Received: 1 July 2022 | Revised: 30 May 2023 | Accepted: 30 October 2023

DOI: 10.1111/sjos.12696

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Accurate bias estimation with applications to focused model selection

Ingrid Dæhlen | **Nils Lid Hjort** | **Ingrid Hobæk Haff**

Department of mathematics, University of Oslo, Oslo, Norway

Correspondence
Ingrid Dæhlen, Department of mathematics, University of Oslo, Oslo, Norway.
Email: ingrdae@math.uio.no

Funding information
Norges Forskningsråd, Grant/Award Number: 237718

Abstract

We derive approximations to the bias and squared bias with errors of order $o(1/n)$ where n is the sample size. Our results hold for a large class of estimators, including quantiles, transformations of unbiased estimators, maximum likelihood estimators in (possibly) incorrectly specified models, and functions thereof. Furthermore, we use the approximations to derive estimators of the mean squared error (MSE) which are correct to order $o(1/n)$. Since the variance of many estimators is of order $O(1/n)$, this level of precision is needed for the MSE estimator to properly take the variance into account. We also formulate a new focused information criterion (FIC) for model selection based on the estimators of the squared bias. Lastly, we illustrate the methods on data containing the number of battle deaths in all major inter-state wars between 1823 and the present day. The application illustrates the potentially large impact of using a less-accurate estimator of the squared bias.

KEYWORDS

asymptotic theory, bias estimation, focused information criterion, misspecified models, model selection, MSE estimation

1 | INTRODUCTION

Bias estimation has a long tradition and an extensive list of techniques. Cox and Snell (1968) give expressions of the bias of maximum likelihood estimators under the assumption of a correctly specified model. The jackknife, a way of estimating bias based on resampling, was introduced in Quenouille (1949), with important extensions in Quenouille (1956) and Tukey (1958). Firth (1993) takes a different approach and proposes techniques for reducing the bias, and hence removing the need for its estimation. In the present article, we build on classic theory by deriving approximations to the bias of a large class of estimators. Our goal will be to give comprehensive formulas throughout the article, and special attention should be given to Section 2.3. Here we derive model-robust expressions for the bias of maximum likelihood estimators. We believe the main theorem of this section serves as a modern extension of the results in Cox and Snell (1968) and is a novel addition to the field of bias estimation and correction.

Our main motivation for deriving precise bias approximations is estimation of mean squared error, or MSE for short. This risk function measures the expected squared L^2-distance from an estimator to the value it is aiming for. In symbols, this means $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta_0)^2$, where $\hat{\theta}$ is an estimator of a parameter θ_0. Famously, the MSE decomposes into the sum of the squared bias and the variance of the estimator evaluated. For many popular and well used estimators, for example, means, medians, and maximum likelihood estimators, the variance component is relatively easy to estimate. The bias component of the MSE, on the other hand, is often more complicated as the true value θ_0 shows up in the expression. Furthermore, the variance of many estimators, for example, quantiles, maximum likelihood estimators, and means, is of order $O(1/n)$, where n is the sample size. Because, of this, estimators of the squared bias must have expected errors of order $o(1/n)$ for the variability to be given sufficient weight in the estimated MSE. To achieve this, the precise bias approximations is crucial.

Recently, estimation of bias and MSE has been developed in great detail within the field of model selection. This is mainly due to the information criterion introduced in Claeskens and Hjort (2003) and later extended by multiple authors. See for instance Zhang and Liang (2011), Claeskens et al. (2006), or Claeskens and Hjort (2008). The criterion is called the focused information criterion, or FIC for short, and ranks models by the quality of their estimate of a prespecified parameter of interest. Quality is measured by MSE, and as a consequence, the formulation of the FIC requires sophisticated estimators of the MSE and squared bias. Because of this, many advanced techniques for estimation of the squared bias have been developed in the literature concerning the FIC. The original versions of the criterion were created for situations where models are nested within a true parametric alternative, with the degree of misspecification decreasing by the order of $O(1/\sqrt{n})$. In Jullum and Hjort (2017), however, a new version of the criterion is given, which gives a general way of estimating the MSE of estimators.

The approximation given in Jullum and Hjort (2017) avoids the assumption of asymptotically correctly specified models. Because of this, the estimators in this article have found use in multiple situations, see for example, Jullum and Hjort (2019), Ko et al. (2019), Claeskens et al. (2019), and Cunen, Walløe, and Hjort (2020). The approximations can, however, be shown to have an error term of order $O(1/n)$ in cases where some or all of the quantities involved are asymptotically biased. This makes the criterion too imprecise to give proper weight to the variance component of the MSE. As a result the variability of estimators can potentially be greatly downplayed in the estimator of the MSE developed in Jullum and Hjort (2017). In the present article, we will build on

the estimators given in Jullum and Hjort (2017) and derive an approximation to the MSE which has expected error order $o(1/n)$.

We will start by considering precise bias estimation for a wide class of estimators. Afterwards, we will discuss MSE estimation and the FIC. In particular, we will consider when and why the expressions given in Jullum and Hjort (2017) need to be corrected. In Section 4.1, we will give a new version of the FIC, extending that of Jullum and Hjort (2017). Lastly, we will illustrate the techniques developed in this article by estimating the MSE of a selection of estimators of the difference in the median number of battle deaths in major inter-state wars before and after the Korean war.

2 | PRECISE BIAS ESTIMATION

Our main topic is precise estimation of the bias. We will work with estimators on the form

$$\hat{\mu} = \mu + \frac{1}{n}\sum_{i=1}^{n} \nu(Y_i) + \epsilon_n. \tag{1}$$

where $Y_1,\ldots,Y_n$ are i.i.d. from a distribution F and $\mathrm{E}\nu(Y) = 0$ when $Y \sim F$. At first glance, considering estimators on the form of (1) only might look slightly restrictive. In practice, however, this condition holds for many commonly used estimators. If, for instance, $\hat{\mu}_1$ is a mean, the above equation holds trivially with $\nu_1(y) = y - \mu_1$. For the p quantile μ_p, $\nu_1(y) = [p - I(y \leq \mu_p)]/f(\mu_p)$ can be used when f, the density in the distribution F, exists. When $\hat{\mu}_1$ is a maximum likelihood estimator of μ in some parametric model f_θ, (1) holds with $\nu_1(y) = \nabla g(\theta_{\mathrm{lf}})^T J(\theta_{\mathrm{lf}})^{-1} u(y, \theta_{\mathrm{lf}})$ where θ_{lf} is the minimizer of the Kullback–Leibler divergence from the true distribution to the parametric family, u is the score function, $J(\theta_{\mathrm{lf}})$ is the Fisher matrix in the model evaluated at θ_{lf} and g is a function mapping θ to the value of μ in the parametric model, see Section 2.3 for details. In addition, expressions for ν_1 when $\hat{\mu}_1$ is a smooth function of the above estimators, can be derived by the delta method. For conditions ensuring that (1) is satisfied for the influence function, see for example, Thm. 5.5 in Shao (2003).

The main goal of this section is to find expressions for and estimators of

$$c = \lim_{n\to\infty} n\,\mathrm{E}(\hat{\mu} - \mu) = \lim_{n\to\infty} n\,\mathrm{E}\epsilon_n.$$

Before we can begin with estimation of the quantity, however, we need to discuss when $n\mathrm{E}\epsilon_n$ converges and c is well-defined. We start with an informal argument which hopefully will provide some intuition about what c is really is and when it exists.

Let F_n be the empirical distribution function of $Y_1,\ldots,Y_n \in \mathbb{R}$ and F the true cumulative distribution function. Assume that $\hat{\mu}$ can be represented as a functional of F_n, that is, that there exists a function T such that $\hat{\mu} = T(F_n)$. If T is sufficiently smooth, a Taylor expansion around F reveals

$$T(F_n) - T(F) = D^{(1)}_{F-F_n} T(F) + (1/2) D^{(2)}_{F-F_n} T(F) + O(\|F_n - F\|^3_\infty), \tag{2}$$

where $D^{(k)}_{F-F_n} T$ is the kth order directional derivative of T in direction $F_n - F$. The expression above is called the von Mises expansion, see for example, chap. 20 of Van der Vaart (1998) for more details.

The Donsker theorem, see for example Van der Vaart (1998, p. 266), ensures $\|F_n - F\|_\infty = O_{pr}(1/\sqrt{n})$. Hence, (2) implies

$$T(F_n) - T(F) = D^{(1)}_{F-F_n} T(F) + (1/2) D^{(2)}_{F-F_n} T(F) + o_{pr}(1/n).$$

The first term on the right-hand side is a sum of influence functions of T and has expectation zero. Under sufficient regularity, we therefore have

$$\mathrm{E}\{n[T(F_n) - T(F)]\} = (n/2)\mathrm{E} D^2 T_{F-F_n}(F) + o(1). \tag{3}$$

Since $D^2 T_{F-F_n}(F)$ is the quadratic term in a Taylor expansion, we would expect it to be of order $O(||F_n - F||^2) = O_{pr}(1/n)$, ensuring that the expression in (3) converges and c exists. A formal statement, proof and precise set of conditions showing that this intuition is indeed correct is given in Thm. 3.1 of Shao (1991).

In principle, we can always estimate c by approximating the right-hand side of (3). This method is quite general and can, in theory, be used for most estimators. In practice, however, this procedure is far from straightforward and, when done in full generality, results in incomprehensible formulas requiring fine-tuning and extensive calculations in each new application. We will therefore instead focus on some situations where precise formulas and/or estimation strategies for c can be found.

We will present four ways to estimate c. First, we will work with functions of unbiased estimators and quantiles. In Section 2.3 we will work with higher-order Taylor expansions of the log-likelihood function to give expressions for the bias of maximum likelihood estimators. Afterwards, we will discuss how c can be estimated when $\hat{\mu}$ is a function of estimators for which approximations to the bias exists. Lastly, we will discuss resampling techniques.

Remark 1. In the following, we will derive multiple approximations to c. Most of the formulas will be obtained by finding estimators with sufficiently small remainder terms, δ_n. Typically these will be of size $o_{pr}(1/n)$, and we will take this to imply that their expected values are of order $o(1/n)$. For this to hold true, a sufficient condition is that $\{n\delta_n\}_n$ is uniformly integrable. See, for example, Billingsley (1999, p. 31) for a proof and details concerning uniform integrability or Shao and Wu (1989) for an alternative set of regularity conditions. To make arguments more transparent, we will not focus on these technical details in the proofs, but take $\mathrm{E}\, o_{pr}(1/n) = o(1/n)$ for granted. All required regularity conditions will, however, be stated in the theorems and full proofs dealing with all technicalities may be found in the Appendix B.

2.1 | Functions of unbiased estimators

The first situation we will consider is when $\hat{\mu}$ is a function of unbiased estimators. The following result gives a formula for c in this case.

Theorem 1. *Let $\hat{a} \in \mathbb{R}^p$ be an unbiased estimator of $a \in \mathbb{R}^p$ such that $\sqrt{n}(\hat{a} - a)$ converges in distribution to a $N(0, \Sigma)$-distributed variable and all components of $\sqrt{n}(\hat{a} - a)$ to the power of three are uniformly integrable. Assume $h : \mathbb{R}^p \to \mathbb{R}$ is a function for which all partial derivatives up to and including order three exist and are continuous in*

a neighbourhood of a. Then, if $\mu = h(a)$ *and* $\hat{\mu} = h(\hat{a})$,

$$c + o(1) = n\mathrm{E}(\hat{\mu} - \mu) = (1/2)\mathrm{Tr}[Hh(a)\Sigma] + o(1), \tag{4}$$

where $Hh(a)$ *is the Hessian matrix of* h *evaluated at the point* a. *Furthermore,* c *can be estimated consistently by*

$$\hat{c} = (1/2)\mathrm{Tr}(Hh(\hat{a})\hat{\Sigma}), \tag{5}$$

where $\hat{\Sigma}$ *is some consistent estimator of* Σ.

Proof. Taylor expanding h around a_0 shows

$$\hat{\mu} - \mu = (\hat{a} - a)^T \nabla h(a) + (1/2)(\hat{a} - a)^T Hh(a)(\hat{a} - a) + \epsilon_n. \tag{6}$$

for some remainder term ϵ_n. Typically, ϵ_n will be of order $O_{pr}(\|\hat{a} - a\|^3)$ which ensures $\mathrm{E}\epsilon_n = O(n^{-3/2})$ under uniform integrability. Under the conditions of the theorem, this is indeed the case. Consult the Appendix B for details. Hence,

$$\mathrm{E}(\hat{\mu} - \mu) = 0 + (1/2)\mathrm{E}[(\hat{a} - a)^T Hh(a)(\hat{a} - a)] + o(1/n).$$

As is shown in the Appendix B, the conditions in the theorem ensures that $\sqrt{n}(\hat{a} - a) \xrightarrow{d} \mathrm{N}(0, \Sigma)$ implies $n\mathrm{Var}\hat{a} \to \Sigma$. Hence, by properties of the trace operator, (4) holds true.

Lastly, since $\hat{a}$ and $\hat{\Sigma}$ are consistent estimators of Σ and a, respectively, $\hat{c}$ defined in (5) converges in probability to c by the continuous mapping theorem. This concludes the proof. ∎

The simplest examples of functions of unbiased estimators are unbiased estimators themselves. For such estimators h is linear and its Hessian zero. Hence, $c = 0$ and needs not be estimated. Example include linear function of means and ordinary least squares regression coefficients.

Another important example is maximum likelihood estimators in exponential families. To see this, let f be a member of the exponential family. Then f is on the form

$$f(x) = A(x)B(\theta)\exp[\omega(\theta)^T T(x)], \tag{7}$$

for some functions A, B, ω, and T and a parameter θ. The maximum likelihood estimator based on a sample $Y_1, \ldots, Y_n$, is the maximizer of the log-likelihood function,

$$\ell_n(\theta) = \sum_{i=1}^{n} \log A(Y_i) + n \log B(\theta) + \omega(\theta)^T \sum_{i=1}^{n} T(Y_i),$$

which is the solution to the following equation

$$0 = \frac{nB'(\theta)}{B(\theta)} + \omega'(\theta)^T \sum_{i=1}^{n} T(Y_i).$$

Solving the equation shows

$$\hat{\theta} = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} T(Y_i)\right),$$

where g is the function

$$g(\theta) = -\omega'(\theta)^{-1}\frac{\partial}{\partial\theta}\log B(\theta) = -\omega'(\theta)^{-1}\frac{B'(\theta)}{B(\theta)}.$$

Now assume the focus parameter takes the value $k(\theta)$ in the family parameterized by (7) and let $\hat{\mu} = k(\hat{\theta})$ be the maximum likelihood estimator of this quantity. Then

$$\hat{\mu} = h[\frac{1}{n}\sum_{i=1}^{n} T(Y_i)],$$

where $h(x) = k[g^{-1}(x)]$. This shows that the maximum likelihood estimators in exponential families are functions of means. By Theorem 1, we get

$$c = (1/2)\mathrm{Tr}\{Hh[\mathrm{E}T(Y)]\mathrm{Var}[T(Y)]\}. \tag{8}$$

It is worth noting that (8) makes no assumptions about the parametric model being correctly specified. Hence, the expression in (8) holds true even when the true distribution of the data is not a member of the parametric family f_θ.

2.2 | Quantiles

We will now consider the case where $\hat{\mu}$ is a quantile. More specifically, for $Y_1,\ldots,Y_n \sim F$ i.i.d. and some $p \in (0,1)$, we will work with $(1-\gamma)Y_{(j)} + \gamma Y_{(j+1)}$, where $Y_{(k)}$ is the kth order statistic and j and γ are functions of p and n. All of the standard quantile estimators in `R` and `Python` are on this form, with j and γ given in Table 1.

We start by considering the case when $Y_1,\ldots,Y_n$ are uniformly distributed. The results are stated in the following lemma.

Lemma 1. *Let $U_1,\ldots,U_n$ be i.i.d. uniformly distributed variables and set $\hat{\mu} = (1-\gamma)U_{(j)} + \gamma U_{(j+1)}$ with γ and j defined by any of the rows in Table 1. Then, c takes the following form*

$$c + o(1) = n\mathrm{E}(\hat{\mu} - p) = n[(j+\gamma)/(n+1) - p], \tag{9}$$

as p is the p-quantile in the uniform distribution.

Proof. By standard results, the jth-order statistic follows a Beta$(j, n-j+1)$-distribution. The expression in (9) follows. ∎

In many cases, the left-hand side of (9) converges as n grows to infinity. For the quantile versions in `R` with the `type` argument equal to 4, 5,…, 9, this is the case, and the corresponding limits

are shown in Table 1. When `type` is equal to 1, 2, or 3, however, the expression to the left in (9) does not converge unless very specific conditions on how n changes from iteration to iteration is assumed. In applications, we recommend reading the limit of (9) off Table 1 when using a quantile estimator corresponding to one in the default `quantile` function in `R` with `type` equal to 4,…, 9. When `type` equal to 1, 2, or 3 is used, the left-hand side of (9) can be estimated directly using j and γ from Table 1. The latter approach can, of course, also be used when `type` equals 4,…, 9, but using the limit in Table 1 leads to relatively simpler formulas.

We will now use Lemma 1 to find an expression for c for more general distributions than the uniform.

Theorem 2. *Let $Y_1,\dots,Y_n$ be i.i.d. from a distribution F with p-quantile μ for some fixed $p \in (0,1)$. Assume that F is invertible and that there exists a neighborhood around μ on which F is thrice continuously differentiable. Let f be the density in the distribution. Then, when $\hat{\mu} = (1-\gamma)Y_{(j)} + \gamma Y_{(j+1)}$, with j and γ defined by either of the rows in Table 1,*

$$c + o(1) = n\mathrm{E}(\hat{\mu} - \mu) = \frac{n}{f(\mu)}\left(\frac{j+\gamma}{n+1} - p\right) - \frac{f'(\mu)p(1-p)}{2f(\mu)^3} + O(n^{-1/2}). \tag{10}$$

provided $n^{3/2}(\hat{\mu}-\mu)^3$ is uniformly integrable. Furthermore, c can be estimated consistently by

$$\frac{n}{\widehat{f(\mu)}}\left(\frac{j+\gamma}{n+1} - p\right) - \frac{\widehat{f'(\mu)}p(1-p)}{2\widehat{f(\mu)}^3}, \tag{11}$$

where $\widehat{f(\mu)}$ and $\widehat{f'(\mu)}$ are some estimators going in probability to $f(\mu)$ and $f'(\mu)$, respectively.

Proof. We sketch the proof here. More details can be found in the Appendix B.

By the probability integral transform, $U_i = F(Y_i)$ is uniformly distributed. Furthermore, $Y_{(i)} = F^{-1}(U_{(i)})$ as F is increasing and preserves ordering of variables.

TABLE 1 Each row in the table describes the nonparametric quantile estimator $(1-\gamma)Y_{(j)} + \gamma Y_{(j+1)}$.

Type	j	γ	Limit
1	$\lfloor pn \rfloor$	$I(np \neq \lfloor pn \rfloor)$	—
2	$\lfloor pn \rfloor$	$(1/2)I(np = \lfloor pn \rfloor) + I(np \neq \lfloor pn \rfloor)$	—
3	$\lfloor pn - 1/2 \rfloor$	$I(np \neq \lfloor pn \rfloor$ or j even)	—
4	$\lfloor pn \rfloor$	$np - \lfloor pn \rfloor$	$-p$
5	$\lfloor pn + 1/2 \rfloor$	$np + 1/2 - \lfloor pn + 1/2 \rfloor$	$1/2 - p$
6	$\lfloor pn + p \rfloor$	$np + p - \lfloor pn + p \rfloor$	0
7	$\lfloor pn + 1 - p \rfloor$	$np + 1 - p - \lfloor pn + 1 - p \rfloor$	$1 - 2p$
8	$\lfloor pn + (p+1)/3 \rfloor$	$np + (p+1)/3 - \lfloor pn + (p+1)/3 \rfloor$	$1/3 - (2/3)p$
9	$\lfloor pn + p/4 \rfloor$	$np + p/4 - \lfloor pn + p/4 \rfloor$	$-(3/4)p$

Notes: The column *Type* gives the `type` argument that should be given to `quantile` in `R` to get the corresponding estimator. To the far right we have displayed c, the limits of (9) when they exist. The symbol $\lfloor \cdot \rfloor$ denotes truncation.

Hence,

$$\hat{\mu} = (1-\gamma)F^{-1}(U_{(j)}) + \gamma F^{-1}(U_{(j+1)}). \tag{12}$$

A Taylor expansion around p shows

$$F^{-1}(U_{(j)}) = \mu + (F^{-1})'(p)(U_{(j)} - p) + (1/2)(F^{-1})''(p)(U_{(j)} - p)^2 + \delta_n'. \tag{13}$$

By arguments that can be found in the Appendix B, $\delta_n' = O_{pr}[(U_{(j)} - p)^3]$. All the standard versions of quantile estimators, have j chosen in such a way that $U_{(j)} - p = O_{pr}(1/\sqrt{n})$. Hence, $\delta_n' = O_{pr}(n^{-3/2})$. A similar result holds for $F^{-1}(U_{(j+1)})$.

By the inverse function theorem $(F^{-1})'(p) = f(\mu)^{-1}$ and $(F^{-1})''(p) = -f'(\mu)f(\mu)^{-3}$. Combining this with (12) and (13), shows

$$\hat{\mu} = \mu + f(\mu)^{-1}(\hat{\mu}_U - p) - f'(\mu)f(\mu)^{-3}[(1-\gamma)(U_{(j)} - p)^2 + \gamma(U_{(j+1)} - p)^2] + \delta_n. \tag{14}$$

where $\delta_n = O_{pr}(n^{-3/2})$ and $\hat{\mu}_U = (1-\gamma)U_{(j)} + \gamma U_{(j+1)}$.

Direct computations reveal

$$n\mathrm{E}[(1-\gamma)(U_{(j)} - p)^2 + \gamma(U_{(j+1)} - p)^2] = p(1-p) + o(1).$$

Furthermore, the arguments in the Appendix B ensure that $\delta_n = o_{pr}(n^{-1})$ implies $\mathrm{E}(\delta_n) = o(n^{-1})$ under the present conditions. Plugging this into (14), shows

$$n\mathrm{E}(\hat{\mu} - \mu) = n\mathrm{E}(\hat{\mu}_U - p)/f(\mu) - f'(\mu)p(1-p)/2f(\mu)^3 + o(1).$$

Combining the above equation with Lemma 1, shows (10). By the continuous mapping theorem (11) is consistent for c. This concludes the proof ∎

If a parametric model is known, we can estimate $f(\mu)$ and $f'(\mu)$ by their maximum likelihood estimate. When working with for instance nested models, one may then fit larger models by maximum likelihood and use the fit to estimate $f(\mu)$ and $f'(\mu)$. Otherwise, kernel density estimation can be used. In the Appendix A, we discuss the latter approach and give suggestion for bandwidths minimizing the MSE of the estimates of $f(\mu)$ and $f'(\mu)$.

2.3 | Maximum likelihood estimators

In some situations, we either know or at least strongly believe, that the true underlying distribution of the data belongs to some parametric family. When this is the case, the maximum likelihood estimator of the focus parameter will be both consistent and asymptotically unbiased for the true value. Hence, it can be used as the consistent estimator, $\hat{\mu}_0$, from Section 3.2. Because of this, we need expressions for c when $\hat{\mu}$ is a maximum likelihood estimator. Such expressions exist already. Cox and Snell (1968) derived formulas for the limit of $n\mathrm{E}(\hat{\theta} - \theta_0)$ when $\hat{\theta}$ is a maximum likelihood estimator in a correctly specified model and θ_0 is the true value of this parameter. In most situations, however, we do not know which, if any, parametric model the distribution of the data belongs to. Using the results of Cox and Snell (1968) will therefore be too optimistic. In this

section, we will derive model robust expressions for c. These can be applied even when we do not believe that the chosen model is the true one.

Let $Y_1,\ldots,Y_n \in \mathbb{R}^d$ be i.i.d. random variables from some underlying distribution with density or probability mass function, g, and assume we fit a parametric model, f_θ, to these data using maximum likelihood. We will let $\hat{\theta} \in \mathbb{R}^p$ denote the maximum likelihood estimator of θ, that is, the maximizer of $\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(Y_i)$. Under the assumptions stated in White (1982), $\hat{\theta}$ converges in probability to the minimizer of the Kullback-Leibler divergence from g to f_θ,

$$\mathrm{KL}(g, f_\theta) = \mathrm{E}_g[\log g(Y) - \log f_\theta(Y)].$$

The minimizer of this expression is in some sense, the "least false" parameter and is the quantity we have denoted by θ_{lf}. By White (1982), we also know that the speed of convergence is $O_{pr}(1/\sqrt{n})$ and that

$$\sqrt{n}(\hat{\theta} - \theta_{\mathrm{lf}}) \xrightarrow{d} N(0, J^{-1}KJ^{-1}),$$

where

$$J = \mathrm{E}_g\left(-\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta=\theta_{\mathrm{lf}}} \log f_\theta(Y)\right) \quad \text{and} \quad K = \mathrm{Var}_g\left(\frac{\partial}{\partial\theta}\Big|_{\theta=\theta_{\mathrm{lf}}} \log f_\theta(Y)\right). \tag{15}$$

In the above, g is added as a subscript to emphasize that we are taking expectations with respect to this distribution.

The normal limit of $\sqrt{n}(\hat{\theta} - \theta_{\mathrm{lf}})$ is proved by finding the root of the one term Taylor expansion of $\nabla\ell_n$ and showing that this is only $o_{pr}(1/\sqrt{n})$ away from $\hat{\theta}$. This allows us to study the behavior of an explicit expression rather than the, in general, only implicitly defined $\hat{\theta}$. For our purposes, however, a linear approximation is not sufficient, as we want to estimate the bias with an error of order $o(1/n)$ and therefore need higher precision than $o_{pr}(1/\sqrt{n})$. This is the main idea in the proof of the following proposition.

Theorem 3. *Let $Y_1,\ldots,Y_n \in \mathbb{R}^d$ be i.i.d. from a distribution F and let f_θ be some parametric family of densities or probability point mass functions indexed by an open set $\Theta \subseteq \mathbb{R}^p$. Furthermore, let $\hat{\theta}$ be the maximum likelihood estimator when fitting this family to the data and θ_{lf} the minimizer of the Kullback–Leibler divergence from F to f_θ. Then, the bias of $\hat{\theta}$ is given by the following formula*,

$$c + o(1) = n\mathrm{E}(\hat{\theta} - \theta_{\mathrm{lf}}) = J^{-1}\left[\begin{pmatrix}\mathrm{Tr}(J^{-1}V_1)\\ \vdots \\ \mathrm{Tr}(J^{-1}V_p)\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\mathrm{Tr}(W_1J^{-1}KJ^{-1})\\ \vdots \\ \mathrm{Tr}(W_pJ^{-1}KJ^{-1})\end{pmatrix}\right] + o(1), \tag{16}$$

where J and K are defined in (15) and

$$V_j = \mathrm{Cov}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_{\mathrm{lf}}} \log f_\theta(Y), \frac{\partial^2}{\partial\theta\partial\theta_j}\Big|_{\theta_{\mathrm{lf}}} \log f_\theta(Y)\right) \text{ and } W_j = \mathrm{E}\left(\frac{\partial^3}{\partial\theta\partial\theta^T\partial\theta_j}\Big|_{\theta_{\mathrm{lf}}} \log f_\theta(Y)\right)$$
$$\text{for } j = 1,\ldots,p, \tag{17}$$

provided the following four regularity conditions hold true:

- *(A1) All partial derivatives up to and including order four of* $\log f_\theta(y)$ *at* θ_{lf} *exists and are continuous for F-almost all y. In addition, with* $Y \sim F$ *we need all fourth order moments of* $\nabla_{\theta_{\text{lf}}} \log f_\theta(Y)$ *and all second order moments of all second- and third-order partial derivatives of* $\log f_\theta(Y)$ *to exist and be finite.*
- *(A2) The matrix J is positive definite.*
- *(A3) All fourth-order powers of all components of* $\sqrt{n}[\hat{\theta} - \theta_{\text{lf}} - (nJ)^{-1}\nabla \ell_n(\theta_{\text{lf}})]$ *are uniformly integrable.*
- *(A4) There exists a function* $m : \mathbb{R}^d \to \mathbb{R}$ *such that in a neighborhood of* θ_{lf} *all fourth-order partial derivatives of* $\log f_\theta(y)$ *are bounded by* $m(y)$ *for F-almost all y, and* $\text{E}m(Y)^4$ *exists and is finite.*

In addition, c can be estimated consistently by replacing the matrices J and K and V_j *and* W_j *for* $j = 1,\ldots,p$ *by the empirical means, covariances, and variances of the corresponding functions evaluated at* $\theta = \hat{\theta}$.

Proof. As mentioned previously, we will only sketch the proof of Theorem 3. For full arguments taking care of all details, consult the Appendix B.

Fix j and let η_j denote the jth component function of $\nabla \ell_n$, that is, $\partial \ell_n / \partial \theta_j$. Taylor expanding the function around θ_{lf}, shows

$$\eta_j(\theta) = \eta_j(\theta_{\text{lf}}) + (\theta - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}}) + (1/2)(\theta - \theta_{\text{lf}})^T H\eta_j(\theta_{\text{lf}})(\theta - \theta_{\text{lf}}) + \epsilon_n(\theta).$$

If η_j is sufficiently regular, we have $\epsilon_n(\theta) = O(n\|\theta - \theta_{\text{lf}}\|^3)$. At $\hat{\theta}$ it therefore holds that

$$0 = \eta_j(\theta_{\text{lf}}) + (\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}}) + (1/2)(\hat{\theta} - \theta_{\text{lf}})^T H\eta_j(\theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}}) + O_{pr}(1/\sqrt{n}), \tag{18}$$

since $\hat{\theta} - \theta_{\text{lf}} = O_{pr}(1/\sqrt{n})$. Here we have used that the $\hat{\theta}$ maximizing ℓ_n satisfies $\eta_j(\hat{\theta}) = 0$.

Since θ_{lf} minimizes the Kullback–Leibler divergence, the derivative of $\theta \mapsto \text{E} \log f_\theta(Y)$ at θ_{lf} is zero. Hence,

$$\text{E}\eta_j(\theta_{\text{lf}}) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_{\text{lf}}} \text{E}\ \log f_\theta(Y_i) = 0.$$

Because of this, taking the expectation of both sides of (18) shows

$$\text{E}[(\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}})] + (1/2)\text{E}[(\hat{\theta} - \theta_{\text{lf}})^T H\eta_j(\theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})] + O(1/\sqrt{n}) = 0. \tag{19}$$

We will work with each term in (19) separately, and use the result to find an expression for $\text{E}(\hat{\theta} - \theta_{\text{lf}})$ with remainder term smaller than $o(1/n)$. In the following, let $b_n = \text{E}(\hat{\theta} - \theta_{\text{lf}})$.

We start with the first term. By definition of η_j and properties of the trace operator,

$$\text{E}[(\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}})] = \text{E}\left((\hat{\theta} - \theta_{\text{lf}})^T \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}})\right) = \text{TrE}\left((\hat{\theta} - \theta_{\text{lf}}) \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}})^T\right).$$

By definition of the covariance, this equals

$$\operatorname{Tr}\left[b_n \mathrm{E}\left(\frac{\partial^2 \ell_n}{\partial\theta_j\partial\theta}(\theta_{\mathrm{lf}})\right)^T\right] + \operatorname{TrCov}\left(\hat{\theta}-\theta_{\mathrm{lf}}, \frac{\partial^2\ell_n}{\partial\theta_j\partial\theta}(\theta_{\mathrm{lf}})\right)$$
$$= b_n^T \mathrm{E}\left(\frac{\partial^2\ell_n}{\partial\theta_j\theta}(\theta_{\mathrm{lf}})\right) + \operatorname{TrCov}\left(\hat{\theta}-\theta_{\mathrm{lf}}, \frac{\partial^2\ell_n}{\partial\theta_j\partial\theta}(\theta_{\mathrm{lf}})\right).$$

We recognize the vector $\mathrm{E}\partial^2\ell_n(\theta_{\mathrm{lf}})/\partial\theta_j\partial\theta$ as $-n$ times the jth column of the Fisher matrix J. Let this be denoted by J^j. Furthermore, White (1982) showed that

$$\hat{\theta}-\theta_{\mathrm{lf}} = J^{-1}n^{-1}\ell_n'(\theta_{\mathrm{lf}}) + o_{pr}(1/\sqrt{n}). \tag{20}$$

Hence,

$$\mathrm{E}[(\hat{\theta}-\theta_{\mathrm{lf}})^T\nabla\eta_j(\theta_{\mathrm{lf}})] = -nb_n^TJ^j + \frac{1}{n}\operatorname{Tr}\left[J^{-1}\operatorname{Cov}\left(\ell_n'(\theta_{\mathrm{lf}}), \frac{\partial^2\ell_n}{\partial\theta_j\theta}(\theta_{\mathrm{lf}})\right)\right] + o(1).$$

Let u be the score function and u_j its jth component function. Then, since the Y_is are independent and identically distributed,

$$\operatorname{Cov}\left(\ell_n'(\theta_{\mathrm{lf}}), \frac{\partial^2\ell_n}{\partial\theta_j\theta}(\theta_{\mathrm{lf}})\right) = \sum_{i=1}^n V_j = nV_j,$$

with V_j defined in (17). Putting all of our results together, shows that the first term of (19) is equal to

$$-nb_n^TJ^j + \operatorname{Tr}(J^{-1}V_j) + o(1). \tag{21}$$

The second term of (19) requires less effort. First notice that

$$(\hat{\theta}-\theta_{\mathrm{lf}})^TH\eta_j(\theta_{\mathrm{lf}})(\hat{\theta}-\theta_{\mathrm{lf}}) = \sqrt{n}(\hat{\theta}-\theta_{\mathrm{lf}})^Tn^{-1}H\eta_j(\theta_{\mathrm{lf}})\sqrt{n}(\hat{\theta}-\theta_{\mathrm{lf}}).$$

By White (1982), $\sqrt{n}(\hat{\theta}-\theta_{\mathrm{lf}}) \xrightarrow{d} U$ where $U \sim \mathrm{N}(0, J^{-1}KJ^{-1})$. Furthermore, by the law of large numbers $H\eta_j(\theta_{\mathrm{lf}})/n = W_j + o_{pr}(1)$, where W_j is defined in (17). Hence, $(\hat{\theta}-\theta_{\mathrm{lf}})^TH\eta_j(\theta_{\mathrm{lf}})(\hat{\theta}-\theta_{\mathrm{lf}}) \xrightarrow{d} U^TW_jU$. This implies

$$\mathrm{E}[(\hat{\theta}-\theta_{\mathrm{lf}})^TH\eta_j(\theta_{\mathrm{lf}})(\hat{\theta}-\theta_{\mathrm{lf}})] = \operatorname{Tr}(W_jJ^{-1}KJ^{-1}) + o(1). \tag{22}$$

Here we have again used properties of the trace operator multiple times.

We are now ready to complete the argument. Combining (19), (21), and (22) for $j = 1,\ldots,p$, shows

$$0 = -nb_n^TJ^j + \operatorname{Tr}(J^{-1}V_j) + \operatorname{Tr}(W_jJ^{-1}KJ^{-1}) + o(1),$$

for $j = 1,\ldots,p$. These p equations hold simultaneously if and only if

$$0 = -nJb_n + \begin{pmatrix}\operatorname{Tr}(J^{-1}V_1)\\ \vdots \\ \operatorname{Tr}(J^{-1}V_p)\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\operatorname{Tr}(W_1J^{-1}KJ^{-1})\\ \vdots \\ \operatorname{Tr}(W_pJ^{-1}KJ^{-1})\end{pmatrix} + o(1).$$

Rearranging the terms and multiplying the equation by $(nJ)^{-1}$, shows (16).

Consistency of the estimators defined in Theorem 3 is shown in the Appendix B. ∎

If one of the parameters of the model itself is chosen as focus parameter, Theorem 3 can be used to estimate c. In most situations, however, $\hat{\mu}$ is a function of $\hat{\theta}$, and the above expressions do not suffice. This is a special case of what we will discuss in Section 2.5. First, we will, however, give extensions of Theorem 3 to the more general case of M- and Z-estimators.

2.3.1 | M- and Z-estimators

The calculations made in the proof of Theorem 3 are not limited to maximum likelihood estimators. Similar arguments can be made to find the bias of a more general class of estimators: M- and Z-estimators. Examples of such estimators include copulas fitted using a two-stage procedure, see Joe (2005) and Ko et al. (2019), and robust M-estimators, see for example, Huber (2009).

M-estimators are maximizers of expressions on the form

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}\zeta(Y_i,\theta), \tag{23}$$

where $\zeta : \mathbb{R}^{d+p} \to \mathbb{R}$. The most obvious example of such estimators is maximum likelihood estimators. In this case $\zeta(y,\theta) = \log f_\theta(y)$. Because of their similarity, M-estimators share many properties with maximum likelihood estimators. It can for instance be shown that the maximizer of (23), $\hat{\theta}$, converges almost surely to θ_0, the maximizer of the limit function $\theta \mapsto \mathrm{E}\zeta(Y,\theta)$, and that $\sqrt{n}(\hat{\theta}-\theta_0)$ has a normal limit, or in particular, that

$$\sqrt{n}(\hat{\theta}-\theta_0) = -V^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\Big|_{\theta=\theta_0}\zeta(Y_i,\theta) + o_{pr}(1),$$

where V is the Hessian matrix of the limit function $\theta \mapsto \mathrm{E}\zeta(Y,\theta)$ at θ_0. This expression is an analogue to (20). For more details and proofs, consult Chap. 4 of Van der Vaart (1998).

Looking at the proof of Theorem 3 we notice that the arguments can be repeated for the more general class of M-estimators. This is achieved by replacing $\log f_\theta(y)$ by the function ζ and using the properties listed in the previous paragraph in place of the corresponding results for maximum likelihood estimators. The argument is very similar and will not be repeated here, but the result is stated in the following theorem.

Theorem 4. *Let $Y_1,\dots,Y_n \in \mathbb{R}^d$ be i.i.d. from a distribution F and $\hat{\theta} \in \mathbb{R}^p$ be an M-estimator maximizing $n^{-1}\sum_{i=1}^{n}\zeta(Y_i,\theta)$. Furthermore, let $\theta_0 \in \mathbb{R}$ be the maximizer of the limit function $\theta \mapsto \mathrm{E}\zeta(Y,\theta)$ where $Y \sim F$. Then the bias of $\hat{\theta}$ is given by the following formula,*

$$c + o(1) = n\mathrm{E}(\hat{\theta}-\theta_0) = J^{-1}\left[\begin{pmatrix}\mathrm{Tr}(J^{-1}V_1)\\ \vdots \\ \mathrm{Tr}(J^{-1}V_p)\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\mathrm{Tr}(W_1J^{-1}KJ^{-1})\\ \vdots \\ \mathrm{Tr}(W_pJ^{-1}KJ^{-1})\end{pmatrix}\right] + o(1),$$

where J and K are defined as

$$J = \mathrm{E}\left(-\frac{\partial^2\zeta}{\partial\theta\partial\theta^T}(Y,\theta_0)\right) \quad \text{and} \quad K = \mathrm{Var}\left(\frac{\partial\zeta}{\partial\theta}(Y,\theta_0)\right).$$

and

$$V_j = \mathrm{Cov}\left(\frac{\partial\zeta}{\partial\theta}(Y,\theta_0), \frac{\partial^2\zeta}{\partial\theta\partial\theta_j}(Y,\theta_0)\right) \quad \text{and} \quad W_j = \mathrm{E}\left(\frac{\partial^3\zeta}{\partial\theta\partial\theta^T\partial\theta_j}(Y,\theta_0)\right) \quad \text{for } j = 1,\ldots,p,$$

provided the regularity conditions (A1–A4) of Theorem 3 hold true after replacing $\log f$ *with* ζ, θ_{lf} *with* θ_0 *and the maximum likelihood estimator with the maximizer of* $n^{-1}\sum_{i=1}^n \zeta(Y_i,\theta)$.

In addition, c can be estimated consistently by replacing the matrices J and K and V_j *and* W_j *for* $j = 1,\ldots,p$ *by the empirical means, covariances, and variances of the corresponding functions evaluated at* $\theta = \hat{\theta}$.

Assume that the function ζ is differentiable in θ. Maximization of (23) can then be rephrased as solving

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial\zeta}{\partial\theta}(Y_i,\theta) = 0.$$

Solutions to equations like the one above, are called Z-estimators. In the proof of Theorem 3, it is the fact that the maximum likelihood estimator is a Z-estimator which is used in the arguments. Because of this, the result holds true also for this class of estimators, as stated in the following theorem. The proof is very similar to that of Theorem 3 and is left out.

Theorem 5. *Let* $Y_1,\ldots,Y_n \in \mathbb{R}^d$ *be i.i.d. from a distribution F and* $\hat{\theta} \in \mathbb{R}^p$ *be a Z-estimator solving* $0 = n^{-1}\sum_{i=1}^n \xi(Y_i,\theta)$. *Furthermore, let* $\theta_0 \in \mathbb{R}$ *be the solution to the limit equation* $0 = \mathrm{E}\zeta(Y,\theta)$ *where* $Y \sim F$. *Then the bias of* $\hat{\theta}$ *is given by the following formula,*

$$c + o(1) = n\mathrm{E}(\hat{\theta} - \theta_0) = J^{-1}\left[\begin{pmatrix}\mathrm{Tr}(J^{-1}V_1)\\ \vdots \\ \mathrm{Tr}(J^{-1}V_p)\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\mathrm{Tr}(W_1J^{-1}KJ^{-1})\\ \vdots \\ \mathrm{Tr}(W_pJ^{-1}KJ^{-1})\end{pmatrix}\right] + o(1),$$

where J and K are defined as

$$J = \mathrm{E}\left(-\frac{\partial\xi}{\partial\theta}(Y,\theta_0)\right) \quad \text{and} \quad K = \mathrm{Var}\xi(Y,\theta_0).$$

and

$$V_j = \mathrm{Cov}\left(\xi(Y,\theta_0), \frac{\partial\xi_j}{\partial\theta}(Y,\theta_0)\right) \quad \text{and} \quad W_j = \mathrm{E}\left(\frac{\partial^2\xi_j}{\partial\theta\partial\theta^T}(Y,\theta_0)\right) \quad \text{for } j = 1,\ldots,p,$$

provided the regularity conditions (A1–A4) of Theorem 3 hold true after replacing $\partial\log f/\partial\theta$ *with* ξ, θ_{lf} *with* θ_0 *and the maximum likelihood estimator with the solution to* $0 = n^{-1}\sum_{i=1}^n \zeta(Y_i,\theta)$. *In the above* ξ_j *is the jth component function of* ξ.

In addition, c can be estimated consistently by replacing the matrices J and K and V_j and W_j for $j = 1,\ldots,p$ by the empirical means, covariances and variances of the corresponding functions evaluated at $\theta = \hat{\theta}$.

Lastly, we remark that Theorem 4 and Theorem 5 hold true even when $\hat{\theta}$ is only "almost" an M- or a Z-estimator. In fact, it is sufficient that $\hat{\theta}$ is an estimator which can be expressed as the root of a function on the form

$$\theta \mapsto \sum_{i=1}^{n} \xi(Y_i, \theta) + \delta_n(\theta),$$

where $\delta_n(\theta) = o_{pr}(1)$ uniformly in a neighbourhood of the root of the limit function $\theta \mapsto \mathrm{E}\zeta(Y, \theta)$. This follows from the fact that $\delta_n(\theta)$ can be incorporated into the remainder term of (18).

2.4 | Resampling techniques

The last way of estimating c we will present is resampling techniques. We will discuss two methods: the bootstrap and the jackknife. The main ideas will be presented briefly here, but additional details, proofs, and discussions can be found in for example, Efron (1982), Efron and Tibshirani (1993), or Shao and Tu (1995).

The bootstrap was first introduced in Efron (1979), and has since become a popular statistical tool. To see how the procedure can be used to estimate c, notice that by (1)

$$\mathrm{E}[n(\hat{\mu} - \mu)] = n\mathrm{E}\epsilon_n = c + o(1).$$

The left-hand side of this equation can be easily estimated with bootstrap. Let $\hat{\mu}_j^*$ be the estimator of μ based on the jth out of B bootstrap samples. Then

$$\hat{c}_{\mathrm{boot}} = \frac{n}{B} \sum_{j=1}^{B} (\hat{\mu}_j^* - \hat{\mu}),$$

is an estimator of c, see for example, Chap. 10 of Efron and Tibshirani (1993).

The bootstrap procedure is quite general and works for most focus parameters. Hence, this technique is an option when all other methods fail. Using the procedure in practice is, however, often inconvenient, as a very large number of bootstrap samples must be drawn to achieve sufficient precision. To see this, notice that the bootstrap estimate $B^{-1}\sum_{j=1}^{B}(\hat{\mu}_j^* - \hat{\mu})$ is a Monte Carlo estimate of its expectation. Since the variance of $\hat{\mu}_j^*$ is of order $O(1/n)$ by (1), the central limit theorem ensures that the Monte Carlo error of $B^{-1}\sum_{j=1}^{B}(\hat{\mu}_j^* - \hat{\mu})$ is of order $O_{pr}(1/\sqrt{nB})$. Hence, the error of $\hat{c}_{\mathrm{boot}}$ is of order $O_{pr}(\sqrt{n}/\sqrt{B})$. For $\hat{c}_{\mathrm{boot}}$ to be consistent we would therefore need $B \gg n$. To achieve a convergence rate of $O_{pr}(1/\sqrt{n})$ for c we would need a million bootstrap samples when $n = 1000$, and even with a mere 50 data points, we would need B to be greater than 2000. Because of this, the computational burden might be too heavy for bootstrap to be practical in many situations.

The jackknife is another resampling technique. It was introduced in Quenouille (1949) and Tukey (1958), and was developed specifically for the purpose of estimating the bias with an error of order $o(1/n)$. It is therefore well suited for estimating c.

The jackknife estimate of the bias of $\hat{\mu}$ is defined as

$$\hat{c}_{\text{jack}} = (n-1)\left(\hat{\mu} - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{-i}\right).$$

Here $\hat{\mu}_{-i}$ is the same estimator of μ as $\hat{\mu}$, but based on $Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n$, for $i = 1,\ldots,n$ rather than the full dataset. More formally, we assume there is a functional T, such that $\hat{\mu} = T(F_n)$ and $\hat{\mu}_{-i} = T(F_n^{-i})$, where F_n^{-i} is the empirical distribution function based on all but the ith data point. Under certain smoothness conditions, $\hat{c}_{\text{jack}}$ is consistent for the true value c. This is shown in Thm. 3.1 of Shao (1991).

The jackknife procedure is a flexible and general tool for bias estimation. That being said, the method requires n applications of the functional T to estimate c. In cases where application of the functional is computationally expensive or n is very large, the jackknife estimate can be slow to calculate. To ease the computational burden, it is possible to use less computer intensive versions of the standard jackknife procedure. Options include the one-step jackknife and the grouped jackknife. See Chap. five of Shao and Tu (1995) for an overview. These methods limit the number of times T needs to be applied, reducing the overall computational cost. In certain situations, however, resampling techniques can be too time consuming to be practical, even with modifications like these. In such cases, the formulas derived in the previous sections can be used instead.

2.5 | Functions of the above

In the previous sections, we derived formulas and described strategies for estimating c in certain situations. We will now show how these results can be used to estimate the bias of even more complex estimators.

Theorem 6. *Let $\hat{\theta}$ be some estimator of θ for which c_θ, the limit of $n\mathrm{E}(\hat{\theta}-\theta)$ is known. Assume furthermore that $\sqrt{n}(\hat{\theta}-\theta)$ converges in distribution to a $N(0,\Sigma)$-distributed variable. and that $h : \mathbb{R}^p \to \mathbb{R}$ is a function for which all partial derivatives up to an including order three exists and are continuous in a neighborhood of θ. Then if $\mu = h(\theta)$ and $\hat{\mu} = h(\hat{\theta})$,*

$$c = \nabla h(\theta)^T c_\theta + (1/2)\mathrm{Tr}[Hh(\theta)\Sigma]. \tag{24}$$

Furthermore, if $\hat{c}_\theta$ is a consistent estimator of c_θ,

$$\nabla h(\hat{\theta})^T \hat{c}_\theta + (1/2)\mathrm{Tr}[Hh(\hat{\theta})\hat{\Sigma}], \tag{25}$$

converges in probability to c.

Notice the similarities with Theorem 1. When $\hat{\theta}$ is unbiased, the first term of (24) reduces to (4).

Proof. As before, we present a proof sketch only. A full argument is similar to that of Theorem 1. Consult the proof of this theorem in the Appendix B for more details.

A Taylor expansion reveals the following

$$\hat{\mu} = \mu + \nabla h(\theta)^T(\hat{\theta} - \theta) + (1/2)(\hat{\theta} - \theta)^T Hh(\theta)(\hat{\theta} - \theta) + o_{pr}(1/n).$$

Let $b_n = \mathrm{E}(\hat{\theta} - \theta)$. Taking expectations on both sides of the above equation, shows

$$n\mathrm{E}(\hat{\mu} - \mu) = \nabla h(\theta)^T n b_n + (n/2)\mathrm{E}[(\hat{\theta} - \theta)^T Hh(\theta)(\hat{\theta} - \theta)] + o(1),$$

which, by standard rules for the trace operator, is equivalent to $n\mathrm{E}(\hat{\mu} - \mu) = \nabla h(\theta)^T n b_n + (1/2)\mathrm{Tr}[Hh(\theta)\Sigma] + o(1)$. By assumption nb_n converges to c_θ, and hence,

$$c = \nabla h(\theta)^T c_\theta + (1/2)\mathrm{Tr}[Hh(\theta)\Sigma].$$

Since $\hat{c}_\theta$ is consistent for c_θ, (25) converges to c by the continuous mapping theorem. ∎

Now that we have discussed c in detail, we are ready to move onto our main application: estimation of the MSE and the FIC.

3 | ESTIMATING MSE

We will start by introducing notation and assumptions. In addition, we will derive equations that show explicitly the quantities needed to estimate the MSE.

3.1 | An approximation to the MSE

Assume we have i.i.d. data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ from a distribution, F. Let $\hat{\mu}_1$ be an estimator of some focus parameter with true value $\mu_0 \in \mathbb{R}$ based on these data, and assume $\hat{\mu}_1$ admits an influence function, ν_1, with a small enough remainder term, in the sense that

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n}\sum_{i=1}^{n} \nu_1(Y_i) + S_n, \tag{26}$$

with $\mathrm{E}\nu_1(Y) = 0$ and $S_n = O_{pr}(1/n)$. Lastly, let c_1 denote the limit of $n\mathrm{E}S_n$.

Direct computations show

$$\mathrm{MSE}(\hat{\mu}_1) = \mathrm{E}(\hat{\mu}_1 - \mu_0)^2 = b^2 + \frac{\tau}{n} + \frac{2bc_1}{n} + 2\mathrm{E}\left(\frac{S_n}{n}\sum_{i=1}^{n}\nu_1(Y_i)\right) + \mathrm{E}S_n^2 + o(1/n), \tag{27}$$

where $b = \mu_1 - \mu_0$ and τ is the variance of $\nu_1(Y)$. By the above assumptions $S_n = o_{pr}(1/\sqrt{n})$. Hence, $S_n^2 = o_{pr}(1/n)$. In addition, $n^{-1}\sum_{i=1}^{n}\nu_1(Y_i) = O_{pr}(1/\sqrt{n})$ by the central limit theorem. Because of this, $n^{-1}S_n\sum_{i=1}^{n}\nu_1(Y_i) = o_{pr}(1/n)$. If we further assume S_n^2 and $S_n n^{-1}\sum_{i=1}^{n}\nu_1(Y_i)$ to be uniformly integrable, the above is sufficient for the last two terms in (27) to be of order $o(1/n)$,

see for example, Billingsley (1999, p. 31). Hence, the MSE of $\hat{\mu}_1$ takes the following form

$$\text{MSE}(\hat{\mu}_1) = b^2 + \tau/n + 2bc_1/n + o(1/n).$$

We summarize our findings in a theorem for easier reference.

Theorem 7. *Assume $Y_1, \ldots, Y_n \in \mathbb{R}^d$ are i.i.d. from a distribution F, and $\hat{\mu}_1$ is an estimator of μ_0 satisfying*

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n}\sum_{i=1}^{n} \nu_1(Y_i) + S_n,$$

for some $S_n = O_{pr}(1/n)$. Furthermore, let $b = \mu_1 - \mu_0$, and take c_1 to be the limit of $n\text{E}S_n$. Then

$$\text{MSE}(\hat{\mu}_1) = b^2 + \tau/n + 2bc_1/n + o(1/n), \tag{28}$$

where τ is the variances of $\nu_1(Y)$, provided this quantity exists and is finite and nS_n is uniformly integrable.

To estimate the MSE of $\hat{\mu}_1$, we need to approximate all quantities in (28). This will be the topic of the next section.

3.2 | Estimating to the correct order

Looking at (28), we notice that the limiting bias, b, is constant while the remaining terms decrease at the speed of $O(1/n)$. Because of this, b will dominate when n gets large. Hence, minimizing the MSE asymptotically reduces to choosing a $\hat{\mu}_1$ with $b = 0$. In practice, however, we do not have infinite data points, and blindly choosing the estimator with the lowest limit bias is therefore rarely useful. Most applications require extremely flexible estimators to ensure that the bias disappears in the limit. Such procedures often have high variance. As a result, their MSEs tend to be high in finite samples. Because of this, many statistical applications instead attempt to find a trade-off between bias and variance. For our MSE estimates to be useful, they therefore need to take both the bias and variance into account. Since the latter is of order $O(1/n)$, we will need to create an approximation with an error of a smaller order. Otherwise, the variance is on the same scale as the error of the estimator, washing out the effect of the variability.

We will create an estimator of the MSE which is only $o(1/n)$ away from (28) in expected value. To do this, we need to estimate three quantities: the asymptotic bias b, the variance τ of $\nu_1(Y)$ and the limit c_1 of $n\text{E}S_n$. If we can find asymptotically unbiased estimators for these quantities, the two last terms in (28) can be estimated with an error of $o(1/n)$, due to the n appearing in the denominator of these terms. We do, however, need to be extra careful with the estimator of the squared bias. This quantity is not divided by n, and hence, we need to make sure that the bias of the estimator is of order $o(1/n)$.

To estimate the asymptotic bias, we need to approximate the distance from the limit of our estimator, μ_1, to the truth, μ_0. This is impossible to do without having some idea of the value of μ_0. Because of this, we need to introduce a second estimator which is consistent for this true value. We will call this estimator $\hat{\mu}_0$ and make similar assumptions about its form to the ones made for

$\hat{\mu}_1$, that is,

$$\hat{\mu}_0 = \mu_0 + \frac{1}{n}\sum_{i=1}^{n} \nu_0(Y_i) + T_n, \tag{29}$$

with T_n being an error term of order $O_{pr}(1/n)$ and $\nu_0 : \mathbb{R}^d \to \mathbb{R}$ a function such that $\mathrm{E}\nu_0(Y) = 0$.

The introduction of $\hat{\mu}_0$ might seem slightly counter-intuitive at first glance, and one might wonder why we would bother with $\hat{\mu}_1$ when we have a consistent estimator available. The answer to this concern is that even though $\hat{\mu}_1$ is inconsistent for μ_0, its MSE might still be lower than that of the consistent estimator $\hat{\mu}_0$, making in preferable for estimation. In addition, $\hat{\mu}_0$ does not need to satisfy all the conditions we require of $\hat{\mu}_1$. The most obvious example of this is when we wish to fit a parametric model to the data. In such cases we might "need" $\hat{\mu}_1$ to be a maximum likelihood estimator, while this condition is not required of $\hat{\mu}_0$. This is a crucial point when working with the FIC, which will be discuss in Section 4.

Since $\hat{\mu}_0$ is consistent for the true value μ_0, the asymptotic bias can be estimated consistently by the following quantity

$$\hat{b} = \hat{\mu}_1 - \hat{\mu}_0 = b + \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i) + R_n,$$

where $\phi = \nu_1 - \nu_0$ and $R_n = S_n - T_n$. Squaring the above expression shows

$$\hat{b}^2 = b^2 + \left(\frac{1}{n}\sum_{i=1}^{n} \phi(Y_i)\right)^2 + R_n^2 + 2\left(\frac{b + R_n}{n}\sum_{i=1}^{n} \phi(Y_i) + bR_n\right).$$

Taking expectations on both sides of the equation and arguing as in Section 3.1, reveals

$$\mathrm{E}\hat{b}^2 = b^2 + \kappa/n + 2b\mathrm{E}R_n + o(1/n),$$

under the assumption of uniform integrability of R_n^2 and $n^{-1}R_n\sum_{i=1}^{n}\phi(Y_i)$. In the above, $\kappa = Var\phi(Y_i)$. Since $R_n = S_n - T_n$, we further have

$$\mathrm{E}\hat{b}^2 = b^2 + 2bc_1/n + \kappa/n - 2b\mathrm{E}T_n + o(1/n).$$

Just as for S_n, $\mathrm{E}T_n$ will typically be of order $O(1/n)$. So let c_0 be the limit of $n\mathrm{E}T_n$. Then, the above equation implies

$$\mathrm{E}\hat{b}^2 = b^2 + 2bc_1/n + \kappa/n - 2bc_0/n + o(1/n).$$

We summarize the results in a theorem for easier reference.

Theorem 8. *Assume $Y_1,\ldots,Y_n \in \mathbb{R}^d$ are i.i.d. from a distribution F, and $\hat{\mu}_1$ and $\hat{\mu}_0$ are estimators satisfying*

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n}\sum_{i=1}^{n} \nu_1(Y_i) + S_n \quad \text{and} \quad \hat{\mu}_0 = \mu_0. + \frac{1}{n}\sum_{i=1}^{n} \nu_0(Y_i) + T_n.$$

for some $S_n, T_n = O_{pr}(1/n)$. Furthermore, let $\hat{b} = \hat{\mu}_1 - \hat{\mu}_0$ and $b = \mu_1 - \mu_0$, and take c_0 and c_1 to be the limits of $n\mathrm{E}T_n$ and $n\mathrm{E}S_n$. Then

$$\mathrm{E}\hat{b}^2 = b^2 + 2bc_1/n + \kappa/n - 2bc_0/n + o(1/n), \tag{30}$$

where κ and τ are the variances of $\phi(Y) = \nu_1(Y) - \nu_0(Y)$ and $\nu_1(Y)$, respectively, provided κ and τ exists and are finite and nS_n and nT_n are uniformly integrable.

The first two terms in (30) appear in (28), the asymptotic approximation we wish to estimate. The last two terms, however, do not and need to be corrected for. If we can find consistent estimators, $\hat{c}_0$ and $\hat{\kappa}$, for c_0 and κ, respectively, the following estimator is only $o(1/n)$ away from $b^2 + 2bc_1/n$ in expected value

$$\hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n. \tag{31}$$

The approximation $\hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n + \hat{\tau}/n$ of the MSE, therefore has an error of order $o(1/n)$ in expectation. Here $\hat{\kappa}$ and $\hat{\tau}$ are the empirical variances of $\phi(Y_1),\ldots,\phi(Y_n)$ and $\nu_1(Y_1),\ldots,\nu_1(Y_n)$, respectively.

Sometimes, (31) will be negative, leading to an estimate of the squared bias which less than zero. By definition, however, $b^2 \geq 0$. Negative estimates are therefore both counter-intuitive and always further away from the true value than 0. Hence, when negative estimates show up, we will truncate them to zero, as in Jullum and Hjort (2017). This leads to the following estimator of the MSE of $\hat{\mu}_1$

$$\widehat{\mathrm{MSE}}(\hat{\mu}_1) = \max\{0, \hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n\} + \hat{\tau}/n. \tag{32}$$

Estimators for b, κ, and τ have already been discussed. By definition c_0 is the limit of $n\mathrm{E}(\hat{\mu}_0 - \mu_0)$, and estimation of this quantity was discussed in length in Section 2.

Remark 2. The estimator in (32) is not the only way to estimate the MSE of $\hat{\mu}_1$. Rather than adding $2\hat{b}\hat{c}_0/n$ to $\hat{b}^2$ to remove the bias, we could subtract $\hat{c}_0/n$ from $\hat{\mu}_0$ before estimating b. This would correct the bias of $\hat{\mu}_0$ and remove the need to add $2\hat{b}\hat{c}_0/n$ to the squared bias estimate. Such a procedure is indeed an alternative to using (32), but does not lead to a significantly different estimator. By correcting the bias of $\hat{\mu}_0$ before using it to estimate the MSE of $\hat{\mu}_1$, we are left with the estimator $\hat{b}^2 + 2\hat{b}\hat{c}_0/n + \hat{c}_0^2/n^2 - \hat{\kappa}/n$ of the squared bias. The expected difference between this expression and (30) is a term of order $O(1/n^2)$. As we are attempting to create estimators with biases of order $o(1/n)$, the difference between the two approaches is negligible.

That being said, estimating and subtracting $\hat{c}_0$ is only one out of many ways to remove bias from $\hat{\mu}_0$. Modifying the score function as in Firth (1993), removes the $O(1/n)$ part of the bias of the maximum likelihood estimator. For M-estimators, similar alterations can be made to the moment equations to achieve approximate unbiasedness, see Kim (2016). For situations where such techniques are available, correcting the bias of $\hat{\mu}_0$ before using it to estimate b is indeed an alternative to the methods suggested in this article.

4 | THE FOCUSED INFORMATION CRITERION

As mentioned previously, estimation of the MSE has been developed recently in the literature concerning the FIC. This criterion attempts to choose the model with the most precise estimate of a prespecified focus parameter, μ, rather than the one fitting the data the best overall. This is achieved by ranking candidate models by the MSE of their estimate of μ. The estimator in (32) can therefore be seen, not only as a way to approximate MSEs, but indeed as an information criterion.

Assume we fit a parametric model f_θ using maximum likelihood. The resulting estimator $\hat{\theta}$ of θ then satisfies the following equation

$$\hat{\theta} = \theta + J(\theta_{\text{lf}})^{-1}\frac{1}{n}\sum_{i=1}^{n} u(Y_i, \theta_{\text{lf}}) + O_{pr}(1/n),$$

where u is the score function and θ_{lf} and $J(\theta_{\text{lf}})$ are as defined in Section 3.1. See for example, Thm. 5.39 in Van der Vaart (1998) for a proof and a set of conditions. Applying the delta method ensures that the focus parameter satisfies (26) with $\nu_1(y) = \nabla\mu(\theta_{\text{lf}})^T J(\theta_{\text{lf}})^{-1}u(y, \theta_{\text{lf}})$, where ∇ denotes the gradient operator. Because of this, the MSE of $\psi(\hat{\theta})$ can be estimated using (32), provided a consistent estimator of the true value exists and satisfies the conditions stated previously in this chapter. Since the FIC of the parametric model is the MSE of $\psi(\hat{\theta})$, we can compute the FIC of f_θ using (32). A related idea is used in Jullum and Hjort (2017), and their formulas are indeed quite similar to ours. We will now discuss the FIC of Jullum and Hjort (2017) in closer detail and use (32) to define a new FIC.

4.1 | The new and old FIC

In Jullum and Hjort (2017), the authors derive a FIC, allowing for comparison of multiple nonnested parametric models and nonparametric alternatives, by evaluating how well a certain pre-specified parameter of interest μ is estimated. This quantity is called the focus parameter, and a model's FIC score is an approximation to the MSE of its estimator $\hat{\mu}_1$ of μ.

To derive the criterion, the authors proceed more or less as we have done in Section 3. They argue that $\hat{b}^2 = (\hat{\mu}_1 - \hat{\mu}_0)^2$ tends to overshoot the actual bias of $\hat{\mu}_1$ and that $\kappa/n = \text{Var}\phi(Y_i)/n$ should be subtracted from the expression ot correct for this. Using our notation, their final formula for the FIC is as follows,

$$\text{FIC}_{\text{old}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n\} + \hat{\tau}/n.$$

This expression is very similar to the one given in (32). The only difference is that the above expression lacks the $2\hat{b}\hat{c}_0/n$ term, which is present in (32). Because of this, FIC_{old} is easier to compute than the expression in (32), but, unless b or c is zero, the expected value of the quantity is $O(1/n)$ away from the true MSE of $\hat{\mu}_1$. Hence, the FIC of Jullum and Hjort (2017) will, in many cases, underemphasize the importance of the variance since the error of the estimated squared bias is on the same scale as this quantity. We therefore propose the following new form of the FIC, which in some sense can be seen as a corrected version of FIC_{old},

$$\text{FIC}_{\text{new}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}/n\} + \hat{\tau}/n.$$

As argued previously, this estimator is only $o(1/n)$ from the true MSE of $\hat{\mu}_1$. With this new version of the FIC, we need to estimate c_0, as was discussed in Section 2. All other quantities can be approximated by the canonical estimators.

4.2 | How much does it matter?

Even though the the FIC from Jullum and Hjort (2017) is slightly off its mark, this version of the FIC has had success in many applications. See, for instance, Cunen, Walløe, and Hjort (2020), Claeskens et al. (2019), Wang and Hobæk Haff (2019), and Hermansen et al. (2015). It is therefore natural to ask "how wrong" this formula really is. In this section, we will present some examples illustrating the differences.

Firstly, notice that when $\hat{\mu}_0$ is unbiased, $c = 0$. When this is the case, the term distinguishing FIC_{old} from FIC_{new} is equal to zero and the two versions of the criterion agree. This happens when $\hat{\mu}_0$ is a mean or some bias corrected estimator. Another perhaps less trivial example is linear combinations of the regression coefficients used in linear regression. These estimators are unbiased for the true values, provided the model is correctly specified. Hence c_0 is zero in this case. Because of this, the use of the FIC in, for instance, Cunen, Walløe, and Hjort (2020) and Claeskens et al. (2019) is unproblematic.

Another situation where FIC_{old} and FIC_{new} give very similar results, is when either b or c are very small. If either $\hat{\mu}_1$ or $\hat{\mu}_0$ are almost unbiased for the true value μ_0, this can happen, and the term distinguishing the two versions of the criterion is too small to have any real effect. In addition, we truncate negative estimates of the squared bias to zero. In some cases, the estimated squared bias will be less than zero regardless of whether we subtract $2\hat{b}\hat{c}_0/n$ or not. Hence, the two criteria can end up being very similar, if not equal, when the bias is small, relatively speaking. An example of such a situation, is when estimating the difference in median life lengths of Roman era Egyptian men and women using data from Pearson (1902). Jullum and Hjort (2017) analyze this data set in their article and use FIC_{old} to choose the model estimating the focus parameter with the highest precision. We repeated their computations in addition to calculating FIC_{new} for the same data. In this example, the two information criteria gave almost identical results. The reason for this, is that the bias estimate is truncated for all but one model. This happens for both FIC_{new} and FIC_{old}, and hence, the two criteria agree for all but the Gompertz model. For this one case, FIC_{new} was 0.022 larger than FIC_{old}. The Gompertz model is, however, by far the worst of the ones considered in Jullum and Hjort (2017) with very high estimated MSE of the focus parameter. Because of this, adding 0.022 to the result, does not change the conclusions we would draw from using the FIC in this case.

The above example might lead one to believe that the difference between the new and old FIC is negligible in all but a few situations. It might therefore be tempting to use FIC_{old} rather than FIC_{new} because of its relatively easier formulation. This, however, is not to be advised, as there in general is no bound on the size of term $2bc/n$. To see this, consider fitting an exponential model to i.i.d. data points $Y_1, \ldots, Y_n \sim \text{Gamma}(k, \sqrt{k})$ for a fixed number k in order to estimate the SD in the distribution, which is 1 for all k. In this situation, the limiting bias in the model will grow with k, and as a result FIC_{old} will loose precision when k is large. A small simulation supports this claim. For a selection of k-values, we simulated one million data sets of size $n = 100$. In each case, we fitted an exponential model and computed FIC_{old} as well as FIC_{new} using the SD as focus parameter. We used the empirical SD as the consistent estimator $\hat{\mu}_0$. In Figure 1, we illustrate the

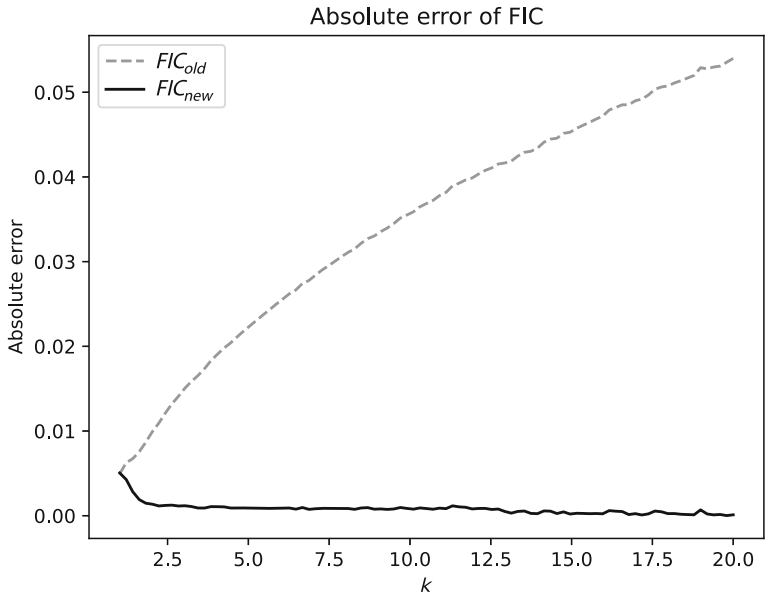

FIGURE 1 The figure displays the absolute value of the difference between the mean of 10^6 simulated FIC-values and the true mean squared error as functions of the shape parameter in the underlying Gamma$(k, \sqrt{k})$-distribution of the data. The error for the old version of the information criterion is shown by a grey dashed line, while the error of new version is plotted as a solid black line. The focus parameter is the SD in the distribution and the empirical SD is used as the consistent estimator $\hat{\mu}_0$. The sample size is $n = 100$.

results. We have plotted the estimated absolute value of the error of both FIC_{old} and FIC_{new} as functions of k, the shape parameter in the true underlying Gamma distribution. The figure shows clearly that not taking the term $2bc/n$ into account, causes the error of FIC_{new} to grow large as k increases.

This is of course only a toy example, but shows that there, at least in theory, is no bound on $2bc/n$ and that in certain situations ignoring its effect can be very problematic. We therefore recommend using FIC_{new} unless working exclusively with unbiased estimators. We will now illustrate this even further by trying out the new and old FIC on a dataset where the two criteria give very different results.

4.3 | Application—battle deaths in inter-state wars

In Pinker (2011) it is claimed that the world is gradually becoming more peaceful. Such statements have since been a heated topic of discussion. See for example, Cirillo and Taleb (2016a, 2016b), Clauset (2018), or Cunen, Hjort, and Nygård (2020). When investigating claims like that of Pinker statistically, proper modeling is important.

In this section, we will use the FIC to evaluate suitability of different models for checking the claim of Pinker based on the Correlates of War (CoW) dataset (Sarkees & Wayman, 2010). This dataset consists of the number of battle deaths in the 95 most recently concluded wars. The number of battle deaths is by no means a complete measure of violence. Hence, using the CoW data set to assess the claim of Pinker is of course somewhat lacking. Nevertheless, we use

these data to estimate the difference in median battle deaths between newer and older inter-state wars.

To compare recent wars to older ones, we need to decide which conflicts should be considered new. Cunen, Hjort, and Nygård (2020) investigate this question. Their analysis finds that the Korean War marks a change in the number of battle deaths in inter-state wars. Building on this, we define the "old" wars to be all conflicts before and including the Korean war. The remaining wars are treated as "new" conflicts.

We consider the data points as independent. Furthermore, we assume that the battle deaths in the recent wars are identically distributed and that the same holds true for the older wars. In our analysis, we consider eight different models and investigate how well they estimate the difference in median deaths between newer and older inter-state wars. For each case, we fit the corresponding model to the data sets consisting of newer and older wars separately. Afterwards, we use the fitted models to estimate the difference between median number of battle deaths in older and more recent wars. To evaluate the different models, we compute $\mathrm{FIC}_{\mathrm{new}}$ for each case. As the consistent estimator, $\hat{\mu}_0$, we use the difference between the empirical median for the older and newer wars. We use the default method in `R` with `type` equal to 7 to estimate this quantity. Since the two datasets are independent, κ/n and τ/n are estimated by the sum of the corresponding quantities for each separate data set. To approximate c_0 and b, we take the difference between the corresponding estimates for older and more recent wars. Afterwards, the squared bias is estimated and truncated to zero in the cases where the approximation is negative. It is worth noting that the forms of κ, τ, b, and c_0 are relatively simple in this case where the focus parameter is the difference between two quantiles. For more complicated functions, for example, $\hat{\mu}_B/\hat{\mu}_A$ where $\hat{\mu}_A$ and $\hat{\mu}_B$ are the median number of battle deaths in newer and older wars, respectively, the delta method and the result of Section 2.5 need to be applied.

We worked with a slightly modified version of the data. In the CoW dataset only wars with more than 1001 battle deaths are registered. To avoid problems with observations on the bounds of the support, we replaced all observations of 1001 with 1001.01. The same trick is used in Cunen, Hjort, and Nygård (2020). See this article for more explanations and discussions. After transforming the data, we fitted eight parametric models. Firstly, Dagum, Log-logistic, Pareto IV, Lomax, log-normal and log-Cauchy distribution were fit to the data shifted by 1001 to the left. Secondly, we fitted log-Gamma and log-Weibull distributions to the data scaled by 1000. The resulting values of $\mathrm{FIC}_{\mathrm{new}}$ for each model can be found in the plot in the left panel of Figure 2 together with the maximum likelihood estimate of the focus parameter for each model. From the figure, we notice that using the log-normal distributions results in the lowest MSE for the focus parameter. With this model, the estimated difference between median battle deaths before and after the Korean War is 2066.

To illustrate the difference between $\mathrm{FIC}_{\mathrm{new}}$ and $\mathrm{FIC}_{\mathrm{old}}$, we also computed the criterion of Jullum and Hjort (2017). The results are displayed in Figure 2. From the plot, we notice that the two methods indeed give very different estimates of the MSE for many of the models, though not all. In fact, using $\mathrm{FIC}_{\mathrm{new}}$ rather than $\mathrm{FIC}_{\mathrm{old}}$ actually leads to different models being chosen. With the model selected by $\mathrm{FIC}_{\mathrm{old}}$ the difference between median number of battle deaths before and after the Korean War is estimated to be 3958, which is almost twice as large as the estimate we got from the model chosen by $\mathrm{FIC}_{\mathrm{new}}$. This shows that the additional term in $\mathrm{FIC}_{\mathrm{new}}$ does have a real implication and is important for analysing the CoW dataset correctly.

The above analysis can be repeated for other quantiles than the median. In the right panel of Figure 2 we have computed and displayed the FIC values for and estimates of the difference in the third quartile in the two distributions. This can be used as a measure of how violence in the

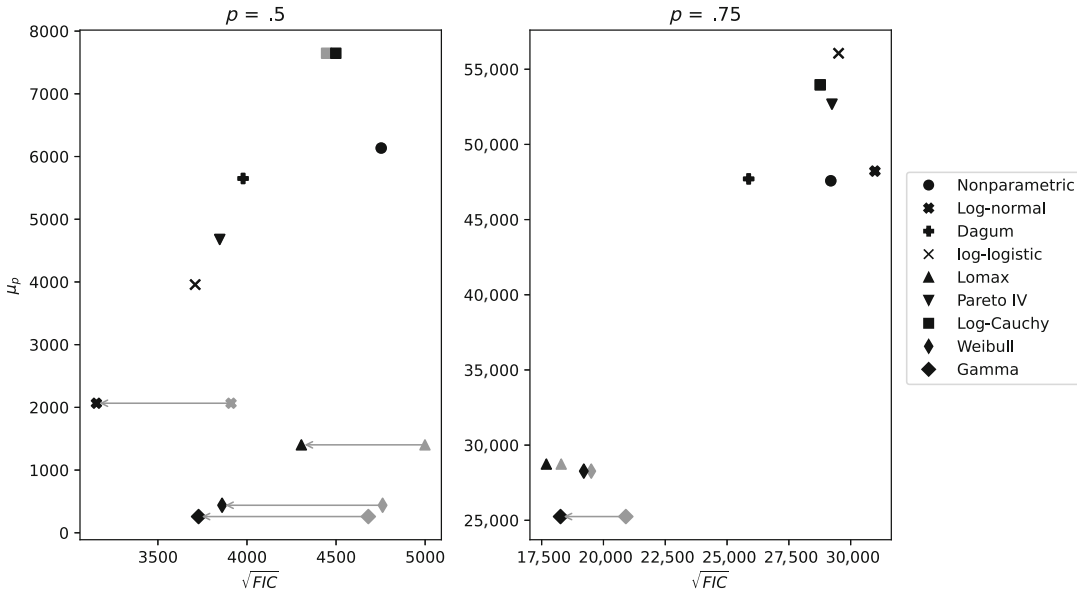

FIGURE 2 The figure shows FIC_{old} (colored grey) and FIC_{new} (colored black) for different models fitted to the correlates of war (CoW) dataset together with the models corresponding estimate of the focus parameter. The focus parameter is the difference in the p-quantile for older and newer wars. The value of p is shown in the title of each plot.

more deadly wars has changed in recent years. From the plot we notice that the two criteria give similar results in this case, with both the new and the old FIC selecting the Lomax model with an estimate of 28,740 for the focus parameter.

The plots in Figure 2 give some insight into how the new and old version of FIC differ. For the case of the median, c_0/n was estimated to be approximately 1227. This is a positive number, and hence the nonparametric estimate of the focus parameter is biased upward. Because of this, we would expect models underestimating the focus parameter compared to the nonparametric estimate, to be "less off" than what the naive estimate $\hat{b}$ suggests. Models estimating the focus parameter to be even larger than $\hat{\mu}_0$, on the other hand, are likely to be more biased than what $\hat{b}$ would lead one to believe. This intuitive idea is what is caught by FIC_{new}, but not FIC_{old}. If $\hat{b}$ and $\hat{c}_0$ have the same signs, $2\hat{b}\hat{c}/n > 0$ and the model is penalized by the criterion developed in this article, whereas when $\hat{\mu}_0$ is biased in the opposite direction of $\text{sign}(\hat{b})$, $2\hat{b}\hat{c}/n < 0$, resulting in a lower value of FIC_{new} than FIC_{old} unless the estimate of the squared bias is truncated to zero.

The above calculations show that there can be real and impactful differences between the old and new FIC. This is, however, not a guarantee that FIC_{new} performs better than FIC_{old} in this example. To investigate whether this is the case, we need to compute the true value of the focus parameter and compare the true MSEs with the corresponding FIC_{new} and FIC_{old} scores. In theory, this is a simple task, but in practice we do now know the true underlying random structure of battle deaths in inter-state wars. Because of this, we cannot check the performance of FIC_{new} and FIC_{old} directly. To get some indication on the quality of the MSE estimates we will therefore instead work with a distribution similar to our data set where the truth is known. To achieve this, we fitted a Lomax model to the data as explained previously. Afterwards, we simulated 10,000 datasets from this model. For each simulated dataset, we computed $\hat{\mu}_1$, FIC_{old} and FIC_{new} for each of the models considered in this section. We used the difference in median number of battle

deaths before and after the Korean war as the focus parameter. Afterwards, we used the 10,000 samples to estimate the expected value of FIC_{new} and FIC_{old} in addition to the true MSE for $\hat{\mu}_1$ for each model. For five out of the eight models considered, the average value of FIC_{new} was closer to the true MSE than FIC_{old}. For the remaining three cases FIC_{old} performed better than FIC_{new}. The wo criteria were more or less equally variable, with $\text{VarFIC}_{\text{new}}/\text{VarFIC}_{\text{old}}$ ranging from 0.655 to 1.186 for the different models.

Comparing only the FIC-scores gives somewhat of a limited view. FIC_{old} and FIC_{new} differ only in their estimate of the squared bias. The estimated variance is identical in the two criteria. Because of this, an over- or underestimated variance might make it beneficial to under- or overestimate the squared bias, respectively, as this "cancels" parts of the error made by the variance estimate. Although this might lead to more precise estimates of the MSE in certain cases, such cancellation happens mostly due to "luck" and is in general not something one should hope or aim for. We therefore believe it is more relevant to compare the quality of the estimates of the squared bias made by each criterion rather than the full MSE scores. Looking only at the estimated squared bias shows that the cancellation described above has happened in two of the cases where FIC_{new} performed better than FIC_{old}. We find that the estimate of the squared bias in the new criterion is indeed closer to the true value than the estimate provided by FIC_{old} in all but one model.

Choosing to fit a Lomax model and simulate from this is of course not the only way one can generate data similar to the number of battle deaths in the 95 most recent and concluded inter-state wars. We could fit any of the models considered in this section or even draw bootstrap samples from the dataset itself. We have tried out a number of these methods, and similar results were found in all cases. For the sake of clarity and brevity, we will therefore not go into depth about all settings, but taking the average over all simulation settings gives the following result. More often than not, FIC_{new} performed better than FIC_{old} when it comes to estimation of the MSE. Furthermore, the squared bias estimate provided by FIC_{new} were on average more precise than the one from FIC_{old} for a little more than five out of eight models. In addition, the average value of $\text{VarFIC}_{\text{new}}/\text{VarFIC}_{\text{old}}$ ranged from 0.457 to 1.025 between the different models.

5 | CONCLUDING REMARKS

In this article, we have only considered the framework of i.i.d. data, but extensions to more complicated situations are certainly possible. For regression and classification settings with responses Y_i and predictors X_i for $i = 1,\ldots,n$, the formulas derived in this article are directly applicable when the covariates are considered random. This is because the tuples $Z_i = (Y_i, X_i)$ for $i = 1,\ldots,n$ can be considered as i.i.d. data points in this case. In addition, results like the Lindeberg–Feller theorem (see e.g., Billingsley, 1995, p. 359) allows most of the formulas derived in this article to be generalized to situations where the covariates are considered nonstochastic. It is, however, worth noting that in regression and classification settings, nonparametric estimators satisfying (29) can be hard to come by. This is especially true when the number of covariates is high. In such cases, a solution can be to replace the true parameter, μ_0, by the least false parameter in a widest model. This is in some sense the closest we can get to the true value, and the FIC should therefore measure the MSE of estimators relative to this quantity. Claeskens et al. (2019) and Cunen, Walløe, and Hjort (2020) take this approach for the simpler FIC of Jullum and Hjort (2017). Modifications along the same lines can be made for the FIC derived in this article as well, but to use the more precise information criterion developed in this article, c_0 needs to be estimated. This can be achieved

using Theorem 3. Further extensions of the iid framework include dependent and censored data . Arguments similar to those given in Hermansen et al. (2015) and Jullum and Hjort (2019), respectively, can be used to generalize the results of this paper to these more complicated data structures.

FIC is a popular research topic and there exists many extensions. Most of which can also be made for the criterion developed in this article as well. We will mention two such extensions, but many other possibilities exist. For an overview of techniques, consult Claeskens and Hjort (2008).

In certain situations, multiple focus parameters are of interest and should be taken into account when performing model selection. One way of achieving this for the FIC is to use average FIC (AFIC). Assume the focus parameters can be written as $\mu(u)$ for u in some index set U. A risk function taking all parameters into account is the following expression,

$$\mathrm{E}\left(\int_U [\hat{\mu}(u) - \mu(u)]^2 \, dW(u)\right) = \int_U \mathrm{bias}[\hat{\mu}(u)]^2 \, dW(u) + \int_U \mathrm{Var}\hat{\mu}(u) \, dW(u),$$

where W is some distribution over U indicating how much each focus parameter matters relative to the others. The risk of the estimators $\hat{\mu}(u)$ for $u \in U$ can therefore be estimated by the following expression

$$\mathrm{AFIC} = \max\left\{0, \int_U \left(\hat{b}(u)^2 - \frac{\hat{\kappa}(u)}{n} + \frac{2\hat{b}(u)\hat{c}_0(u)}{n}\right) dW(u)\right\} + \frac{1}{n}\int_U \hat{\tau}(u) \, dW(u).$$

Here $\hat{b}(u)$, $\hat{c}_0(u)$, $\hat{\kappa}(u)$a, and $\hat{\tau}(u)$ are the estimators discussed previously in this article computed for the estimator $\hat{\mu}(u)$. Notice that AFIC does not equal the integral of the individual FIC scores as we truncate possibly negative estimates of the integrated squared bias rather than the individual bias estimates for each $u \in U$ to zero. For more information about and variants of AFIC, see Claeskens and Hjort (2008, pp. 179–183).

FIC scores can be used for more than strictly selecting the optimal model for estimating the focus parameter. An alternative approach, is to use model averaging. Assume we have p models with estimates $\hat{\mu}_1,\ldots,\hat{\mu}_p$ of the focus parameter. Rather than using the FIC to choose the estimator with the lowest MSE, we could use the weighted average $\hat{\mu}_{\mathrm{avg}} = \sum_{j=1}^p w_j\hat{\mu}_j$, where w_j are weights summing to one. The weights should be related to the FIC scores such that $\hat{\mu}_j$s with low MSE are given more weight than $\hat{\mu}_j$s for which $\mathrm{FIC}(\hat{\mu}_j)$ is large. Inspired by Eq. 6.1 in Jullum and Hjort (2017), we suggest the following form for the weights

$$w_j = \frac{\exp[-\lambda \mathrm{FIC}(\hat{\mu}_j)]}{\sum_{k=1}^p \exp[-\lambda \mathrm{FIC}(\hat{\mu}_k)]},$$

for some tuning parameter λ.

ACKNOWLEDGMENTS

This research is funded in part by The Norwegian Research Council 237718 through the Big Insight Centre for research-driven innovation. The authors are grateful for support from Centre for Advanced Research, Oslo, via the Stability and Change project, led by Hjort. We appreciate the anonymous reviewers for their careful reading, constructive comments, and suggestions, which contributed to improving our the manuscript.

ORCID

Ingrid Dæhlen https://orcid.org/0000-0002-2029-4451

REFERENCES

Billingsley, P. (1995). *Probability and measure* (3rd ed.). Wiley.

Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons.

Cirillo, P., & Taleb, N. N. (2016a). *The decline of violent conflicts: What do the data really say*? [Conference presentation]. The Nobel Foundation Symposium 161: The Causes of Peace, 1–26.

Cirillo, P., & Taleb, N. N. (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, *452*, 29–45.

Claeskens, G., Croux, C., & Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, *62*, 972–979.

Claeskens, G., Cunen, C., & Hjort, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, *7*, 415.

Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*, 900–916.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science*. *Advances*, *4*, eaao3580.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society Series B (Methodological)*, *30*, 248–275.

Cunen, C., Hjort, N. L., & Nygård, H. M. (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research*, *57*, 221–234.

Cunen, C., Walløe, L., & Hjort, N. L. (2020). Focused model selection for linear mixed models with an application to whale ecology. *The Annals of Applied Statistics*, *14*, 872–904.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics*, *24*, 2431–2461.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Hermansen, G. H., Hjort, N. L., & Jullum, M. (2015). *Parametric or nonparametric: The FIC approach for stationary time series* [Conference presentation]. Proceedings of the 60th World Statistics Congress of the International Statistical Institute, 4827–4832.

Hjort, N. L., & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, *23*, 882–904.

Hjort, N. L., & Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, *24*, 1619–1647.

Huber, P. J. (2009). *Robust statistics* (2nd ed.). Wiley.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*, 401–419.

Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, *27*, 951–981.

Jullum, M., & Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Analysis*, *25*, 406–438.

Kim, K. I. (2016). Higher order bias correcting moment equation for M-estimation and its higher order efficiency. *Econometrics*, *4*, 48.

Ko, V., Hjort, N. L., & Hobæk Haff, I. (2019). Focused information criteria for copulas. *Scandinavian Journal of Statistics*, *46*, 1117–1140.

Lindstrøm, T. L. (2017). *Spaces: An introduction to real analysis* (Vol. *29*). American Mathematical Society.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, *24*, 1602–1618.

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, *1*, 261–264.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Viking.

Quenouille, M. H. (1949). *Approximate tests of correlation in time-series 3*. In *Mathematical proceedings of the Cambridge philosophical society* (Vol. *45*, pp. 483–484). Cambridge University Press.
Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360.
Sarkees, M. R., & Wayman, F. (2010). *Resort to war: 1816 - 2007*. CQ Press.
Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, *1*, 185–202.
Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer Science & Business Media.
Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer Science & Business Media.
Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, *17*, 1176–1197.
Singh, R. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika*, *66*, 177–180.
Tukey, J. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, *29*, 614.
Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge university press.
Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall/CRC.
Wang, Y., & Hobæk Haff, I. (2019). Focussed selection of the claim severity distribution. *Scandinavian Actuarial Journal*, *2019*, 129–142.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.
Zhang, X., & Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, *39*, 174–200.

How to cite this article: Dæhlen, I., Hjort, N. L., & Hobæk Haff, I. (2024). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics*, *51*(2), 724–759. https://doi.org/10.1111/sjos.12696

APPENDIX A. KERNEL DENSITY ESTIMATION

In Section 2.2, we derived an estimator for c when $\hat{\mu}$ is a quantile. To use the formula in (10), however, we need consistent estimators of both the density f and its derivative f' at μ_p.

The classic kernel density estimators are

$$\hat{f}(x) = \frac{1}{nh_0}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_0}\right) \quad \text{and} \quad \hat{f}'(x) = -\frac{1}{nh_1^2}\sum_{i=1}^{n} K'\left(\frac{x_i - x}{h_1}\right). \tag{A1}$$

In the above, h_1 and h_2 are called the bandwidths and K is the kernel. In principle, the kernel can be any nonnegative and differentiable function, but to simplify calculations we will assume it is symmetric. The obvious choice for such a kernel is the standard normal density, ϕ.

The standard strategy is to minimize the asymptotic integrated MSE. Minimizing the L^2 distance between $\hat{f}$ and $\hat{f}'$ and the corresponding true values is reasonable when a good overall fit is desired. In our situation, however, we are only interested in using the kernel density estimates to approximate $f(\mu_p)$ and $f'(\mu_p)$. Hence, we should choose bandwidths minimizing the MSE of these estimators and not the full integrated MSE. Arguing as in Singh (1979), one can show that this is achieved by

$$h_1 = \left(\frac{R(K)f(\mu_p)}{S(K)^2 f^{(2)}(\mu_p)^2}\right)^{1/5} n^{-1/5} \quad \text{and} \quad h_2 = \left(\frac{3R(K')f(\mu_p)}{S(K)^2 f^{(3)}(\mu_p)^2}\right)^{1/7} n^{-1/7},$$

where $f^{(k)}$ denote the kth derivative of f and

$$R(H) = \int_{\mathbb{R}} H(u)^2\, du \quad \text{and} \quad S(H) = \int_{\mathbb{R}} u^2 H(u)\, du.$$

See Chap. 2 of Wand and Jones (1995) for more details.

To estimate the optimal bandwidths, we need to know the value of $f(\mu_p)$, $f^{(2)}(\mu_p)$ and $f^{(3)}(\mu_p)$. These quantities are, however, unknown by definition of the problem. To estimate them, we therefore suggest fitting parametric models to the data. This can lead to inconsistent estimators of $f(\mu_p)$, $f^{(2)}(\mu_p)$ and $f^{(3)}(\mu_p)$, but this will only result in bandwidths being slightly off from the optimal ones. The estimators in (A1) will still converge in probability to the true values, $f(\mu_p)$ and $f'(\mu_p)$ when inconsistent estimators of $f(\mu_p)$, $f^{(2)}(\mu_p)$, and $f^{(3)}(\mu_p)$ are used to find optimal bandwidths.

We recommend the following rule of thumb: If the distribution has a heavy tail, fit a Pareto distribution to the data. Otherwise, fit a normal distribution. These distributions are chosen mainly because their maximum likelihood estimators have closed form expressions, allowing $f(\mu_p)$, $f^{(2)}(\mu_p)$, and $f^{(3)}(\mu_p)$ to be estimated without numerical optimization.

Combining all of the above, leads to the following rule of thumb for estimating $f(\mu_p)$ and $f'(\mu_p)$:

$$\hat{f}(x) = \frac{1}{nh_0}\sum_{i=1}^{n}\phi\left(\frac{x_i - x}{h_0}\right) \quad \text{and} \quad \hat{f}'(x) = \frac{1}{nh_1^2}\sum_{i=1}^{n}\phi'\left(\frac{x_i - x}{h_1}\right).$$

where ϕ is the standard normal density and

$$\begin{aligned} h_0 &= \hat{\mu}_p\left[2\sqrt{\pi}(Y_{(1)}/\mu_p)^{\hat{\alpha}}\hat{\alpha}(\hat{\alpha}+1)^2(\hat{\alpha}+2)^2 n\right]^{-1/5} \quad \text{and} \quad h_1 \\ &= \hat{\mu}_p\left[(4/3)\sqrt{\pi}(Y_{(1)}/\mu_p)^{\hat{\alpha}}\hat{\alpha}(\hat{\alpha}+1)^2(\hat{\alpha}+2)^2(\hat{\alpha}+3)^2 n\right]^{-1/7}, \end{aligned}$$

when the distribution of Y is heavy tailed and

$$\begin{aligned} h_0 &= \sigma^2 \exp\left[(\hat{\mu}_p - \overline{Y})^2/10\hat{\sigma}^2\right]\left\{\sqrt{2}\hat{\sigma}\left[(\hat{\mu}_p - \overline{Y})^2 - \hat{\sigma}^2\right]^2 n\right\}^{-1/5} \quad \text{and} \\ h_1 &= \sigma^2 \exp\left[(\hat{\mu}_p - \overline{Y})^2/14\hat{\sigma}^2\right]\left\{(2\sqrt{2}/3)\hat{\sigma}(\hat{\mu}_p - \overline{Y})^2\left[(\hat{\mu}_p - \overline{Y})^2 - 3\hat{\sigma}^2\right]^2 n\right\}^{-1/7}, \end{aligned}$$

otherwise. In the above, $\hat{\sigma}$, $\overline{Y}$, and $Y_{(1)}$ are the empirical SD, mean, and minimum value in the sample, respectively. In addition, $\hat{\mu}_p$ is the empirical quantile and

$$\hat{\alpha} = \left[\frac{1}{n}\sum_{i=1}^{n}\log(Y_i/Y_{(1)})\right]^{-1}.$$

Lastly, we would like to remark that there exists more sophisticated methods for estimating $f(\mu_p)$ and $f'(\mu_p)$ than the ones presented here. For instance, rather than using a purely nonparametric or parametric estimate, a semi-parametric approach is possible. Parametric and nonparametric models can be combined as in, for example, Hjort and Glad (1995) and Efron and Tibshirani (1996). Another approach is to use local likelihoods, see Hjort and Jones (1996) and Loader (1996).

APPENDIX B. COMPLETE PROOFS

We will use the following notation. If A is a matrix $A_{j,k}$ denotes the (j,k)th component of A. Furthermore A^j is notation for the jth column vector of A. Similarly, a_j denotes the jth component of a if a is a vector. Lastly we will use $\nabla_{x_0} f(x)$ and $H_{x_0} f(x)$ as the gradient and hessian matrix of f at x_0, respectively.

B.1 Uniform integrability

To give full proofs and sets of conditions for the theorems in the article, we will need to show that multiple quantities are uniformly integrable. We give the definition here, for the sake of completion.

Definition 1 (Uniformly integrable, Billingsley (1999)). A sequence of random variables $(X_n)_n \subseteq \mathbb{R}$ is uniformly integrable if the following holds,

$$\lim_{K\to\infty} \sup_n \mathrm{E}(|X_n| \ : \ |X_n| \geq K) = 0.$$

We also state a lemma containing multiple results concerning uniform integrability. These results are not new and proofs can be found in most textbooks on the subject. Properties we will use are nevertheless included here fore easier reference in the arguments that follow.

Lemma 2. *In the following, let $(X_n)_n$, $(Y_n)_n \subseteq \mathbb{R}$ be two sequences of random variables. Furthermore, let $A_n \in \mathbb{R}^{p\times q}$ and $B_n \in \mathbb{R}^{q\times p'}$ be random matrices and $C \in \mathbb{R}^{q\times p''}$ a fixed matrix.*

(i) *X_n is uniformly integrable if and only if $|X_n|$ is uniformly integrable.*
(ii) *Let $a, b \in \mathbb{R}$ and assume that X_n and Y_n are uniformly integrable, then $aX_n + bY_n$ is uniformly integrable*
(iii) *Let X_n^2 and Y_n^2 be uniformly integrable, then X_nY_n is uniformly integrable.*
(iv) *If X_n^p is uniformly integrable, X_n^q is uniformly integrable for all $1 \leq q < p$.*
(v) *If X_n converges in distribution to X and $\mathrm{E}|X_n|$ converges to $\mathrm{E}|X|$, X_n is uniformly integrable.*
(vi) *If each component of A_n is uniformly integrable, each component of CA_n is uniformly integrable as well.*
(vii) *If the square of each component of A_n and B_n both are uniformly square integrable, each component of A_nB_n is uniformly integrable.*

Proof. (i) holds by definition. (ii) follows from the triangle inequality and (iii) from Hölder's inequality. Since $\mathrm{E}X_n^p$ converges to $\mathrm{E}X^p$ by uniform integrability of X_n^p, $\sup_n \mathrm{E}(|X_n|^{q+\epsilon}) < \infty$ where $\epsilon = (p-q)/q > 0$. By (3.18) in Billingsley (1999, p. 31), this implies that X_n^q is uniformly integrable, proving (iv). Case (v) is a consequence of Thm. 3.6 in Billingsley (1999, p. 32), the continuous mapping theorem and case (i). The two remaining points follow from (ii) and (iii) respectively. ∎

B.2 Functions of unbiased estimators

We start with the simplest case: functions of unbiased estimators.

Theorem 9 (Functions of unbiased estimators). *Let $\hat{a} \in \mathbb{R}^p$ be an estimator of a_0 such that $\sqrt{n}(\hat{a} - a_0)$ converges in distribution to some U with finite second moment. Assume that all components of $\sqrt{n}(\hat{a} - a_0)$ to the power of three are uniformly integrable and that there exists a neighborhood $\mathcal{N}$ of a_0 on which $h : \mathbb{R}^p \to \mathbb{R}$ has continuous partial derivatives up to order 3. Then*

$$\lim_{n\to\infty} n\,\mathrm{E}[h(\hat{a}) - h(a_0)] = \frac{1}{2}\mathrm{Tr}[Hh(a_0)\mathrm{E}UU^T].$$

Proof. Since all components of $\sqrt{n}(\hat{a} - a_0)$ are uniformly square integrable, case (vii) of lemma 2 ensures that all components of $\mathrm{E}n(\hat{a} - a_0)(\hat{a} - a_0)^T$ are uniformly integrable as well. Hence, $\mathrm{E}n(\hat{a} - a_0)(\hat{a} - a_0)^T$ converges to $\mathrm{E}UU^T$ and $n\mathrm{E}(\hat{a} - a_0)^T Hh(a_0)(\hat{a} - a_0) = Hh(a_0)\mathrm{E}UU^T + o(1)$. To show the theorem, we therefore only need to prove that, $n\mathrm{E}\epsilon_n(\hat{a}) = o(1)$, where $\epsilon_n(\hat{a})$ is the remainder term in (6).

Choose a compact subset K of $\mathcal{N}$ with a_0 in its interior. Since $\hat{a}$ converges in probability to a_0, it lies in the interior of K with probability tending to one. Without loss of generality, we will therefore assume that $\hat{a}$ lies in the interior of K. Since all third-order partial derivatives of h are continuous on this set, the extreme value theorem guarantees that they are bounded by some $M \in \mathbb{R}$. Hence, by standard results for Taylor expansions of multivariate functions,

$$|\epsilon_n(\hat{a})| \leq \frac{M}{3!}\|\hat{a} - a_0\|_1^3.$$

See for example, Coroll. 6.5.8 in Lindstrøm (2017). In the above, $\|\cdot\|_1$ denotes the L^1 norm on $\mathbb{R}^p$. Since all components of $\sqrt{n}(\hat{a} - a_0)$ to the power of three are uniformly integrable, and $n^{3/2}\|\hat{a} - a_0\|_1^3$ is $O_{\mathrm{Pr}}(1)$ by the continuous mapping theorem, we have $\mathrm{E}|\epsilon_n(\hat{a})| = O(n^{-3/2})$. ∎

For the special case of empirical means, the theorem simplifies somewhat.

Corollary 1 (Functions of means). *Let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be random variables with empirical mean $\hat{\mu}$ and expected value μ_0. Assume further that the expected value of all components of Y_i to the power of four exists, and that h is a function with continuous partial derivatives up to order three in a neighborhood of μ_0. Then*

$$\lim_{n\to\infty} n\,\mathrm{E}[h(\hat{\mu}) - h(\mu_0)] = \frac{1}{2}\mathrm{Tr}[Hh(\mu_0)\mathrm{Var}Y_i].$$

Proof. By the central limit theorem $\sqrt{n}(\hat{\mu} - \mu_0)$ converges to a normal distribution with variance $\mathrm{Var}Y_i$. Because of this, the result follows from Theorem 9 provided all components of $\sqrt{n}(\hat{\mu} - \mu_0)$ to the power of three are uniformly integrable. By case (iv) of lemma 2, this holds true when all components of $\sqrt{n}(\hat{\mu} - \mu_0)$ to the power of four are uniformly integrable. We will show this latter statement.

Fix j. By the continuous mapping theorem, $n^2[(\hat{\mu})_j - (\mu_0)_j]^4$ converges in distribution to U^4 where $U \sim \mathrm{N}[0, (\mathrm{Var}Y_i)_{jj}]$. Furthermore, utilizing that $Y_1, \ldots, Y_n$ are independent and have mean zero one can show

$$\mathrm{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i)_j-(\mu_0)_j]\right)^4=$$
$$\frac{1}{n^2}\sum_{i=1}^{n}\{\mathrm{E}[(Y_i)_j-(\mu_0)_j]^4-3(\mathrm{Var}Y_i)^2_{j,j}\}+3(\mathrm{Var}Y_i)^4_{j,j}\to 3(\mathrm{Var}Y_i)^2_{j,j},$$

which is equal to the fourth moment of the jth component of U. Hence, $n^2\left[(\hat{\mu})_j-(\mu_0)_j\right]^4$ is uniformly integrable and Theorem 9 can be applied. ∎

B.3 Quantiles

The second case is quantiles.

Theorem 10 (Quantiles). *Let $Y_1,\ldots,Y_n$ be i.i.d. with distribution F and let μ_p be the true p-quantile in the distribution. Assume $|\sqrt{n}(\hat{\mu}_p-\mu_p)|^3$ is uniformly integrable and that there exists a neighbourhood $\mathcal{N}$ of μ_p on which $F:\mathbb{R}^p\to\mathbb{R}$ has continuous partial derivatives up to order 3, then*

$$n\mathrm{E}(\hat{\mu}_p-\mu_p)=\frac{n}{f(\mu_p)}\left(\frac{j+\gamma}{n+1}-p\right)-\frac{f'(\mu_p)p(1-p)}{2f(\mu_p)^3}+O(n^{-3/2}),$$

where f is the density function in the distribution F and $\hat{\mu}_p$ is the empirical quantile defined by $\hat{\mu}_p=(1-\gamma)Y_{(j)}+\gamma Y_{(j+1)}$ with j and γ defined by either of the rows in Table 1.

Proof. The only thing we need to prove is that $\mathrm{E}\delta_n=o(1/n)$, where δ_n is defined in (14) of the article. The rest of the argument is similar to those given in the article. Since δ_n is a sum of remainder terms, it suffices to show that the expected value of each of them are $o(1/n)$. We will focus on δ_n', defined in Equation (13) of the article. The other case is similar.

By standard results for quantiles, $\hat{\mu}_p$ is consistent for μ_p and $\sqrt{n}(\hat{\mu}_p-\mu_p)$ converges in distribution to a $\mathrm{N}[0,p(1-p)/f(\mu_p)]$ distribution. This allows us to argue as in the proof of Theorem 9, to show

$$n^{3/2}\mathrm{E}\delta_n'\le\frac{M}{3!}\mathrm{E}|\sqrt{n}(\hat{\mu}_p-\mu_p)|^3.$$

Since $|\sqrt{n}(\hat{\mu}_p-\mu_p)|^3$ is uniformly integrable by assumption and $|\sqrt{n}(\hat{\mu}_p-\mu_p)|^3=O(1)$ by the continuous mapping theorem, the above equation ensures $\mathrm{E}\delta_n'=O(n^{-3/2})$. ∎

B.4 Maximum likelihood estimators

We will now prove Theorem 3 of the article. We start by showing a lemma.

Lemma 3. *Let $Y_1,\ldots,Y_n$ be i.i.d. data points from a distribution F and $\mathcal{F}=\{\log f_\theta\mid\theta\in\Theta\}$ a family of densities indexed by an open set Θ. Let θ_{lf} denote the minimizer of the Kullback–Leibler divergence from F to $\mathcal{F}$ and $\hat{\theta}$ the maximum likelihood estimator. Furthermore, let*

$$J_n=-\frac{1}{n}H_{\theta_{\mathrm{lf}}}\ell_n(\theta)\quad\text{and}\quad J=-H_{\theta_{\mathrm{lf}}}\mathrm{E}\log f_\theta(Y),$$

and

$$W_n = \frac{1}{n}\frac{\partial^3 \ell_n}{\partial\theta\partial\theta^T\partial\theta^j}(\theta_{\text{lf}}) \quad \text{and} \quad W^j = \frac{\partial^3}{\partial\theta\partial\theta^T\partial\theta_j}\Big|_{\theta_{\text{lf}}} \text{E}\log f_\theta(Y),$$

where ℓ_n denotes the likelihood function based on $Y_1,\ldots,Y_n$ and $Y \sim F$. Then the following quantities are uniformly integrable:

(1) $n^2(\hat{\theta}_j - (\theta_{\text{lf}})_j)^4$ *for each* $j = 1,\ldots,p$
(2) $n\left[(J_n)_{j,k} - J_{j,k}\right]^2$ *for each* $j,k = 1,\ldots,p$
(3) $n[(W_n)_{j,k} - W_{j,k}]^2$ *for each* $j,k = 1,\ldots,p$,

provided the following assumptions hold true:

(A1) All partial derivatives up to order 4 of $\log f_\theta(y)$ *at* θ_{lf} *exists and are continuous for almost all* y. *In addition, all fourth moments of* $\nabla_{\theta_{\text{lf}}}\log f_\theta(Y)$ *and all second moments of all second- and third-order partial derivatives of* $\log f_\theta(Y)$ *at* θ_{lf} *exists and are finite.*
(A2) The matrix J *is positive definite.*
(A3) All fourth-order powers of $\sqrt{n}[\hat{\theta} - \theta_{\text{lf}} - (nJ)^{-1}\nabla\ell_n(\theta_{\text{lf}})]$ *are uniformly integrable.*

Proof. We start with (1). Fix j and introduce the two variables

$$N_n = n^{-1/2}\left(J^{-1}\right)^j\nabla\ell_n(\theta_{\text{lf}}) \quad \text{and} \quad \delta_n = \sqrt{n}\left\{\hat{\theta}_j - (\theta_{\text{lf}})_j - \left[(nJ)^{-1}\right]^j\nabla\ell_n(\theta_{\text{lf}})\right\}.$$

Then,

$$\left[\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\right]^4 = (N_n + \delta_n)^4 = N_n^4 + \delta_n^4 + 4N_n\delta^3 + 4N_n^3\delta_n + 6N_n^2\delta_n^2.$$

We will show that each term is uniformly integrable. By case (ii) of Lemma 2, this ensures that $n^2(\hat{\theta}_j - (\theta_{\text{lf}})_j)^4$ is uniformly integrable.

By the central limit theorem, N_n converges in distribution to a random variable $N \sim \text{N}(0,\sigma^2)$ where

$$\sigma = \left[(J^{-1})^j\right]^T\text{Var}[\nabla_{\theta_{\text{lf}}}\log f_\theta(Y_i)](J^{-1})^j.$$

Furthermore, elementary computations show

$$\text{E}N_n^4 = \frac{1}{n}(\text{E}\{\left[(J^{-1})^j\right]^T\nabla_{\theta_{\text{lf}}}\log f_\theta(Y_i)\}^4 - 3\sigma^4) + 3\sigma^4 \to 3\sigma^4 = EN^4,$$

since all fourth-order moments of $\nabla_{\theta_{\text{lf}}}\log f_\theta(Y_i)$ exist and are finite. Hence, by cases (v) and (iv) of Lemma 2, N_n^p is uniformly integrable for $p \le 4$.

By assumption and case (iv) of Lemma 2, δ_n^p is uniformly integrable for $p \le 4$. The remaining terms now follow directly. Since δ_n^2 and N_n^2 are uniformly integrable, case (iii) of Lemma 2 ensures that $N_n\delta_n$ is uniformly integrable. This, together with uniform integrability of N_n^4 and δ_n^4 and yet another application of case (iii) in Lemma 2, shows that $N_n\delta_n^3$ and $N_n^3\delta_n$ are uniformly integrable.

We will now prove (2). The argument for (3) is similar and hence omitted. Fix j and k. By the law of large numbers $n\left[(J_n)_{j,k} - J_{j,k}\right]^2$ converges almost surely to $\mathrm{Var} H_{\theta_{\mathrm{lf}}} \log f_\theta(Y_i)_{j,k}$. Furthermore, the expected value of $n\left[(J_n)_{j,k} - J_{j,k}\right]^2$ is equal to $\mathrm{Var} H_{\theta_{\mathrm{lf}}} \log f_\theta(Y_i)_{j,k}$. Hence, by case (v) of Lemma 2, $n\left[(J_n)_{j,k} - J_{j,k}\right]^2$ is uniformly integrable. ∎

We are now ready to prove Theorem 3 of the article.

Theorem 11. *Let all quantities be defined as in Lemma 2 and assume that (A1–A3) hold true. Then the bias of $\hat\theta - \theta_{\mathrm{lf}}$ is on the form given in Theorem 3 of the article, provided the following additional assumption*:

- *There exists a neighborhood, $\mathcal{N}$ of θ_{lf} and an integrable function m such that all fourth-order partial derivatives are bounded by $m(y)$ and $\mathrm{E} m(Y)^4$ exists and is finite.*

Proof. By the mean value theorem, the following holds for all $i, j, k = 1, \ldots, p$,

$$\left\| \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \log f_\theta(y) - \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\Big|_{\theta_{\mathrm{lf}}} \log f_\theta(y) \right\| \le \left\| \frac{\partial^4}{\partial\theta\partial\theta_i\partial\theta_j\partial\theta_k}\Big|_{\theta^*} \log f_\theta(y)(\theta - \theta_{\mathrm{lf}}) \right\|,$$

for some θ^* on the line segment between θ and θ_{lf}. Hence, if we assume without loss of generality that $\mathcal{N}$ is convex, the existence of m as described in (A4), ensures that for all $\theta \in \mathcal{N}$,

$$\left\| \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \log f_\theta(y) - \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\Big|_{\theta_{\mathrm{lf}}} \log f_\theta(y) \right\| \le C m(y) \|\theta - \theta_{\mathrm{lf}}\|,$$

for some $C > 0$. If we further, without loss of generality, assume that $\mathcal{N}$ is compact we can use the above equation and the triangle inequality to show

$$\left\| \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \log f_\theta(y) \right\| = \left\| \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\Big|_{\theta_0} \log f_\theta(y) \right\| + C' m(y),$$

for some positive $C' \in \mathbb{R}$. The right-hand side does not depend on θ and has finite expectation, hence all third-order partial derivatives of $\log f_\theta(y)$ are bounded by a fixed integrable function. This together with condition (A1) and (A2) of Lemma 3, ensures that $\hat\theta$ exists with probability tending to 1, $\hat\theta$ converges in probability to θ_{lf} and that

$$\hat\theta = \theta_{\mathrm{lf}} + (nJ)^{-1} \ell'_n(\theta_{\mathrm{lf}}) + o_{\mathrm{Pr}}(1). \tag{B1}$$

This follows from Thms. 5.41 and 5.42 in Van der Vaart (1998). Since $\hat\theta$ converges to θ_{lf} in probability, the probability of $\hat\theta$ lying in $\mathcal{N}$ goes to one. In the following, we will therefore assume that $\hat\theta \in \mathcal{N}$ without loss of generality.

To show Theorem 3 in the article we need to prove that the expected value of the remainder term $\epsilon_n(\hat\theta)$ is of order $O(1/\sqrt{n})$, and that each component of $W_n^j n(\hat\theta - \theta_{\mathrm{lf}})(\hat\theta - \theta_{\mathrm{lf}})^T$ and $\sqrt{n}(\hat\theta - \theta_{\mathrm{lf}})\sqrt{n}\left(J_n^j - J^j\right)^T$ is uniformly integrable. The rest of the argument is similar to that given in the article.

We start with the remainder term $\epsilon_n(\hat{\theta})$. Since m bounds all fourth order partial derivatives of $\log f_\theta(y)$ in $\mathcal{N}$ for almost all y, standard results concerning remainder terms in Taylor expansions can be used to show

$$|\epsilon_n(\hat{\theta})| \leq \frac{1}{3!}\sum_{i=1}^{n} m(Y_i)\|\hat{\theta} - \theta_{\text{lf}}\|_1^3.$$

Taking expectations on both sides of the inequality shows,

$$\mathrm{E}\sqrt{n}|\epsilon_n(\hat{\theta})| \leq \frac{1}{6}\mathrm{E}\left[m(Y)\|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^3\right].$$

Furthermore, applying Hölder's inequality twice, reveals

$$\mathrm{E}\sqrt{n}|\epsilon_n(\hat{\theta})| \leq \frac{1}{6}\left(\mathrm{E}m_2(Y)^4\right)^{1/4}\left(\mathrm{E}\|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^4\right)^{3/4}.$$

By (B1) $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) = O(1)$. Applying the continuous mapping theorem, ensures $n^2\|\hat{\theta} - \theta_{\text{lf}}\|_1^4 = O(1)$. Furthermore, uniform integrability of $n^2\|\hat{\theta} - \theta_{\text{lf}}\|_1^4$ can be shown using case (1) of lemma 3 and case (ii) and (iii) of Lemma 2. Hence, $\mathrm{E}\|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^4 = O(1)$, showing that $\mathrm{E}\epsilon_n(\hat{\theta}) = O(1/\sqrt{n})$.

Uniform integrability of each component of $W_n^j n(\hat{\theta} - \theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})^T$ holds true as long as the square of each component of W_n^j is uniformly integrable and each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ to the power of four is uniformly integrable. This follows from case (vii) of Lemma 2. Each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ to the power of four is uniformly integrable by (1) in Lemma 3. For W_n^j, notice that the square of each component of W_n^j is uniformly integrable by (3) in Lemma 3. Hence, $W_n^j n(\hat{\theta} - \theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})^T$ is uniformly integrable.

For the last case, notice that each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\sqrt{n}\left(J_n^j - J^j\right)^T$ is uniformly integrable by case (vii) in Lemma 2 if the square of each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ and $\sqrt{n}(J_n^j - J^j)$ is uniformly integrable. By by case (1) of Lemma 3 and case (iv) of Lemma 2, the square of each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ is uniformly integrable, while the square of each component of $\sqrt{n}(J_n^j - J^j)$ is uniformly integrable by case (2) of Lemma 3. ∎

To show consistency of the estimator of c defined in Theorem 3, we first state and prove a lemma.

Lemma 4. *Let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F, $\hat{\theta} \in \mathbb{R}^p$ a consistent estimator of $\theta_0 \in \mathbb{R}^p$ and $\psi : \mathbb{R}^{d+p} \to \mathbb{R}^q$ a function. Then $n^{-1}\sum_{i=1}^n \psi(Y_i, \hat{\theta})$ converges in probability to $\mathrm{E}\psi(Y, \theta_0)$ where $Y \sim F$, if there exists an F-integrable function $m : \mathbb{R}^{d+p} \to \mathbb{R}$ and a neighborhood $\mathcal{N}$ for θ_0 such that $\|\psi(y, \theta)\| \leq m(y)$ for F-almost all y and all $\theta \in \mathcal{N}$ and $\psi(y, \theta)$ is continuous in θ with probability 1 for all $\theta \in \mathcal{N}$.*

Proof. Since $\hat{\theta}$ is consistent for θ_0, there exists a compact subset K of $\mathcal{N}$ such that $\hat{\theta} \in K$ with probability tending to 1. We will assume $\hat{\theta} \in K \subseteq \mathcal{N}$, without loss of generality.

Since K is compact and ψ is continuous in θ with probability 1 on K, the existence of m as in the theorem, guarantees that the uniform law of large numbers holds true on

K. Hence, $\sup_{\theta \in K} \| n^{-1} \sum \psi(Y_i, \theta) - \mathrm{E}\psi(Y, \theta) \|$ converges in probability to 0. This implies that $\| n^{-1} \sum \psi(Y_i, \hat{\theta}) - \mathrm{E}\psi(Y, \theta_0) \| \xrightarrow{pr} 0$, showing the theorem. ∎

Theorem 12. *Let all quantities be defined as in Lemma 3, and assume that condition (A1–A4) of lemma 3 and Theorem 11 hold true. Then, replacing J, K, V_j, and W_j for $j = 1,\ldots,p$ in Theorem 3 with the empirical variances, covariance or means of the corresponding functions evaluated at $\hat{\theta}$ yields a consistent estimator of c.*

Proof. By the continuous mapping theorem, it suffices to show that the proposed estimators of J, K, V_j, and W_j for $j = 1,\ldots,p$ are consistent.

Arguing as in the beginning of the proof of Theorem 11, one can show that there exists functions $m_1(y)$, $m_2(y)$ and $m_3(y)$ with finite expectation bounding all first-, second-, and third-order partial derivatives of $\log f_\theta(y)$ in a neighborhood of θ_{lf}, respectively. This, combined with Lemma 4, ensures that the estimators for J and W_j for $j = 1,\ldots,p$ are consistent. The estimator of the matrix K is

$$\frac{1}{n} \sum_{i=1}^{n} \nabla \log f_{\hat{\theta}}(Y_i) \nabla \log f_{\hat{\theta}}(Y_i)^T.$$

The existence of m_1 guarantees that the norm of the above expression is bounded by $C m_1(y)^2$ for some $C > 0$. Direct computations show $m(y) = C_1 g(y) + C_2 m(Y)$ where $g(y)$ is some linear combination of first-, second-, and third-order partial derivatives of $\log f_\theta(y)$ at θ_{lf}. By condition (A1), $\mathrm{E}g(Y)^2$ exists and is finite and hence, Hölder's inequality implies

$$\mathrm{E}m_1(Y)^2 \leq \mathrm{E}g(Y)^2 + 2(\mathrm{E}g(Y)^2 \mathrm{E}m(Y)^2)^{1/2} + \mathrm{E}m(Y)^2.$$

Each term on the right-hand side of this equation is finite, showing that $\mathrm{E}m_1(Y)^2 < \infty$. By lemma 4, the estimator of K is consistent for the true value of the matrix. The argument for the estimators of V_j for each $j = 1,\ldots,p$ is similar and left out. ∎