$, which means that the absolute value of the bias of the MPSE decreases at the rate of $O((\log n)/n)$. By contrast, $n$bias of the MLE and JMMPSE remains constant throughout the range of n investigated, which means that the biases of the MLE and JMMPSE decrease at the rate of $O(1/n)$. The results for the rate parameter agree well with the example in the previous section. The upper panels show the nMSEs, in which we observe that the MSEs of the MLE, JMMPSE, and MPSE all converge to the same value as n increases. The difference among the methods can be observed only when the sample size is small. For rate parameter λ, (A) the MSE for small samples is largest for the ML, followed by JMMPS and the MPS. For scale parameter η, however, (B) the MSE for small samples is largest in the MPS. By contrast, the MSE in the ML and JMMPS have practically no small sample effects. These two tendencies will be repeatedly observed in what follows, and we call (A) 'pattern A' and (B) 'pattern B.'

Normal distribution: Figure 2 shows the results for the normal distribution. We compared the two cases in which (a) location parameter μ_0 is known and variance σ_0^2 is unknown, and (b) both location parameter μ_0 and variance σ_0^2 are unknown. Similar to the exponential distribution, we observe that the bias of the MPSE is of $O((\log n)/n)$, whereas those of the MLE and JMMPSE are of $O(1/n)$. For both cases (a) and (b), the MSE exhibits pattern B, that is, we expect a larger MSE for the MPSE than the MLE or JMMPSE in small samples. Another interesting difference between the MLE and JMMPSE is that in case (a), the MLE is unbiased and the JMMPSE is biased; however, in case (b), the MLE is biased and the JMMPSE is unbiased. It

would be worth investigating the reason for this to understand the difference between the ML and JMMPS, but we do not examine this point further in this paper.

Pareto distribution: The results for the Pareto distribution are shown in Figure 3. For the scale parameter η (Figure 3(a)), we have $\tilde{F}_\eta(0) = -\eta$, that is, the scale parameter η is of non-zero-BC. We observe that the bias and MSE of $\hat{\eta}$ exhibit different patterns from the exponential or normal distributions. Regarding the bias, in the lower panel, the MPSE is unbiased, whereas the MLE and JMMPSE are biased, although the biases of the MLE and JMMPSE remain of $O(1/n)$. The results for the MSE are shown in the upper panel. Note that we plotted n^2 MSE rather than n MSE. For all the ML, MPS, and JMMPS, n^2 MSE remains constant, which means that the MSE decreases as $O(1/n^2)$, and the n^2 MSEs of the three methods do not converge to the same value. Additionally, we observe that the n^2 MSEs are highest in the ML and lowest in the MPS. We call this 'pattern C.' Figure 3(b) shows the shape parameter α, which is zero-BC. Both the bias and MSE behave similarly to the results for the rate parameter of the exponential distribution; that is, the rate of bias convergence of the MPSE is of $O((\log n)/n)$ and the MSE exhibits pattern A.

Gamma distribution: Figure 4 shows the results for the Gamma distribution. Figure 4(a) shows the scale parameter η, and Figure 4(b) shows the shape parameter α. For both parameters, the biases of the MLE and JMMPSE are of $O(1/n)$, and that of the MPSE is of $O((\log n)/n)$. The MSE of $\hat{\eta}$ exhibits pattern B and that of $\hat{\alpha}$ exhibits pattern A. Additionally, we observe that the JMMPSE of η is unbiased.

Beta distribution: For the beta distribution, the results for $(\alpha_0, \beta_0) = (0.5, 1)$ are shown in Figure 5. Figure 5(a) shows $\hat{\alpha}$ and Figure 5(b) shows $\hat{\beta}$. The bias, in addition to the MSE, exhibits the same tendencies for both parameters: the biases of the MLE and JMMPSE are of $O(1/n)$, and the bias of the MPSE is of $O((\log n)/n)$; the MSE exhibits pattern A. Although we do not present the results here, we verified that these tendencies are observed with other combinations of parameters (e.g., (0.5, 0.5), (0.5, 2), (1, 2), (2, 2)).

Uniform distribution: The results for the uniform distribution are shown in Figure 6. For this distribution, the lower a and upper b parameters are both non-zero-BC. The results exhibit the same tendency as those of the scale parameter of the Pareto distribution. For both a and b, the MPSE are unbiased, the MLE and JMMPSE are biased, and the magnitude of the bias is JMMPSE $<$ MLE. Regarding the MSE, it converges as $O(1/n^2)$, the MPSE is the smallest, and the MLE is the largest (pattern C). These simulation results are consistent with the analytical results in the previous section.

Cauchy distribution: We investigated the case in which both location μ and scale η parameters are unknown, and the results are shown in Figure 7. We only show the results for the scale parameter α because there is no difference among the methods for the location parameter. The MLE and JMMPSE are unbiased, whereas the MPSE is biased, the bias is of $O(1/n)$, and the MSE exhibits pattern B.

We summarize the results in Table 1. In the table, we include the results that we do not show in the figures; all of them exhibit no difference among the methods. We call this pattern (for MSE) 'pattern O.' As expected from the results in Sections 3 and 4, whether the parameter is zero-BC (Eq. (11)) significantly affects the asymptotic behavior of the bias and the small sample behavior of the MSE of the estimators.

Regarding the choice of method, we should note that zero-BC and non-zero-BC are characteristics not only of the distribution but also of the parameter. For example, the Pareto distribution has both types of parameters: the scale parameter is non-zero-BC and the shape parameter is zero-BC. In such cases, the choice of method is based on which parameter we consider is most important.

6 Discussion

We applied the MWR to minimize the KLD to demonstrate that the point collocation method leads to the ML and the Galerkin method leads to JMMPS. Moreover, the traditional MPS was shown to be good for cases in which the boundary values of $\tilde{F}_\theta$ are non-zero; that is, from the viewpoint of the MWR, the counterpart of the ML is JMMPS and not the MPS. This proximity of JMMPS to the ML was verified by the results in Sections 4 and 5: the results for the MLE and JMMPSE are qualitatively the same if we ignore the difference between unbiasedness and bias of $O(1/n)$; the bias and the MSE of JMMPSE always lie between those of the MLE and MPSE.

We demonstrated that the behaviors of the bias and MSE are strongly affected by whether the parameter satisfies zero-BC. For zero-BC parameters, the bias convergence rate is always better for the ML and JMMPS than the MPS; however, the small sample MSE sometimes exhibits pattern A, in which the MPSE outperforms the MLE and JMMPSE, and sometimes pattern B, in which the MLE and JMMPSE are better than the MPSE. For non-zero-BC parameters, we always observed that the MPS is the best among the three methods: the MPSE is always unbiased, and the MSE of the MPSE is always smaller than the others for any size n.

Regarding the MPS, we observed that (a) the performance of the MPS for non-zero-BC parameters is significantly better than that of the ML and JMMPS, and (b) for the uniform distribution, the MPS, together with the JMMPS, has a PSE whose solution is relevant, whereas for the ML, we have no LE. (a) and (b) should be added to the list of advantages of MPS that has been established by researchers (e.g., Ekström, 2008).

In this paper, we only investigated distributions for which the ML works. This is one of the major limitations of this paper because the MPS and JMMPS are most likely to be used for distributions for which the ML does not work. The relative performance of the MPS and JMMPS for those distributions is of theoretical and practical importance.

With this KLD-MWR approach, we can consider the ML, MPS, and JMMPS from a unified point of view. This approach provides us with a new tool for investigating the difference among estimators and helping us in the choice of estimators. Thus far, we have examined only the point collocation and Galerkin methods. However, the MWR allows more methods, including the subdomain and least square; hence, applying these methods of the MWR to KLD may also yield insights.

References

Cheng R C H, Amin N A K (1983) Estimating parameters in continuous distributions with a shifted origin. J R Stat Soc B 45(3):394–403

Ekström M (2001) Consistency of generalized maximum spacing estimates. Scand J Stat 28(2):343–354

Ekström M (2008) Alternatives to maximum likelihood estimation based on spacings and the Kullback-Leibler divergence. J Stat Plan Infer 138(6):1778–1791

Finlayson B A (1972) The method of weighted residuals and variational principles, with application in fluid mechanics, heat and mass transfer. Academic Press, New York

Huang C, Lin J G (2014) Modified maximum spacings method for generalized extreme value distribution and applications in real data analysis Metrika 77(7):867–894

Jiang R (2013) A modified MPS method for fitting the 3-parameter Weibull distribution. QR2MSE 2013 - Proceedings of 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, 983–985

Ranneby B (1984) The maximum spacing method. An estimation method related to the maximum likelihood method. Scand J Stat 11:93–112

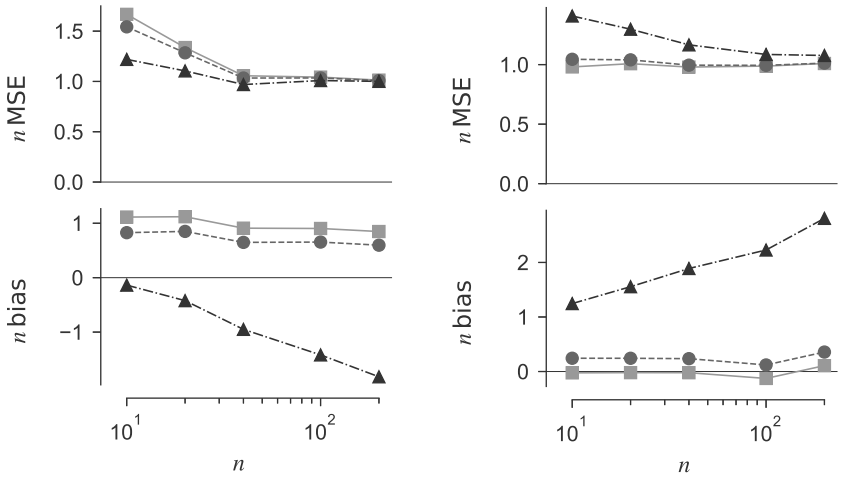
Smith R L (1985) Maximum likelihood estimation in a class of nonregular cases. Biometrika 72(1):67–90

Young L C (2019) Orthogonal collocation revisited. Computer Methods in Applied Mechanics and Engineering 345: 1033–1076

Table 1: Comparison of the bias convergence rate and MSE patterns of estimators

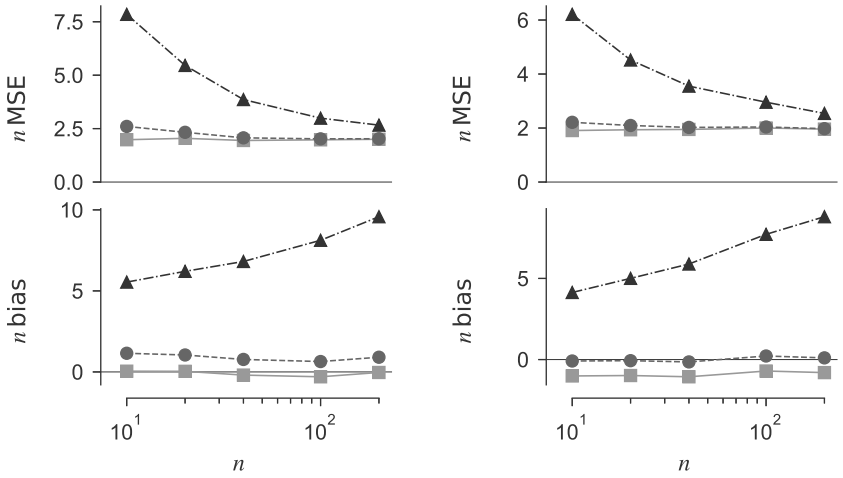
Distribution	cdf	Exponential family	Unknown params.	Investigated param.	Zero-BC	Bias convergence rate			MSE patterns[1]
						ML	JMMPS	MPS	
Exponential	$1-\exp(-\lambda x)$	Yes	λ	λ	Yes	$O(1/n)$	$O(1/n)$	$O((\log n)/n)$	A
Exponential	$1-\exp(-1/\eta)$	Yes	η	η	Yes	Unbiased	$O(1/n)$	$O((\log n)/n)$	B
Normal	$\frac{1}{2}\left\{1+\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right\}$	Yes	σ^2	σ^2	Yes	Unbiased	$O(1/n)$	$O((\log n)/n)$	B
Normal			μ,σ^2	μ	Yes	Unbiased	Unbiased	Unbiased	O
				σ^2	Yes	$O(1/n)$	Unbiased	$O((\log n)/n)$	B
Pareto	$1-(\eta/x)^\alpha$	Yes	η,α	η	No	$O(1/n)$	$O(1/n)$	Unbiased	C
				α	Yes	$O(1/n)$	$O(1/n)$	$O((\log n)/n)$	A
Gamma	$\frac{1}{\Gamma(\alpha)}\int_0^{x/\eta} t^{\alpha-1}e^{-t}\,dt$	Yes	η,α	η	Yes	$O(1/n)$	Unbiased	$O((\log n)/n)$	B
				α	Yes	$O(1/n)$	$O(1/n)$	$O((\log n)/n)$	A
Beta	$\frac{1}{B(\alpha,\beta)}\int_0^{x} t^{\alpha-1}(1-t)^{\beta-1}\,dt$	Yes	α,β	α	Yes	$O(1/n)$	$O(1/n)$	$O((\log n)/n)$	A
				β	Yes	$O(1/n)$	$O(1/n)$	$O((\log n)/n)$	A
Uniform	$a+x(b-a)$	No	a,b	a	No	$O(1/n)$	$O(1/n)$	Unbiased	C
				b	No	$O(1/n)$	$O(1/n)$	Unbiased	C
Cauchy	$\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\eta}\right)$	No	μ,η	μ	Yes	Unbiased	Unbiased	Unbiased	O
				η	Yes	Unbiased	Unbiased	$O(1/n)$	B

1: Pattern A, like Figure 1(a); pattern B, like Figure 1(b); pattern C, like Figure 3(a); pattern O, no difference among the methods.



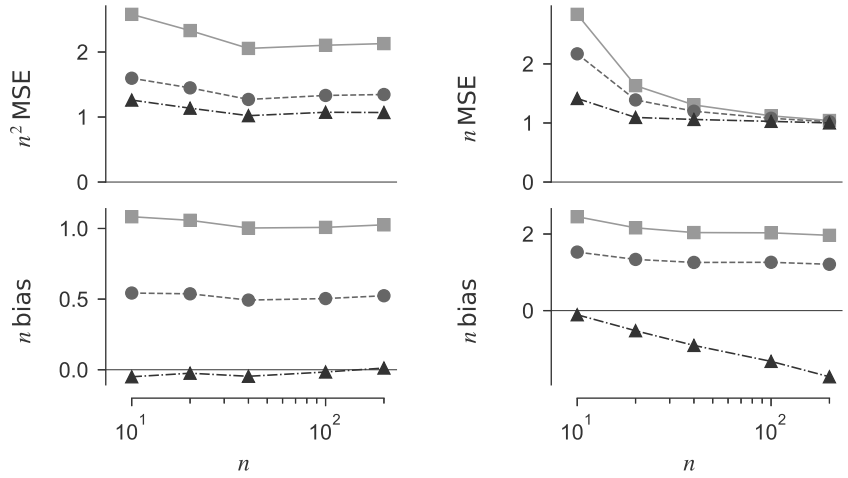
(a) $\hat{\lambda}$ for $\lambda_0 = 1$ (b) $\hat{\eta}$ for $\eta_0 = 1$

Fig. 1: Exponential distribution: simulated $n$bias and nMSE. Comparison of parameterization with (a) the rate parameter λ and (b) the scale parameter η; symbols are squares: ML, circles: JMMPS, triangles: MPS



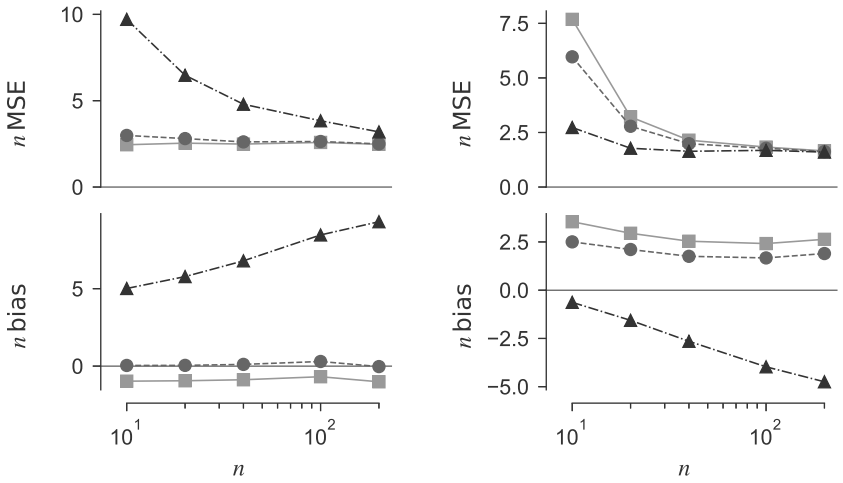
(a) $\hat{\sigma^2}$ for $\sigma_0^2 = 1$ ($\mu_0 = 0$ known) (b) $\hat{\sigma^2}$ for $(\mu_0, \sigma_0^2) = (0, 1)$

Fig. 2: Normal distribution: simulated $n$bias and nMSE of the variance σ^2. Comparison of the cases (a) when μ_0 is known and σ_0^2 unknown, and (b) when μ_0 and σ_0^2 are unknown; symbols are squares: ML, circles: JMMPS, triangles: MPS



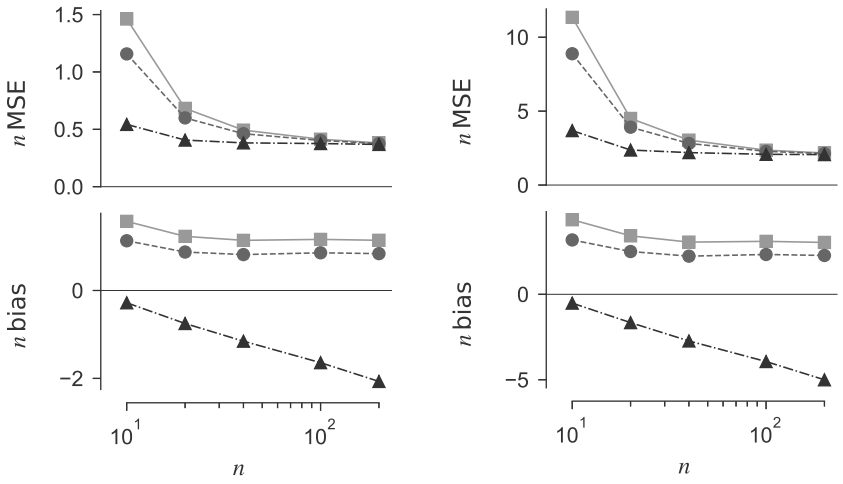
(a) $\hat{\eta}$ for $(\eta_0, \alpha_0) = (1,1)$ (b) $\hat{\alpha}$ for $(\eta_0, \alpha_0) = (1,1)$

Fig. 3: Pareto distribution: (a) $n$bias and n^2MSE of scale parameter η, and (b) $n$bias and nMSE of shape parameter α; symbols are squares: ML, circles: JMMPS, triangles: MPS



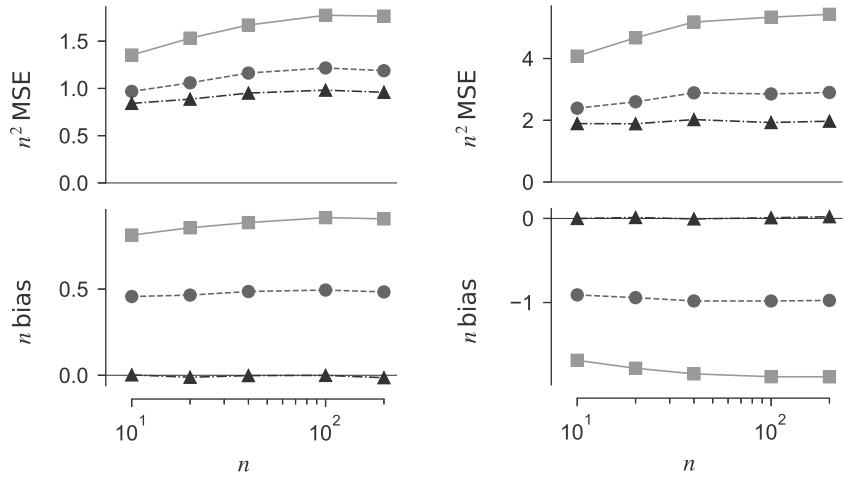
(a) $\hat{\eta}$ for $(\eta_0, \alpha_0) = (1,1)$ (b) $\hat{\alpha}$ for $(\eta_0, \alpha_0) = (1,1)$

Fig. 4: Gamma distribution: $n$bias, and nMSE of (a) scale parameter η, and (b) shape parameter α; symbols are squares: ML, circles: JMMPS, triangles: MPS



(a) $\hat{\alpha}$ for $(\alpha_0, \beta_0) = (0.5, 1)$ (b) $\hat{\beta}$ for $(\alpha_0, \beta_0) = (0.5, 1)$

Fig. 5: Beta distribution: $n$bias and nMSE of the shape parameters (a) α and (b) β; symbols are squares: ML, circles: JMMPS, triangles: MPS



(a) $\hat{a}$ for $(a_0, b_0) = (0, 1)$ (b) $\hat{b}$ for $(a_0, b_0) = (0, 1)$

Fig. 6: Uniform distribution: $n$bias and n^2MSE of (a) lower a and (b) upper b parameters; symbols are squares: ML, circles: JMMPS, triangles: MPS

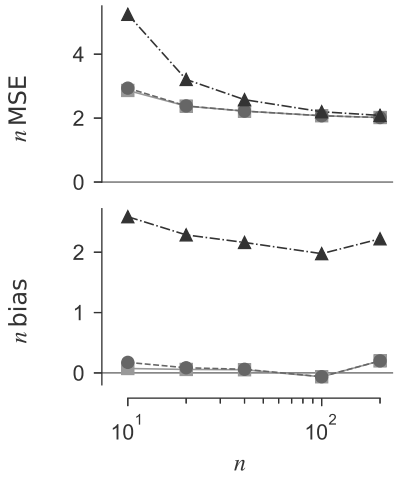


Fig. 7: Cauchy distribution: $n$bias, and nMSE of the scale parameter η $\hat{\eta}$ for $(\mu_0, \eta_0) = (1, 1)$; symbols are squares: ML, circles: JMMPS, triangles: MPS