Robust parameter estimation based on the $\mathcal{K}$-divergence: Supplementary Material

Yair Sorek and Koby Todros

Ben-Gurion University of the Negev

This supplementary material document is organized as follows. In Section I, we state and prove some basic properties of the $\mathcal{K}$-divergence. In Sections II and III, we show that, under some regularity conditions, the M$\mathcal{K}$DE is consistent, asymptotically normal and unbiased w.r.t. the maximizer of the $\mathcal{K}$-divergence over the hypothesized parametric class. In Section IV, we show that the M$\mathcal{K}$DE is Fisher consistent. The influence function of the M$\mathcal{K}$DE is derived in Section V. Finally, in Section VI, we present the fixed-point algorithms used for implementation of the M$\mathcal{K}$DE and the other compared estimators in Sec. 5.

I. Properties of the $\mathcal{K}$-divergence

The following theorem states some basic properties of the $\mathcal{K}$-divergence.

Theorem 1. *The $\mathcal{K}$-divergence (1) satisfies the following properties:*

1) ***Non-negativity:*** *$\mathcal{K}_h[G||F] \geq 0$, where equality holds if and only if $G = F$.*
2) ***Relation to the KLD:*** *assume that the expectation $\mathbb{E}[|\log \frac{g(\mathbf{x})}{f(\mathbf{x})}|; G]$ is finite. Then,*

$$\mathcal{K}_h[G||F] \xrightarrow[h\to\infty]{} \mathbb{E}\left[\log \frac{g(\mathbf{x})}{f(\mathbf{x})}; G\right] \triangleq \mathcal{K}_\infty[G||F], \tag{S-1}$$

3) ***Small bandwidth asymptotics:*** *assume that the expectation $\mathbb{E}[|\log \frac{g(\mathbf{x})}{f(\mathbf{x})}|^2; G]$ is finite and that the densities $g(\cdot)$ and $f(\cdot)$ are square integrable w.r.t. Lebesgue's measure λ. Then,*

$$\mathcal{K}_h[G||F] \xrightarrow[h\to 0]{} \mathbb{E}\left[\bar{\psi}_G(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x})}; G\right] + \log \mathbb{E}[\bar{\psi}_G(\mathbf{x}); F] \triangleq \mathcal{K}_0[G||F], \tag{S-2}$$

   where the weight function $\bar{\psi}_G(\mathbf{r}) \triangleq g(\mathbf{r})/\mathbb{E}[g(\mathbf{x}); G]$.
4) ***Invariance****: The $\mathcal{K}$-divergence is invariant to deterministic translation of $\mathbf{x}$. Furthermore, invariance under deterministic unitary transformation of $\mathbf{x}$ is guaranteed when the kernel function is spherical, i.e., $K(\mathbf{x}) = S(\|\mathbf{x}\|)$, where $S(\cdot)$ is a strictly positive scalar function.*

[Proofs of these properties appear in Subsections I-A-I-D below]

A. Non-negativity

In Lemma 1 below we show that under the assumption of a strictly positive kernel function

$$\mathcal{K}_h[G||F] = \mathcal{D}[\tilde{G}_h||\tilde{F}_h], \tag{S-3}$$

where $\mathcal{D}[\cdot||\cdot]$ denotes the KLD, $\tilde{G}_h$ and $\tilde{F}_h$ are probability distributions with density functions

$$\tilde{g}_h(\mathbf{r}) \triangleq g(\mathbf{r})\psi_G(\mathbf{r};h) \;\text{ and }\; \tilde{f}_h(\mathbf{r}) \triangleq f(\mathbf{r})\psi_F(\mathbf{r};h), \tag{S-4}$$

respectively, $\psi_G(\mathbf{r};h)$ is defined in Eq. (2) and $\psi_F(\mathbf{r};h) \triangleq \frac{(K_h*g)(\mathbf{r})}{\mathbb{E}[(K_h*g)(\mathbf{x});F]}$. We remark that the expectations $\mathbb{E}[(K_h*g)(\mathbf{x});G]$ in Eq. (2) and $\mathbb{E}[(K_h*g)(\mathbf{x});F]$ in $\psi_F(\mathbf{r};h)$ are strictly positive and finite. This is property follows from the boundedness and the strict positiveness of the kernel function $K(\cdot)$. By [s1, Th. 8.6.1] the KLD in (S-3) satisfies:

$$\mathcal{D}[\tilde{G}_h||\tilde{F}_h] \geq 0, \;\; \text{where equality holds} \iff \tilde{G}_h = \tilde{F}_h. \tag{S-5}$$

In Lemma 2, stated below, we prove that

$$\tilde{G}_h = \tilde{F}_h \iff G = F. \tag{S-6}$$

Hence, by (S-3), (S-5) and (S-6) it follows that $\mathcal{K}_h[G||F] \geq 0$, where equality holds if and only if $G = F$. □

Lemma 1. *The equality in (S-3) holds under the assumption of a strictly positive kernel function.*

Proof. The $\mathcal{K}$-divergence (1) can be rewritten as:

$$\mathcal{K}_h[G||F] = \mathbb{E}\left[\psi_G(\mathbf{x},h)\log\frac{g(\mathbf{x})\psi_G(\mathbf{x},h)}{f(\mathbf{x})\psi_F(\mathbf{x},h)};G\right] - \mathbb{E}\left[\psi_G(\mathbf{x},h)\log\frac{\psi_G(\mathbf{x},h)}{\psi_F(\mathbf{x},h)};G\right] + \log\mathbb{E}\left[\psi_G(\mathbf{x},h);F\right], \tag{S-7}$$

where $\psi_G(\mathbf{r};h)$ and $\psi_F(\mathbf{r};h)$ are defined in Eq. (2) and below Eq. (S-4), respectively. Note that since the kernel function is strictly positive, then $\psi_G(\mathbf{r};h)$ and $\psi_F(\mathbf{r};h)$ are strictly positive as well, which implies that the quotient $\psi_G(\mathbf{r},h)/\psi_F(\mathbf{r},h)$ is well defined. Also notice that $\tilde{g}_h(\mathbf{r}) \triangleq g(\mathbf{r})\psi_G(\mathbf{r};h)$ and $\tilde{f}_h(\mathbf{r}) \triangleq f(\mathbf{r})\psi_F(\mathbf{r};h)$ are non-negative functions that integrate to 1. Hence, these functions are viable density functions. Therefore, the expectation

$$\mathbb{E}\left[\psi_G(\mathbf{x},h)\log\frac{g(\mathbf{x})\psi_G(\mathbf{x},h)}{f(\mathbf{x})\psi_F(\mathbf{x},h)};G\right] = \mathbb{E}\left[\log\frac{\tilde{g}_h(\mathbf{x})}{\tilde{f}_h(\mathbf{x})};\tilde{G}_h\right] \triangleq \mathcal{D}[\tilde{G}_h||\tilde{F}_h], \tag{S-8}$$

where $\tilde{G}_h$ and $\tilde{F}_h$ are the probability distributions associated with $\tilde{g}_h(\mathbf{r})$ and $\tilde{f}_h(\mathbf{r})$, respectively. Additionally, using the definitions of $\psi_G(\mathbf{r};h)$ and $\psi_F(\mathbf{r};h)$, one can verify that

$$\mathbb{E}\left[\psi_G(\mathbf{x},h)\log\frac{\psi_G(\mathbf{x},h)}{\psi_F(\mathbf{x},h)};G\right] = -\log\mathbb{E}\left[\psi_G(\mathbf{x},h);F\right]. \tag{S-9}$$

Hence, the relation (S-3) follows directly from (S-7)-(S-9). □

Lemma 2. *The relation in (S-6) holds under the assumption of a strictly positive kernel function.*

Proof.

$$\begin{aligned}
\tilde{G}_h = \tilde{F}_h \quad &\iff \quad \tilde{g}_h(\mathbf{r}) = \tilde{f}_h(\mathbf{r}) \;\; \lambda - a.e., \\
&\overset{(a)}{\iff} \quad g(\mathbf{r})\psi_G(\mathbf{r};h) = f(\mathbf{r})\psi_F(\mathbf{r};h) \;\; \lambda - a.e., \\
&\overset{(b)}{\iff} \quad \frac{g(\mathbf{r})}{\mathbb{E}[(K_h * g)(\mathbf{x}); G]} = \frac{f(\mathbf{r})}{\mathbb{E}[(K_h * g)(\mathbf{x}); F]} \;\; \lambda - a.e., \\
&\overset{(c)}{\iff} \quad g(\mathbf{r}) = f(\mathbf{r}) \;\; \lambda - a.e., \\
&\iff \quad G = F,
\end{aligned} \tag{S-10}$$

where (a) follows from (S-4) and (b) stems from the definitions of $\psi_G(\mathbf{r};h)$ and $\psi_F(\mathbf{r};h)$ and the strict-positiveness of the convolution $(K_h * g)(\mathbf{r})$, that arise from the assumption that the kernel function $K(\cdot)$ is strictly positive. We shall now prove the equivalence in (c). If $\frac{g(\mathbf{r})}{\mathbb{E}[(K_h * g)(\mathbf{x}); G]} = \frac{f(\mathbf{r})}{\mathbb{E}[(K_h * g)(\mathbf{x}); F]}$ $\lambda - a.e.$, then integration of both sides of the equality yields $\mathbb{E}[(K_h * g)(\mathbf{x}); G] = \mathbb{E}[(K_h * g)(\mathbf{x}); F]$, implying that $g(\mathbf{r}) = f(\mathbf{r})$ $\lambda - a.e.$ The converse is trivial. □

B. Relation to the KLD

By Eqs. (1) and (2) it follows that

$$\lim_{h\to\infty} \mathcal{K}_h[G||F] = \frac{\lim\limits_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) g(\mathbf{r}) \log \frac{g(\mathbf{r})}{f(\mathbf{r})} d\lambda(\mathbf{r})}{\lim\limits_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r})} + \log \frac{\lim\limits_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) f(\mathbf{r}) d\lambda(\mathbf{r})}{\lim\limits_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r})}, \tag{S-11}$$

where $\eta_h(\mathbf{r}) \triangleq \int_{\mathbb{R}^p} K\left(\frac{\mathbf{r}-\boldsymbol{s}}{h}\right) g(\boldsymbol{s}) d\lambda(\boldsymbol{s})$. Since the kernel function $K(\cdot)$ is bounded, continuous and strictly positive it follows from the dominated convergence theorem (DCT) [s2, Cor. 2.3.12] that

$$\lim_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r}) = \lim_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) f(\mathbf{r}) d\lambda(\mathbf{r}) = K(0) > 0, \tag{S-12}$$

Additionally, if $\mathbb{E}[|\log \frac{g(\mathbf{x})}{f(\mathbf{x})}|; G]$ is finite, the DCT also implies that

$$\lim_{h\to\infty} \int_{\mathbb{R}^p} \eta_h(\mathbf{r}) g(\mathbf{r}) \log \frac{g(\mathbf{r})}{f(\mathbf{r})} d\lambda(\mathbf{r}) = K(0) \times \mathbb{E}\left[\log \frac{g(\mathbf{x})}{f(\mathbf{x})}; G\right]. \tag{S-13}$$

Therefore, the relation in Eq. (S-1) follows directly from (S-11), (S-12), (S-13) and the definition of the KLD. □

C. Small bandwidth asymptotics

By Eqs. (1) and (2) it follows that

$$\lim_{h\to 0} \mathcal{K}_h[G||F] = \frac{\lim\limits_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) \log \frac{g(\mathbf{r})}{f(\mathbf{r})} d\lambda(\mathbf{r})}{\lim\limits_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r})} + \log \frac{\lim\limits_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) f(\mathbf{r}) d\lambda(\mathbf{r})}{\lim\limits_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r})}, \tag{S-14}$$

where $\zeta_h(\mathbf{r}) \triangleq (K_h * g)(\mathbf{r})$. By the triangle and Cauchy-Schwartz's inequalities, it follows that:

$$\begin{aligned}
\left| \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r}) - \int_{\mathbb{R}^p} g^2(\mathbf{r}) d\lambda(\mathbf{r}) \right| \quad &= \quad \left| \int_{\mathbb{R}^p} (\zeta_h(\mathbf{r}) - g(\mathbf{r})) g(\mathbf{r}) d\lambda(\mathbf{r}) \right| \\
&\leq \quad \sqrt{\int_{\mathbb{R}^p} |\zeta_h(\mathbf{r}) - g(\mathbf{r})|^2 d\lambda(\mathbf{r})} \sqrt{\int_{\mathbb{R}^p} g^2(\mathbf{r}) d\lambda(\mathbf{r})}.
\end{aligned} \tag{S-15}$$

Therefore, since $K(\cdot)$ is integrable, $\int_{\mathbb{R}^p} K(\mathbf{r})d\lambda(\mathbf{r}) = 1$ and $g(\cdot)$ is square integrable, it follows from [s3, Th. 8.14] that

$$\lim_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) d\lambda(\mathbf{r}) = \int_{\mathbb{R}^p} g^2(\mathbf{r}) d\lambda(\mathbf{r}). \tag{S-16}$$

Similarly, since $f(\cdot)$ is square integrable and $\mathbb{E}[|\log \frac{g(\mathbf{x})}{f(\mathbf{x})}|^2; G] < \infty$ it can be shown that

$$\lim_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) f(\mathbf{r}) d\lambda(\mathbf{r}) = \int_{\mathbb{R}^p} g(\mathbf{r}) f(\mathbf{r}) d\lambda(\mathbf{r}). \tag{S-17}$$

and

$$\lim_{h\to 0} \int_{\mathbb{R}^p} \zeta_h(\mathbf{r}) g(\mathbf{r}) \log \frac{g(\mathbf{r})}{f(\mathbf{r})} d\lambda(\mathbf{r}) = \int_{\mathbb{R}^p} g^2(\mathbf{r}) \log \frac{g(\mathbf{r})}{f(\mathbf{r})} d\lambda(\mathbf{r}). \tag{S-18}$$

Hence, relation (S-2) follows directly from (S-14), (S-16)-(S-18) and the definition of $\bar{\psi}_G(\cdot)$ stated below (S-2). □

D. Invariance

Define the random vector:

$$\mathbf{y} \triangleq \mathbf{a}(\mathbf{x}), \tag{S-19}$$

where $\mathbf{a} : \mathbb{R}^p \to \mathbb{R}^p$ is a continuously differentiable one-to-one mapping. In the following, $G_{\mathbf{x}}$, $F_{\mathbf{x}}$, $g_{\mathbf{x}}(\cdot)$ and $f_{\mathbf{x}}(\cdot)$ will denote the probability distributions of $\mathbf{x}$ and their associate density functions. The probability distributions of $\mathbf{y}$ and their densities will be denoted by $G_{\mathbf{y}}$, $F_{\mathbf{y}}$, $g_{\mathbf{y}}(\cdot)$ and $f_{\mathbf{y}}(\cdot)$. In Lemma 3 below, we show that

$$\mathcal{K}_h[G_{\mathbf{y}}||F_{\mathbf{y}}] = \frac{\int_{\mathbb{R}^p} \zeta_{\mathbf{a},h}(\boldsymbol{t}) g_{\mathbf{x}}(\boldsymbol{t}) \log \frac{g_{\mathbf{x}}(\boldsymbol{t})}{f_{\mathbf{x}}(\boldsymbol{t})} d\lambda(\boldsymbol{t})}{\int_{\mathbb{R}^p} \zeta_{\mathbf{a},h}(\boldsymbol{t}) g_{\mathbf{x}}(\boldsymbol{t}) d\lambda(\boldsymbol{t})} + \log \frac{\int_{\mathbb{R}^p} \zeta_{\mathbf{a},h}(\boldsymbol{t}) f_{\mathbf{x}}(\boldsymbol{t}) d\lambda(\boldsymbol{t})}{\int_{\mathbb{R}^p} \zeta_{\mathbf{a},h}(\boldsymbol{t}) g_{\mathbf{x}}(\boldsymbol{t}) d\lambda(\boldsymbol{t})}, \tag{S-20}$$

where

$$\zeta_{\mathbf{a},h}(\boldsymbol{t}) \triangleq \int_{\mathbb{R}^p} K_h\left(\mathbf{a}(\boldsymbol{t}) - \mathbf{a}(\boldsymbol{\tau})\right) g_{\mathbf{x}}(\boldsymbol{\tau}) d\lambda(\boldsymbol{\tau}). \tag{S-21}$$

Hence, by Eqs. (1), (2), (S-20) and (S-21), it follows that the equality

$$\mathcal{K}_h[G_{\mathbf{y}}||F_{\mathbf{y}}] = \mathcal{K}_h[G_{\mathbf{x}}||F_{\mathbf{x}}] \tag{S-22}$$

holds when

$$K_h\left(\mathbf{a}(\boldsymbol{t}) - \mathbf{a}(\boldsymbol{\tau})\right) = K_h\left(\boldsymbol{t} - \boldsymbol{\tau}\right). \tag{S-23}$$

Notice that the relation in (S-23) is satisfied when $\mathbf{a}(\mathbf{x}) = \mathbf{x} + \mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^p$ is a deterministic shift parameter. Additionally, (S-23) holds when $\mathbf{a}(\mathbf{x}) = \mathbf{U}\mathbf{x}$, such that $\mathbf{U}$ is a deterministic unitary matrix, and the kernel function is spherical, i.e., $K(\mathbf{x}) = S(\|\mathbf{x}\|)$, where $S(\cdot)$ is a strictly positive scalar function. □

Lemma 3. *The relation in (S-20) holds under the transformation in (S-19).*

Proof. Let $\mathbf{b}(\cdot)$ denote the inverse of $\mathbf{a}(\cdot)$. By the change of variables formulae [s4, Ch. 2.2.5], it follows that

$$g_{\mathbf{y}}(\mathbf{r}) = g_{\mathbf{x}}(\mathbf{b}(\mathbf{r}))|\det[\nabla \mathbf{b}(\mathbf{r})]| \quad \text{and} \quad f_{\mathbf{y}}(\mathbf{r}) = f_{\mathbf{x}}(\mathbf{b}(\mathbf{r}))|\det[\nabla \mathbf{b}(\mathbf{r})]|. \tag{S-24}$$

Since $\mathbf{a}(\cdot)$ is continuously differentiable, then also $\mathbf{b}(\cdot)$. Therefore, let $\boldsymbol{t} \triangleq \mathbf{b}(\mathbf{r})$. By [s2, Th. 4.4.6], it follows that

$$|\det[\nabla \mathbf{b}(\mathbf{r})]| d\lambda(\mathbf{r}) = d\lambda(\boldsymbol{t}). \tag{S-25}$$

The relation in (S-20) can now be easily verified using (1), (S-24) and (S-25). □

II. Consistency

In the following theorem, we show that, under some regularity assumptions, $\hat{\boldsymbol{\theta}}_h$ (6) is a consistent estimator of $\boldsymbol{\theta}_h^*$ (7) that represents the best fitting parameter, in the sense of minimum $\mathcal{K}$-divergence.

Theorem 2 (Consistency)**.** *Assume that the following regularity conditions are satisfied:*

(A-1) The parameter space $\boldsymbol{\Theta}$ is compact.

(A-2) The $\mathcal{K}$-divergence $\mathcal{K}_h[G||F_{\boldsymbol{\theta}}]$ has a unique minimum over $\boldsymbol{\Theta}$.

(A-3) The expectation $\mathbb{E}[K_h(\mathbf{x}-\mathbf{r}); F_{\boldsymbol{\theta}}]$ is continuous over $\boldsymbol{\Theta}$ λ-a.e.

(A-4) The expectation $\mathbb{E}[\psi_G(\mathbf{x},h)|\log f(\mathbf{x};\boldsymbol{\theta})|; G]$ is finite.

(A-5) There exist functions $u:\mathbb{R}^p\to\mathbb{R}_+$ and $v:\mathbb{R}_+\to\mathbb{R}_+$, such that $\mathbb{E}[\psi_G(\mathbf{x},h)u(\mathbf{x}); G]<\infty$, $v(\cdot)$ is continuous at the origin, $v(0)=0$ and for any $\boldsymbol{\theta},\boldsymbol{\theta}'\in\boldsymbol{\Theta}$, it holds that

$$|\log\rho(\mathbf{r},\boldsymbol{\theta},\boldsymbol{\theta}')|\le u(\mathbf{r})v(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|)\ \ \lambda-a.e., \tag{S-26}$$

where $\rho(\mathbf{r},\boldsymbol{\theta},\boldsymbol{\theta}')\triangleq f(\mathbf{r};\boldsymbol{\theta})/f(\mathbf{r};\boldsymbol{\theta}')$.

Then,

$$\hat{\boldsymbol{\theta}}_h\xrightarrow[N\to\infty]{p}\boldsymbol{\theta}_h^*, \tag{S-27}$$

where "$\xrightarrow{p}$" denotes convergence in probability [s2].

Proof. By Eqs. (1) and (2), it follows that minimization of $\mathcal{K}_h[G||F_{\boldsymbol{\theta}}]$ over $\boldsymbol{\Theta}$ amounts to maximization of the deterministic objective function:

$$\bar{\mathcal{J}}_h(\boldsymbol{\theta})\triangleq\mathbb{E}[\psi_G(\mathbf{x},h)\log f(\mathbf{x};\boldsymbol{\theta}); G]-\log\mathbb{E}[(K_h*g)(\mathbf{x}); F_{\boldsymbol{\theta}}]. \tag{S-28}$$

Therefore, under Assumption (A-2), $\bar{\mathcal{J}}_h(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}=\boldsymbol{\theta}_h^*$ (7). In Proposition 1, stated below, we show that under (A-1) and (A-3)-(A-5), the random objective function $\mathcal{J}_h(\boldsymbol{\theta})$ (3) converges uniformly in probability to $\bar{\mathcal{J}}_h(\boldsymbol{\theta})$ as $N\to\infty$. Furthermore, Assumptions (A-3) and (A-5) imply that $\mathcal{J}_h(\boldsymbol{\theta})$ must be continuous over the compact parameter space $\boldsymbol{\Theta}$ w.p.1. Hence, by [s5, Th. 3.4] we conclude that the convergence in (S-27) holds. □

Proposition 1. *Assume that conditions (A-1) and (A-3)-(A-5) are satisfied. Then,*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|\mathcal{J}_h(\boldsymbol{\theta})-\bar{\mathcal{J}}_h(\boldsymbol{\theta})|\xrightarrow[N\to\infty]{p}0. \tag{S-29}$$

Proof. Using Eqs. (7), (S-28) and the triangle inequality, one can verify that:

$$|\mathcal{J}_h(\boldsymbol{\theta})-\bar{\mathcal{J}}_h(\boldsymbol{\theta})|\le A(\boldsymbol{\theta})+B(\boldsymbol{\theta}), \tag{S-30}$$

where

$$A(\boldsymbol{\theta})\triangleq\left|\frac{\sum_{n=1}^N\sum_{m\neq n}^N K_h(\mathbf{x}_n-\mathbf{x}_m)\log f(\mathbf{x}_n;\boldsymbol{\theta})}{\sum_{n=1}^N\sum_{m\neq n}^N K_h(\mathbf{x}_n-\mathbf{x}_m)}-\frac{\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\log f(\mathbf{x};\boldsymbol{\theta}); G\times G]}{\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}'); G\times G]}\right|, \tag{S-31}$$

$\mathbf{x}'$ is an independent copy of $\mathbf{x}$, such that the product $G\times G$ denotes their joint probability distribution,

$$B(\boldsymbol{\theta})\triangleq|\log T(\boldsymbol{\theta})|, \tag{S-32}$$

$$T(\boldsymbol{\theta}) \triangleq \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{c(\mathbf{x}_n;\boldsymbol{\theta})}{\mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G]} \tag{S-33}$$

and

$$c(\mathbf{r};\boldsymbol{\theta}) \triangleq \mathbb{E}[K_h(\mathbf{x}''-\mathbf{r});F_{\boldsymbol{\theta}}]. \tag{S-34}$$

Note that since $K_h(\cdot)$ is strictly positive, the statistic $T(\boldsymbol{\theta})$ (S-33) must be strictly positive over $\boldsymbol{\Theta}$ w.p.1. In Lemmas 4 and 5, stated below, we show that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} A(\boldsymbol{\theta}) \xrightarrow[N\to\infty]{p} 0 \tag{S-35}$$

and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} B(\boldsymbol{\theta}) \xrightarrow[N\to\infty]{p} 0. \tag{S-36}$$

Hence, the relation in (S-29) follows directly from (S-30), (S-35) and (S-36). □

Lemma 4. *Relation (S-35) holds under Assumptions (A-1), (A-4) and (A-5).*

Proof. Using (S-31) and the triangle inequality, one can verify that

$$A(\boldsymbol{\theta}) = |A_1(\boldsymbol{\theta}) + A_2(\boldsymbol{\theta})| \le |A_1(\boldsymbol{\theta})| + |A_2(\boldsymbol{\theta})|, \tag{S-37}$$

where

$$A_1(\boldsymbol{\theta}) \triangleq \frac{\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}^{N} Z_h(\mathbf{x}_n,\mathbf{x}_m;\boldsymbol{\theta}) - \mathbb{E}[Z_h(\mathbf{x},\mathbf{x}';\boldsymbol{\theta});G\times G]}{\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m)}, \tag{S-38}$$

$$A_2(\boldsymbol{\theta}) \triangleq \mathbb{E}[Z_h(\mathbf{x},\mathbf{x}';\boldsymbol{\theta});G\times G]\left(\frac{1}{\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m)} - \frac{1}{\mathbb{E}\left[K_h(\mathbf{x}-\mathbf{x}');G\times G\right]}\right), \tag{S-39}$$

and

$$Z_h(\mathbf{r},\boldsymbol{s};\boldsymbol{\theta}) \triangleq K_h(\mathbf{r}-\boldsymbol{s})\log f(\mathbf{r};\boldsymbol{\theta}). \tag{S-40}$$

Since $K_h(\cdot)$ is symmetric and bounded, it follows from [s6, Th. 5.4.A] that

$$\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m) \xrightarrow[N\to\infty]{w.p.1} \mathbb{E}\left[K_h(\mathbf{x}-\mathbf{x}');G\times G\right]. \tag{S-41}$$

Furthermore, the series

$$\frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{m\neq n}^{N} Z_h(\mathbf{x}_n,\mathbf{x}_m;\boldsymbol{\theta}) = \frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{m=n+1}^{N} W_h(\mathbf{x}_n,\mathbf{x}_m;\boldsymbol{\theta}), \tag{S-42}$$

where

$$W_h(\mathbf{r},\boldsymbol{s};\boldsymbol{\theta}) \triangleq \frac{1}{2}\left(Z_h(\mathbf{r},\boldsymbol{s};\boldsymbol{\theta}) + Z_h(\boldsymbol{s},\mathbf{r};\boldsymbol{\theta})\right) \tag{S-43}$$

is a symmetrized version of $Z_h(\mathbf{r},\boldsymbol{s};\boldsymbol{\theta})$, i.e., $W_h(\mathbf{r},\boldsymbol{s};\boldsymbol{\theta}) = W_h(\boldsymbol{s},\mathbf{r};\boldsymbol{\theta})$, and

$$\mathbb{E}[W_h(\mathbf{x},\mathbf{x}';\boldsymbol{\theta});G\times G] = \mathbb{E}[Z_h(\mathbf{x},\mathbf{x}';\boldsymbol{\theta});G\times G]. \tag{S-44}$$

We shall assume that the expectation

$$\mathbb{E}[|W_h(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})|;G\times G] < \infty \tag{S-45}$$

and that there exist functions $b : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ and $v : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\mathbb{E}[b(\mathbf{x}, \mathbf{x}'); G \times G] < \infty$, $v(\cdot)$ is continuous at the origin, $v(0) = 0$ and for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, it holds that

$$|W_h(\mathbf{r}, \boldsymbol{s}; \boldsymbol{\theta}) - W_h(\mathbf{r}, \boldsymbol{s}; \boldsymbol{\theta}')| \leq b(\mathbf{r}, \boldsymbol{s}) v(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|) \quad \lambda \times \lambda - a.e. \tag{S-46}$$

Under these assumptions, it follows from [s7, Cor. 4.1] that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{2}{N(N-1)} \sum\nolimits_{n=1}^{N} \sum\nolimits_{m=n+1}^{N} W_h(\mathbf{x}_n, \mathbf{x}_m; \boldsymbol{\theta}) - \mathbb{E}[W_h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}); G \times G] \right| \xrightarrow[N \to \infty]{p} 0. \tag{S-47}$$

and $\mathbb{E}[W_h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}); G \times G]$ is continuous over $\boldsymbol{\Theta}$. Hence, it follows from (S-42), (S-44) and (S-47) that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{N(N-1)} \sum\nolimits_{n=1}^{N} \sum\nolimits_{m=n+1}^{N} Z_h(\mathbf{x}_n, \mathbf{x}_m; \boldsymbol{\theta}) - \mathbb{E}[Z_h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}); G \times G] \right| \xrightarrow[N \to \infty]{p} 0. \tag{S-48}$$

and $\mathbb{E}[Z_h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}); G \times G]$ is continuous over $\boldsymbol{\Theta}$. Therefore, by (S-38), (S-39), (S-41), (S-48), the continuity of $\mathbb{E}[Z_h(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}); G \times G]$, the compactness of $\boldsymbol{\Theta}$ and Mann-Wald's Theorem , we conclude that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |A_1(\boldsymbol{\theta})| \xrightarrow[N \to \infty]{p} 0 \quad \text{and} \quad \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |A_2(\boldsymbol{\theta})| \xrightarrow[N \to \infty]{p} 0. \tag{S-49}$$

Hence, the relation in (S-35) follows directly from (S-37) and (S-49).

To complete the proof, we need to show now that the conditions in (S-45) and (S-46) hold. First, using (S-40) and (S-44) one can easily verify that since the kernel function $K_h(\cdot)$ is strictly positive and bounded, the condition in (S-45) is satisfied under Assumption (A-4). Additionally, by (S-40) and (S-43) it follows that

$$\begin{aligned} |W_h(\mathbf{r}, \boldsymbol{s}; \boldsymbol{\theta}) - W_h(\mathbf{r}, \boldsymbol{s}; \boldsymbol{\theta}')| &= \frac{1}{2} K_h(\mathbf{r} - \boldsymbol{s}) |\log \rho(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\theta}') + \log \rho(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\theta}')| \\ &\leq \frac{1}{2} K_h(\mathbf{r} - \boldsymbol{s}) |\log \rho(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\theta}')| + \frac{1}{2} K_h(\mathbf{r} - \boldsymbol{s}) |\log \rho(\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\theta}')| \\ &\leq \frac{1}{2} K_h(\mathbf{r} - \boldsymbol{s})(u(\mathbf{r}) + u(\boldsymbol{s})) v(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|), \end{aligned} \tag{S-50}$$

where $\rho(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq f(\mathbf{r}; \boldsymbol{\theta}) / f(\mathbf{r}; \boldsymbol{\theta}')$ and the last inequality in (S-50) follows from Assumption (A-5). Hence, the function $b(\cdot, \cdot)$ in (S-50) is given by:

$$b(\mathbf{r}, \boldsymbol{s}) = \frac{1}{2} K_h(\mathbf{r} - \boldsymbol{s})(u(\mathbf{r}) + u(\boldsymbol{s})). \tag{S-51}$$

We need to show that in this case the expectation $\mathbb{E}[b(\mathbf{x}, \mathbf{x}'); G \times G]$ is indeed finite. In Assumption (A-5) it is stated that $\mathbb{E}[\psi_G(\mathbf{x}, h) u(\mathbf{x}); G] < \infty$. Hence, by Eq. (2) it follows that $\mathbb{E}[b(\mathbf{x}, \mathbf{x}'); G \times G] = \mathbb{E}[K_h(\mathbf{x} - \mathbf{x}') u(\mathbf{x}); G \times G] < \infty$ when $\mathbb{E}[\psi_G(\mathbf{x}, h) u(\mathbf{x}); G] < \infty$. □

Lemma 5. *Relation (S-36) holds under assumptions Assumptions (A-1) and (A-3)*

Proof. Let $\mu_T(\boldsymbol{\theta}) \triangleq \mathbb{E}[T(\boldsymbol{\theta}); P_{T(\boldsymbol{\theta})}]$. By (S-33), and the assumption that $\{\mathbf{x}_n\}_{n=1}^N$ are identically distributed, we have that $\mu_T(\boldsymbol{\theta}) = 1$. Hence, the mean-value theorem [s8, Th. 3.4] implies that

$$\log T(\boldsymbol{\theta}) = \log \mu_T(\boldsymbol{\theta}) + \left. \frac{d \log T}{dT} \right|_{T = T^*(\boldsymbol{\theta})} (T(\boldsymbol{\theta}) - \mu_T(\boldsymbol{\theta})) = \frac{T(\boldsymbol{\theta}) - 1}{T^*(\boldsymbol{\theta})}, \tag{S-52}$$

where $T^*(\boldsymbol{\theta})$ lies in the line segment connecting $T(\boldsymbol{\theta})$ and $\mu_T(\boldsymbol{\theta}) = 1$. Note that since $T(\boldsymbol{\theta}) > 0$ w.p. 1, then $T^*(\boldsymbol{\theta})$ must be strictly positive w.p. 1. Therefore, by (S-32) and (S-52)

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} B(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |T(\boldsymbol{\theta}) - 1| \times (\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} T^*(\boldsymbol{\theta}))^{-1} \tag{S-53}$$

where $(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} T^*(\boldsymbol{\theta}))^{-1} < \infty$ w.p.1. Using (S-33), one can verify that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |T(\boldsymbol{\theta}) - 1| \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{\mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G]} \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{N}\sum_{n=1}^{N} c(\mathbf{x}_n;\boldsymbol{\theta}) - \mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G] \right|. \tag{S-54}$$

Under Assumption (A-3), the function $c(\mathbf{r};\boldsymbol{\theta})$ (S-34) is continuous over $\boldsymbol{\Theta}$ λ-a.e. Additionally, since the kernel function $K_h(\cdot)$ is bounded, it follows from (S-34) that $c(\mathbf{r};\boldsymbol{\theta})$ is bounded over $\mathbb{R}^p \times \boldsymbol{\Theta}$. Hence, since $\{\mathbf{x}_n\}_{n=1}^N$ are i.i.d. samples from G, it follows from the uniform weak law of large numbers [s9] that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{N}\sum_{n=1}^{N} c(\mathbf{x}_n;\boldsymbol{\theta}) - \mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G] \right| \xrightarrow[N\to\infty]{p} 0. \tag{S-55}$$

and that $\mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G]$ is continuous over the parameter space $\boldsymbol{\Theta}$. The latter continuity property, and Assumption (A-1) (compactness of $\boldsymbol{\Theta}$) imply that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{\mathbb{E}[c(\mathbf{x};\boldsymbol{\theta});G]} < \infty. \tag{S-56}$$

Hence, by (S-54)-(S-56) we conclude that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |T(\boldsymbol{\theta}) - 1| \xrightarrow[N\to\infty]{p} 0. \tag{S-57}$$

We now analyze the convergence of $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} T^*(\boldsymbol{\theta})$ in (S-53). Note that

$$|\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} T^*(\boldsymbol{\theta}) - 1| \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |T^*(\boldsymbol{\theta}) - 1| \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |T(\boldsymbol{\theta}) - 1|, \tag{S-58}$$

where the second inequality in (S-58) follows from the property that $T^*(\boldsymbol{\theta})$ lies in the line segment connecting $T(\boldsymbol{\theta})$ and $\mu_T(\boldsymbol{\theta}) = 1$. Hence, by (S-57) we conclude that

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} T^*(\boldsymbol{\theta}) \xrightarrow[N\to\infty]{p} 1. \tag{S-59}$$

Therefore, the relation in (S-36) follows directly from (S-53), (S-57), (S-59) and Mann-Wald's Theorem [s10]. □

III. Asymptotic normality and unbiasedness

In the following theorem, we show that, under some regularity conditions, $\hat{\boldsymbol{\theta}}_h$ (6) is asymptotically normal and unbiased w.r.t. $\boldsymbol{\theta}_h^*$ (7) that represents the best fitting parameter, in the sense of minimum $\mathcal{K}$-divergence. Furthermore, we obtain a closed form expression for the asymptotic MSE matrix.

Theorem 3 (Asymptotic normality and unbiasedness)**.** *Assume that the following regularity conditions hold:*

(B-1) $\hat{\boldsymbol{\theta}}_h \xrightarrow[N\to\infty]{p} \boldsymbol{\theta}_h^*$, *where*

(B-2) $\boldsymbol{\theta}_h^*$ *lies in the interior of* $\boldsymbol{\Theta}$ *which is assumed to be compact.*

(B-3) The density function $f(\mathbf{r};\boldsymbol{\theta})$ *is twice continuously differentiable over* $\boldsymbol{\Theta}$ λ*-a.e.*

(B-4) There exist λ*-integrable functions* $a(\mathbf{r})$, $\{b_k(\mathbf{r})\}_{k=1}^p$ *and* $\{c_{k,j}(\mathbf{r})\}_{k,j=1}^p$, *such that* $f(\mathbf{r};\boldsymbol{\theta}) \leq a(\mathbf{r})$, $|\partial f(\mathbf{r};\boldsymbol{\theta})/\partial\theta_k| \leq b_k(\mathbf{r})$, $k = 1,\ldots,p$, *and* $|\partial^2 f(\mathbf{r};\boldsymbol{\theta})/\partial\theta_k\partial\theta_j| \leq c_{k,j}(\mathbf{r})$, $k,j = 1,\ldots,p$, *where* θ_k *denotes the* k*-th coordinate of* $\boldsymbol{\theta}$.

(B-5) The expectations $\mathbb{E}[\xi_h(\mathbf{x})\|\mathbf{q}(\mathbf{x};\boldsymbol{\theta})\|^2;G]$ *and* $\mathbb{E}[\xi_h(\mathbf{x})\|\mathbf{q}(\mathbf{x};\boldsymbol{\theta})\|^2;F_{\boldsymbol{\theta}}]$ *are finite over the parameter space* $\boldsymbol{\Theta}$, *where* $\xi_h(\mathbf{r}) \triangleq (K_h^2 * g)(\mathbf{r})$ *and* $\mathbf{q}(\mathbf{r};\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} \log f(\mathbf{r};\boldsymbol{\theta})$ *is the score-function.*

(B-6) *The expectations* $\{\mathbb{E}[\zeta_h(\mathbf{x})|[\mathbf{H}(\mathbf{x};\boldsymbol{\theta})]_{k,j}|;G]\}_{k,j=1}^{p}$ *are finite over the parameter space* $\boldsymbol{\Theta}$*, where* $\zeta_h(\mathbf{r}) \triangleq (K_h * g)(\mathbf{r})$ *and* $\mathbf{H}(\mathbf{r};\boldsymbol{\theta}) \triangleq \nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{r};\boldsymbol{\theta})$ *is the Hessian of the log-likelihood function.*

(B-7) *For any* $k \in \{1,\ldots,p\}$*, there exist functions* $u_k : \mathbb{R}^p \to \mathbb{R}_+$ *and* $v_k : \mathbb{R}_+ \to \mathbb{R}_+$*, such that* $\mathbb{E}[\psi_G(\mathbf{x},h)u_k(\mathbf{x});G] < \infty$*,* $v_k(\cdot)$ *is continuous at the origin,* $v_k(0) = 0$ *and for any* $\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}$*, it holds that*

$$|\Delta_{\mathbf{q},k}(\mathbf{r};\boldsymbol{\theta},\boldsymbol{\theta}')| \leq u_k(\mathbf{r})v_k(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|) \;\; \lambda - a.e., \tag{S-60}$$

where $\Delta_{\mathbf{q},k}(\mathbf{r};\boldsymbol{\theta},\boldsymbol{\theta}') \triangleq [\mathbf{q}(\mathbf{x};\boldsymbol{\theta})]_k - [\mathbf{q}(\mathbf{x};\boldsymbol{\theta}')]_k$.

(B-8) *For any* $k,j \in \{1,\ldots,p\}$*, there exist functions* $u_{k,j} : \mathbb{R}^p \to \mathbb{R}_+$ *and* $v_{k,j} : \mathbb{R}_+ \to \mathbb{R}_+$*, such that* $\mathbb{E}[\psi_G(\mathbf{x},h)u_{k,j}(\mathbf{x});G] < \infty$*,* $v_{k,j}(\cdot)$ *is continuous at the origin,* $v_{k,j}(0) = 0$ *and for any* $\boldsymbol{\theta},\boldsymbol{\theta}' \in \boldsymbol{\Theta}$*, it holds that*

$$|\Delta_{\mathbf{H},k,j}(\mathbf{r};\boldsymbol{\theta},\boldsymbol{\theta}')| \leq u_{k,j}(\mathbf{r})v_{k,j}(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|) \;\; \lambda - a.e., \tag{S-61}$$

where $\Delta_{\mathbf{H},k,j}(\mathbf{r};\boldsymbol{\theta},\boldsymbol{\theta}') \triangleq [\mathbf{H}(\mathbf{x};\boldsymbol{\theta})]_{k,j} - [\mathbf{H}(\mathbf{x};\boldsymbol{\theta}')]_{k,j}$.

Then,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*) \xrightarrow[N\to\infty]{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_h^*, h)\right), \tag{S-62}$$

where "$\xrightarrow{d}$*" denotes convergence in distribution [s2]. The covariance matrix in (S-62) takes the form:*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta},h) = \mathbf{C}^{-1}(\boldsymbol{\theta},h)\mathbf{D}(\boldsymbol{\theta},h)\mathbf{C}^{-1}(\boldsymbol{\theta},h), \tag{S-63}$$

where

$$\mathbf{C}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\psi_G(\mathbf{x},h)\nabla^2_{\boldsymbol{\theta}} \log f(\mathbf{x};\boldsymbol{\theta});G] - \nabla^2_{\boldsymbol{\theta}} \log u(\boldsymbol{\theta},h),$$

$$\mathbf{D}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\mathbf{v}(\mathbf{x},\boldsymbol{\theta},h)\mathbf{v}^T(\mathbf{x},\boldsymbol{\theta},h);G],$$

$$\mathbf{v}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \psi_G(\mathbf{r},h)\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h) + \mathbf{d}(\mathbf{r},\boldsymbol{\theta},h) - \mathbf{z}(\mathbf{r},\boldsymbol{\theta},h),$$

$$\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \nabla_{\boldsymbol{\theta}} \log f(\mathbf{r};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log u(\boldsymbol{\theta},h),$$

$$\mathbf{d}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \mathbb{E}[\mathbf{c}(\mathbf{x},\boldsymbol{\theta},h)\varphi_h(\mathbf{r}-\mathbf{x});G],$$

$$\mathbf{z}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)}\nabla_{\boldsymbol{\theta}} \log \frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)},$$

$$v(\mathbf{r},\boldsymbol{\theta},h) \triangleq (K_h * f)(\mathbf{r};\boldsymbol{\theta}), \;\; u(\boldsymbol{\theta},h) \triangleq \mathbb{E}[(K_h * g)(\mathbf{x});F_{\boldsymbol{\theta}}] \;\; \text{and} \;\; \varphi_h(\mathbf{r}) \triangleq K_h(\mathbf{r})/\mathbb{E}[(K_h * g)(\mathbf{x});G]$$

Proof. We begin by stating the following remark.

Remark 1. *Throughout the proof we shall assume that integration and differentiation operations can be interchanged. Using [s11, Th. 2.40] it can be shown that this assumption is justified under conditions (B-3), (B-4) and the boundedness of the kernel function* $K_h(\cdot)$.

We now proceed to the proof. By Assumptions (B-1) and (B-2) the estimator $\hat{\boldsymbol{\theta}}_h$ is weakly consistent and the estimand $\boldsymbol{\theta}_h^*$ lies in the interior of $\boldsymbol{\Theta}$, which is assumed to be compact. Therefore, $\hat{\boldsymbol{\theta}}_h$ lies in an open neighbourhood $\mathcal{U} \subset \boldsymbol{\Theta}$ of $\boldsymbol{\theta}_h^*$ with sufficiently high probability as N gets large, i.e., it does not lie on the boundary of $\boldsymbol{\Theta}$. Hence, $\hat{\boldsymbol{\theta}}_h$ is a maximum point of the objective function $\mathcal{J}_h(\boldsymbol{\theta})$ (7) whose gradient satisfies:

$$\nabla \mathcal{J}_h(\hat{\boldsymbol{\theta}}_h) = \mathbf{0}. \tag{S-64}$$

Using (3), one can verify that

$$\nabla \mathcal{J}_h(\boldsymbol{\theta}) = b^{-1}(\boldsymbol{\theta})\mathbf{a}(\boldsymbol{\theta}), \tag{S-65}$$

where

$$\mathbf{a}(\boldsymbol{\theta}) \triangleq \frac{1}{(N-1)N^2}\sum_{n=1}^{N}\sum_{m\neq n}^{N}\sum_{l=1}^{N} K_h(\mathbf{x}_n - \mathbf{x}_m)\mathbf{h}(\mathbf{x}_n, \mathbf{x}_l; \boldsymbol{\theta}), \tag{S-66}$$

$$b(\boldsymbol{\theta}) \triangleq \frac{1}{(N-1)N^2}\sum_{n=1}^{N}\sum_{m\neq n}^{N}\sum_{l=1}^{N} K_h(\mathbf{x}_n - \mathbf{x}_m)d(\mathbf{x}_l; \boldsymbol{\theta}), \tag{S-67}$$

$$\mathbf{h}(\mathbf{r}, \boldsymbol{t}; \boldsymbol{\theta}) \triangleq \mathbf{q}(\mathbf{r}; \boldsymbol{\theta})d(\boldsymbol{t}; \boldsymbol{\theta}) - \mathbf{w}(\boldsymbol{t}; \boldsymbol{\theta}), \tag{S-68}$$

$$d(\boldsymbol{t}; \boldsymbol{\theta}) \triangleq \mathbb{E}[K_h(\mathbf{x} - \boldsymbol{t}); F_{\boldsymbol{\theta}}], \tag{S-69}$$

$$\mathbf{w}(\boldsymbol{t}; \boldsymbol{\theta}) \triangleq \mathbb{E}[K_h(\mathbf{x} - \boldsymbol{t})\mathbf{q}(\mathbf{x}; \boldsymbol{\theta}); F_{\boldsymbol{\theta}}] \tag{S-70}$$

and $\mathbf{q}(\boldsymbol{\theta})$ is the score function defined in (B-5). Therefore, by the strict positivity of the kernel function $K_h(\cdot)$, it follows that the equality in (S-64) is equivalent to

$$\mathbf{a}(\hat{\boldsymbol{\theta}}_h) = \mathbf{0}. \tag{S-71}$$

Using Assumptions (B-3), (B-4) and the dominated convergence Theorem [s2, Cor. 2.3.12], one can verify that $\mathbf{a}(\boldsymbol{\theta})$ is continuous over $\boldsymbol{\Theta}$ w.p.1. Hence, the mean-value theorem [s8, Th. 3.4] implies that

$$\mathbf{0} = \mathbf{a}(\hat{\boldsymbol{\theta}}_h) = \mathbf{a}(\boldsymbol{\theta}_h^*) + \mathbf{F}(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*), \tag{S-72}$$

where

$$\mathbf{F}(\boldsymbol{\theta}) \triangleq \frac{d\mathbf{a}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{1}{(N-1)N^2}\sum_{n=1}^{N}\sum_{m\neq n}^{N}\sum_{l=1}^{N} K_h(\mathbf{x}_n - \mathbf{x}_m)\frac{d\mathbf{h}(\mathbf{x}_n, \mathbf{x}_l; \boldsymbol{\theta})}{d\boldsymbol{\theta}} \tag{S-73}$$

and $\tilde{\boldsymbol{\theta}}$ lies in the line segment connecting $\hat{\boldsymbol{\theta}}_h$ and $\boldsymbol{\theta}_h^*$.

In Proposition 2, stated below, we show that

$$\mathbf{F}(\tilde{\boldsymbol{\theta}}) \xrightarrow[N\to\infty]{p} \eta(\boldsymbol{\theta}_h^*, h)\mathbf{C}(\boldsymbol{\theta}_h^*, h), \tag{S-74}$$

where

$$\eta(\boldsymbol{\theta}, h) \triangleq \mathbb{E}[\psi_G(\mathbf{x}, h); F_{\boldsymbol{\theta}}] \times \mathbb{E}^2[K_h(\mathbf{x} - \mathbf{x}'); G \times G] \tag{S-75}$$

and the matrix function $\mathbf{C}(\boldsymbol{\theta}, h)$, defined below Eq. (S-63), is non-singular by assumption. Note that strict positivity and finiteness of $\eta(\boldsymbol{\theta}_h^*, h)$ follows from the assumption that the kernel function $K_h(\cdot)$ is strictly positive and bounded. Hence, by Mann-Wald's Theorem [s10]

$$\mathbf{F}^{-1}(\tilde{\boldsymbol{\theta}}) \xrightarrow[N\to\infty]{p} \eta^{-1}(\boldsymbol{\theta}_h^*, h)\mathbf{C}^{-1}(\boldsymbol{\theta}_h^*, h), \tag{S-76}$$

which implies that $\mathbf{F}(\tilde{\boldsymbol{\theta}})$ is invertible with sufficiently high probability as N gets large. Therefore, by (S-72) the equality

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*) = -\mathbf{F}^{-1}(\tilde{\boldsymbol{\theta}})\sqrt{N}\mathbf{a}(\boldsymbol{\theta}_h^*) \tag{S-77}$$

holds with sufficiently large probability as N gets large. Furthermore, by Proposition 3 stated below

$$\sqrt{N}\mathbf{a}(\boldsymbol{\theta}_h^*) \xrightarrow[N\to\infty]{d} \mathcal{N}(\mathbf{0}, \eta^2(\boldsymbol{\theta}_h^*, h)\mathbf{D}(\boldsymbol{\theta}_h^*, h)), \tag{S-78}$$

where $\mathbf{D}(\boldsymbol{\theta}, h)$ is defined below Eq. (S-63). Thus, by (S-76)-(S-78) and Slutskey's Theorem, the relation in (S-62) holds. □

Proposition 2. *The relation in (S-74) holds under assumptions (B-1)-(B-8).*

Proof. First, in Lemma 6 stated below we prove that

$$\mathbf{F}(\tilde{\boldsymbol{\theta}}) \xrightarrow[N\to\infty]{p} \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*), \tag{S-79}$$

where

$$\bar{\mathbf{F}}(\boldsymbol{\theta}) \triangleq \mathbb{E}\left[K_h(\mathbf{x}-\mathbf{x}')\frac{d\mathbf{h}(\mathbf{x},\mathbf{x}'';\boldsymbol{\theta})}{d\boldsymbol{\theta}}; G\times G\times G\right]. \tag{S-80}$$

Next, in Lemma 8, we show that

$$\bar{\mathbf{F}}(\boldsymbol{\theta}_h^*) = \eta(\boldsymbol{\theta}_h^*, h)\mathbf{C}(\boldsymbol{\theta}_h^*, h). \tag{S-81}$$

□

Lemma 6. *The relation in (S-79) holds under assumptions (B-1)-(B-8).*

Proof. Notice that

$$\begin{aligned}\|\mathbf{F}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)\| &= \|\mathbf{F}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}}) + \bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)\| \\ &\le \|\mathbf{F}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}})\| + \|\bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)\| \\ &\le \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{F}(\boldsymbol{\theta}) - \bar{\mathbf{F}}(\boldsymbol{\theta})\| + \|\bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)\|.\end{aligned} \tag{S-82}$$

In Lemma 7, stated below, we show that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{F}(\boldsymbol{\theta}) - \bar{\mathbf{F}}(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0 \tag{S-83}$$

and that $\bar{\mathbf{F}}(\boldsymbol{\theta})$ is continuous over $\boldsymbol{\Theta}$. Now, recall that $\tilde{\boldsymbol{\theta}}$ lies in the line segment connecting $\hat{\boldsymbol{\theta}}_h$ and $\boldsymbol{\theta}_h^*$. Furthermore, by Assumption (B-1) we have that $\hat{\boldsymbol{\theta}}_h \xrightarrow[N\to\infty]{p} \boldsymbol{\theta}_h^*$. Therefore, since $\bar{\mathbf{F}}(\boldsymbol{\theta})$ is continuous, it follows from Mann-Wald's Theorem that

$$\|\bar{\mathbf{F}}(\tilde{\boldsymbol{\theta}}) - \bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)\| \xrightarrow[N\to\infty]{p} 0. \tag{S-84}$$

Hence, the relation in (S-79) follows directly from (S-82)-(S-84). □

Lemma 7. *The relation in (S-83) holds under Assumptions (B-2)-(B-8).*

Proof. Using (S-68) and (S-73), one can verify that $\mathbf{F}(\boldsymbol{\theta})$ can be written as:

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_1(\boldsymbol{\theta})\mathbf{A}_2(\boldsymbol{\theta}) + \mathbf{A}_3(\boldsymbol{\theta})\mathbf{A}_4(\boldsymbol{\theta}) - \mathbf{A}_5\mathbf{A}_6(\boldsymbol{\theta}), \tag{S-85}$$

where

$$\begin{aligned}
\mathbf{A}_1(\boldsymbol{\theta}) &\triangleq \frac{1}{(N-1)N}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m)\mathbf{H}(\mathbf{x}_n;\boldsymbol{\theta}), \\
\mathbf{A}_2(\boldsymbol{\theta}) &\triangleq \frac{1}{N}\sum_{l=1}^{N} d(\mathbf{x}_l;\boldsymbol{\theta}), \\
\mathbf{A}_3(\boldsymbol{\theta}) &\triangleq \frac{1}{(N-1)N}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m)\mathbf{q}(\mathbf{x}_n;\boldsymbol{\theta}), \\
\mathbf{A}_4(\boldsymbol{\theta}) &\triangleq \frac{1}{N}\sum_{l=1}^{N} \mathbf{w}^T(\mathbf{x}_l;\boldsymbol{\theta}), \\
\mathbf{A}_5 &\triangleq \frac{1}{(N-1)N}\sum_{n=1}^{N}\sum_{m\neq n}^{N} K_h(\mathbf{x}_n-\mathbf{x}_m), \\
\mathbf{A}_6(\boldsymbol{\theta}) &\triangleq \frac{1}{N}\sum_{l=1}^{N} \boldsymbol{\Gamma}(\mathbf{x}_l;\boldsymbol{\theta}),
\end{aligned} \tag{S-86}$$

$\mathbf{q}(\mathbf{r};\boldsymbol{\theta})$ and $\mathbf{H}(\mathbf{r};\boldsymbol{\theta})$ are the gradient and Hessian of the log-likelihood defined in (B-5) and (B-6), respectively, $d(\mathbf{r};\boldsymbol{\theta})$ and $\mathbf{w}(\mathbf{r};\boldsymbol{\theta})$ are given in Eqs. (S-69) and (S-70), respectively, and $\boldsymbol{\Gamma}(\mathbf{r};\boldsymbol{\theta}) \triangleq \frac{\partial \mathbf{w}(\mathbf{r};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Additionally, by (S-68) and (S-80), it follows that

$$\bar{\mathbf{F}}(\boldsymbol{\theta}) = \bar{A}_1(\boldsymbol{\theta})\bar{A}_2(\boldsymbol{\theta}) + \bar{A}_3(\boldsymbol{\theta})\bar{A}_4(\boldsymbol{\theta}) - \bar{A}_5\bar{A}_6(\boldsymbol{\theta}), \tag{S-87}$$

where

$$\begin{aligned}
\bar{\mathbf{A}}_1(\boldsymbol{\theta}) &\triangleq \mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{H}(\mathbf{x};\boldsymbol{\theta}); G\times G], \\
\bar{\mathbf{A}}_2(\boldsymbol{\theta}) &\triangleq \mathbb{E}[d(\mathbf{x};\boldsymbol{\theta}); G], \\
\bar{\mathbf{A}}_3(\boldsymbol{\theta}) &\triangleq \mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\boldsymbol{\theta}); G\times G], \\
\bar{\mathbf{A}}_4(\boldsymbol{\theta}) &\triangleq \mathbb{E}^T[\mathbf{w}(\mathbf{x};\boldsymbol{\theta}); G], \\
\bar{\mathbf{A}}_5 &\triangleq \mathbb{E}[K_h(\mathbf{x}-\mathbf{x}'); G\times G], \\
\bar{\mathbf{A}}_6(\boldsymbol{\theta}) &\triangleq \mathbb{E}[\boldsymbol{\Gamma}(\mathbf{x};\boldsymbol{\theta}); G].
\end{aligned} \tag{S-88}$$

Therefore, using (S-85), (S-87), the identity

$$\mathbf{AB}-\mathbf{CD} = (\mathbf{A}-\mathbf{C})(\mathbf{B}-\mathbf{D}) + (\mathbf{A}-\mathbf{C})\mathbf{D} + \mathbf{C}(\mathbf{B}-\mathbf{D}),$$

that holds for matrices pairs $(\mathbf{A},\mathbf{C})$ and $(\mathbf{B},\mathbf{D})$ with identical within pair dimensions, and the triangle inequality, one can verify that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{F}(\boldsymbol{\theta}) - \bar{\mathbf{F}}(\boldsymbol{\theta})\| \leq B_1 + B_2 + B_3, \tag{S-89}$$

where

$$\begin{aligned}
B_1 \triangleq\ & \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_1(\boldsymbol{\theta}) - \bar{\mathbf{A}}_1(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_2(\boldsymbol{\theta}) - \bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \\
+\ & \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_1(\boldsymbol{\theta}) - \bar{\mathbf{A}}_1(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \\
+\ & \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\bar{\mathbf{A}}_1(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_2(\boldsymbol{\theta}) - \bar{\mathbf{A}}_2(\boldsymbol{\theta})\|,
\end{aligned} \tag{S-90}$$

$$\begin{aligned} B_2 \;\triangleq\; & \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_2(\boldsymbol{\theta}) - \bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_3(\boldsymbol{\theta}) - \bar{\mathbf{A}}_3(\boldsymbol{\theta})\| \\ + \;& \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_2(\boldsymbol{\theta}) - \bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\bar{\mathbf{A}}_3(\boldsymbol{\theta})\| \\ + \;& \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \times \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_3(\boldsymbol{\theta}) - \bar{\mathbf{A}}_3(\boldsymbol{\theta})\| \end{aligned} \tag{S-91}$$

and

$$\begin{aligned} B_3 \;\triangleq\; & \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_4(\boldsymbol{\theta}) - \bar{\mathbf{A}}_4(\boldsymbol{\theta})\| \times \|\mathbf{A}_5 - \bar{\mathbf{A}}_5\| \\ + \;& \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_4(\boldsymbol{\theta}) - \bar{\mathbf{A}}_4(\boldsymbol{\theta})\| \times \|\bar{\mathbf{A}}_5\| \\ + \;& \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\bar{\mathbf{A}}_4(\boldsymbol{\theta})\| \times \|\mathbf{A}_5 - \bar{\mathbf{A}}_5\|. \end{aligned} \tag{S-92}$$

By the definitions of $\mathbf{A}_2(\boldsymbol{\theta})$, $\mathbf{A}_4(\boldsymbol{\theta})$, $\mathbf{A}_6(\boldsymbol{\theta})$ in (S-86), the definitions of $\bar{\mathbf{A}}_2(\boldsymbol{\theta})$, $\bar{\mathbf{A}}_4(\boldsymbol{\theta})$, $\bar{\mathbf{A}}_6(\boldsymbol{\theta})$ in (S-88), Assumptions (B-3), (B-4) and the uniform weak law of large numbers [s9], it follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_2(\boldsymbol{\theta}) - \bar{\mathbf{A}}_2(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0, \;\; \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_4(\boldsymbol{\theta}) - \bar{\mathbf{A}}_4(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0, \;\; \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_6(\boldsymbol{\theta}) - \bar{\mathbf{A}}_6(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0 \tag{S-93}$$

and $\mathbf{A}_2(\boldsymbol{\theta})$, $\mathbf{A}_4(\boldsymbol{\theta})$ and $\mathbf{A}_6(\boldsymbol{\theta})$ are continuous over $\boldsymbol{\Theta}$. Furthermore, since the kernel function $K_h(\cdot)$ is bounded and symmetric, it follows from [s6, Th. 5.4.A] that

$$\|\mathbf{A}_5 - \bar{\mathbf{A}}_5\| \xrightarrow[N\to\infty]{p} 0. \tag{S-94}$$

Now, similarly to the proof of equality (S-47), in Lemma 4, it can be shown that when Assumptions (B-3)-(B-8) are satisfied it follows from [s7, Cor. 4.1] that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_1(\boldsymbol{\theta}) - \bar{\mathbf{A}}_1(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0, \;\; \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{A}_3(\boldsymbol{\theta}) - \bar{\mathbf{A}}_3(\boldsymbol{\theta})\| \xrightarrow[N\to\infty]{p} 0, \tag{S-95}$$

and $\mathbf{A}_1(\boldsymbol{\theta})$ and $\mathbf{A}_3(\boldsymbol{\theta})$ are continuous over $\boldsymbol{\Theta}$. Therefore, the relation in (S-83) follows directly from (S-89)-(S-95), the continuities of $\mathbf{A}_1(\boldsymbol{\theta})$, $\mathbf{A}_2(\boldsymbol{\theta})$, $\mathbf{A}_3(\boldsymbol{\theta})$, $\mathbf{A}_4(\boldsymbol{\theta})$ and $\mathbf{A}_6(\boldsymbol{\theta})$ and the compactness of $\boldsymbol{\Theta}$ that follows from Assumption (B-2). Additionally, by (S-85), the continuity of $\mathbf{F}(\boldsymbol{\theta})$ is a consequence of the continuities of $\mathbf{A}_1(\boldsymbol{\theta})$, $\mathbf{A}_2(\boldsymbol{\theta})$, $\mathbf{A}_3(\boldsymbol{\theta})$, $\mathbf{A}_4(\boldsymbol{\theta})$ and $\mathbf{A}_6(\boldsymbol{\theta})$. □

Lemma 8. *The relation in (S-81) holds.*

Proof. Using Eqs. (S-80), (S-87), (S-88) and the definition of $\psi_G(\mathbf{r}, h)$ in Eq. (2), one can verify that

$$\begin{aligned} \bar{\mathbf{F}}(\boldsymbol{\theta}) \;=\; & \Big(\mathbb{E}\left[\psi_G(\mathbf{x}, h)\mathbf{H}(\mathbf{x};\boldsymbol{\theta}); G\right] \times \mathbb{E}\left[\psi_G(\mathbf{x}, h); F_{\boldsymbol{\theta}}\right] \\ + \;& \mathbb{E}\left[\psi_G(\mathbf{x}, h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta}); G\right] \times \mathbb{E}\left[\psi_G(\mathbf{x}, h)\mathbf{q}^T(\mathbf{x};\boldsymbol{\theta}); F_{\boldsymbol{\theta}}\right] \\ - \;& \mathbb{E}\left[\psi_G(\mathbf{x}, h)\frac{\nabla^2_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})}; F_{\boldsymbol{\theta}}\right] \Big) \mathbb{E}^2[K_h(\mathbf{x} - \mathbf{x}'); G \times G]. \end{aligned} \tag{S-96}$$

By relation (S-108) we have that $\mathbb{E}[K_h(\mathbf{x} - \mathbf{x}')\mathbf{h}(\mathbf{x}, \mathbf{x}''; \boldsymbol{\theta}_h^*)); G \times G \times G] = \mathbf{0}$. Hence, using (S-68)-(S-70), and the definition of $\psi_G(\mathbf{r}, h)$ in Eq. (2), one can verify that

$$\mathbb{E}\left[\psi_G(\mathbf{x}, h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta}_h^*); F_{\boldsymbol{\theta}_h^*}\right] = \mathbb{E}\left[\psi_G(\mathbf{x}, h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta}_h^*); G\right] \times \mathbb{E}\left[\psi_G(\mathbf{x}, h); F_{\boldsymbol{\theta}_h^*}\right]. \tag{S-97}$$

Using again the definition of $\psi_G(\mathbf{r},h)$ in Eq. (2), it can be easily shown that for any scalar function $v(\cdot)$, such that the expectation $\mathbb{E}\left[\psi_G(\mathbf{x},h)v(\mathbf{x});F_{\boldsymbol{\theta}}\right]$ is finite, it holds that

$$\frac{\mathbb{E}\left[\psi_G(\mathbf{x},h)v(\mathbf{x});F_{\boldsymbol{\theta}}\right]}{\mathbb{E}\left[\psi_G(\mathbf{x},h);F_{\boldsymbol{\theta}}\right]}=\mathbb{E}\left[\psi_F(\mathbf{x},h)v(\mathbf{x});F_{\boldsymbol{\theta}}\right], \tag{S-98}$$

where $\psi_F(\mathbf{r},\boldsymbol{\theta},h)\triangleq(K_h*g)(\mathbf{r})/\mathbb{E}[(K_h*g)(\mathbf{x});F_{\boldsymbol{\theta}}]$. Therefore, by (S-97) and (S-98) we conclude that

$$\mathbb{E}\left[\psi_G(\mathbf{x},h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta}_h^*);G\right]=\mathbb{E}\left[\psi_F(\mathbf{x},h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta}_h^*);F_{\boldsymbol{\theta}_h^*}\right]. \tag{S-99}$$

Hence, relations (S-96)-(S-99) and the definition of $\mathbf{H}(\cdot,\cdot)$ in (B-6) imply that

$$\bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)=\left(\mathbb{E}\left[\psi_G(\mathbf{x},h)\nabla_{\boldsymbol{\theta}}^2\log f(\mathbf{x};\boldsymbol{\theta}_h^*);G\right]-\mathbf{B}(\boldsymbol{\theta}_h^*;h)\right)\eta(\boldsymbol{\theta}_h^*,h), \tag{S-100}$$

where

$$\begin{aligned}\mathbf{B}(\boldsymbol{\theta};h)&\triangleq\mathbb{E}\left[\psi_F(\mathbf{x},h)\frac{\nabla_{\boldsymbol{\theta}}^2 f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta})};F_{\boldsymbol{\theta}}\right]-\mathbb{E}\left[\psi_F(\mathbf{x},h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta});F_{\boldsymbol{\theta}}\right]\mathbb{E}\left[\psi_F(\mathbf{x},h)\mathbf{q}^T(\mathbf{x};\boldsymbol{\theta});F_{\boldsymbol{\theta}}\right]\\&=\nabla_{\boldsymbol{\theta}}^2\log E[(K_h*g)(\mathbf{x});F_{\boldsymbol{\theta}}]\end{aligned} \tag{S-101}$$

and the second equality in (S-101) follows from the definition of $\psi_F(\cdot,\cdot,\cdot)$ below (S-98) and Remark 1. Hence, relations (S-100) and (S-101) imply that $\bar{\mathbf{F}}(\boldsymbol{\theta}_h^*)=\eta(\boldsymbol{\theta}_h^*,h)\mathbf{C}(\boldsymbol{\theta}_h^*,h)$. □

Proposition 3. *The relation in (S-78) holds under conditions (B-2)-(B-5).*

Proof. Using (S-66), one can verify that

$$\sqrt{N}\mathbf{a}(\boldsymbol{\theta})=\frac{N-2}{N}\sqrt{N}\mathbf{a}_1(\boldsymbol{\theta})+\frac{1}{\sqrt{N}}\mathbf{a}_2(\boldsymbol{\theta})+\frac{1}{\sqrt{N}}\mathbf{a}_3(\boldsymbol{\theta}), \tag{S-102}$$

where

$$\mathbf{a}_1(\boldsymbol{\theta})\triangleq\frac{1}{(N-2)(N-1)N}\sum_{n=1}^N\sum_{m\neq n}^N\sum_{l\neq n,m}^N K_h(\mathbf{x}_n-\mathbf{x}_m)\mathbf{h}(\mathbf{x}_n,\mathbf{x}_l;\boldsymbol{\theta}), \tag{S-103}$$

$$\mathbf{a}_2(\boldsymbol{\theta})\triangleq\frac{1}{(N-1)N}\sum_{n=1}^N\sum_{m\neq n}^N K_h(\mathbf{x}_n-\mathbf{x}_m)\mathbf{h}(\mathbf{x}_n,\mathbf{x}_n;\boldsymbol{\theta}), \tag{S-104}$$

$$\mathbf{a}_3(\boldsymbol{\theta})\triangleq\frac{1}{(N-1)N}\sum_{n=1}^N\sum_{m\neq n}^N K_h(\mathbf{x}_n-\mathbf{x}_m)\mathbf{h}(\mathbf{x}_n,\mathbf{x}_m;\boldsymbol{\theta}). \tag{S-105}$$

First, we shall analyze convergence in probability of $\mathbf{a}_2(\boldsymbol{\theta})$ and $\mathbf{a}_3(\boldsymbol{\theta})$. Using Assumptions (B-4), (B-5), the boundedness of the kernel function and [s6, Th. 5.4.A], it can be shown that

$$\mathbf{a}_2(\boldsymbol{\theta})\xrightarrow[N\to\infty]{p}\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{h}(\mathbf{x},\mathbf{x};\boldsymbol{\theta});G\times G]<\infty \tag{S-106}$$

and

$$\mathbf{a}_3(\boldsymbol{\theta})\xrightarrow[N\to\infty]{p}\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{h}(\mathbf{x},\mathbf{x}';\boldsymbol{\theta});G\times G]<\infty. \tag{S-107}$$

Next, we shall analyze convergence in distribution of $\sqrt{N}\mathbf{a}_1(\boldsymbol{\theta}_h^*)$. We begin by calculating the expected value of $\mathbf{a}_1(\boldsymbol{\theta}_h^*)$. As shown at the beginning of Sec. II, $\boldsymbol{\theta}_h^*$ is the maximizer of the deterministic objective $\bar{\mathcal{J}}_h(\boldsymbol{\theta})$ defined in Eq. (S-28). Hence, by Assumption (B-2) and Eqs. (S-28), (S-68), (S-69) and (S-103) it follows that

$$\begin{aligned}\mathbb{E}[\mathbf{a}_1(\boldsymbol{\theta}_h^*);P_{\mathbf{a}_1(\boldsymbol{\theta}_h^*)}]&=\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{h}(\mathbf{x},\mathbf{x}'';\boldsymbol{\theta}_h^*));G\times G\times G]\\&=\nabla\bar{\mathcal{J}}_h(\boldsymbol{\theta}_h^*)\times\mathbb{E}^2\left[K_h(\mathbf{x}-\mathbf{x}');G\times G\right]\times\mathbb{E}\left[d(\mathbf{x}'';\boldsymbol{\theta}_h^*);G\right]=\mathbf{0}.\end{aligned} \tag{S-108}$$

Now, define the statistic

$$T \triangleq \boldsymbol{\beta}^T \mathbf{a}_1(\boldsymbol{\theta}_h^*), \tag{S-109}$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ is an arbitrary deterministic coefficient vector. In the following, we shall express T as a normalized U-statistic [s6, Ch. 5]. Using (S-103) and (S-109), one can verify that

$$T = \frac{U}{6}, \tag{S-110}$$

where

$$U \triangleq \frac{6}{(N-2)(N-1)N} \sum_c w(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \mathbf{x}_{i_3}), \tag{S-111}$$

$$w(\mathbf{r}, \boldsymbol{s}, \boldsymbol{t}) \triangleq \boldsymbol{\beta}^T \boldsymbol{\xi}(\mathbf{r}, \boldsymbol{s}, \boldsymbol{t}), \tag{S-112}$$

$$\boldsymbol{\xi}(\mathbf{r}, \boldsymbol{s}, \boldsymbol{t}) \triangleq \mathbf{z}(\mathbf{r}, \boldsymbol{s}, \boldsymbol{t}) + \mathbf{z}(\mathbf{r}, \boldsymbol{t}, \boldsymbol{s}) + \mathbf{z}(\boldsymbol{s}, \mathbf{r}, \boldsymbol{t}) + \mathbf{z}(\boldsymbol{s}, \boldsymbol{t}, \mathbf{r}) + \mathbf{z}(\boldsymbol{t}, \mathbf{r}, \boldsymbol{s}) + \mathbf{z}(\boldsymbol{t}, \boldsymbol{s}, \mathbf{r}), \tag{S-113}$$

$$\mathbf{z}(\mathbf{r}, \boldsymbol{s}, \boldsymbol{t}) \triangleq K_h(\mathbf{r} - \boldsymbol{s}) \mathbf{h}(\mathbf{r}, \boldsymbol{t}; \boldsymbol{\theta}_h^*) \tag{S-114}$$

and $\sum_c$ denotes the summation over the $\binom{N}{3}$ combinations of distinct elements $\{i_1, i_2, i_3\}$ from $\{1, \ldots, N\}$. Note that U is a U-statistic with symmetric kernel $w(\cdot, \cdot, \cdot)$. Also note that by (S-108)-(S-110), U is a zero-mean statistic. Therefore, assume that

$$\mathbb{E}[w^2(\mathbf{x}, \mathbf{x}', \mathbf{x}''); G \times G \times G] < \infty \tag{S-115}$$

and

$$\xi_1 > 0, \tag{S-116}$$

where

$$\xi_1 \triangleq \text{var}[w_1(\mathbf{x}); G], \tag{S-117}$$

$$w_1(\mathbf{r}) \triangleq \mathbb{E}[w(\mathbf{r}, \mathbf{x}', \mathbf{x}''); G \times G] = \boldsymbol{\beta}^T \boldsymbol{\zeta}(\mathbf{r}), \tag{S-118}$$

and

$$\boldsymbol{\zeta}(\mathbf{r}) \triangleq \mathbb{E}[\boldsymbol{\xi}(\mathbf{r}, \mathbf{x}', \mathbf{x}''); G \times G]. \tag{S-119}$$

By [s6, Th. 5.5.1-A], it follows that

$$\sqrt{N} U \xrightarrow[N \to \infty]{d} \mathcal{N}(0, 9\xi_1). \tag{S-120}$$

In Lemma 9, stated below, we show that

$$\xi_1 = 4\eta^2(\boldsymbol{\theta}_h^*, h) \boldsymbol{\beta}^T \mathbf{D}(\boldsymbol{\theta}_h^*, h) \boldsymbol{\beta}, \tag{S-121}$$

where $\mathbf{D}(\boldsymbol{\theta}, h)$ and $\eta(\boldsymbol{\theta}, h)$ are defined below Eq. (S-63) and in Eq. (S-75), respectively. Hence, by (S-108)-(S-110), (S-120) and the Cramér-Wold Device [s12, Th. 11.2.3] we conclude that

$$\sqrt{N} \mathbf{a}_1(\boldsymbol{\theta}_h^*) \xrightarrow[N \to \infty]{d} \mathcal{N}(\mathbf{0}, \eta^2(\boldsymbol{\theta}_h^*, h) \mathbf{D}(\boldsymbol{\theta}_h^*, h)). \tag{S-122}$$

Thus, the relation in (S-78) follows directly from (S-102), (S-106), (S-107), (S-122) and Slutskey's Theorem [s2, Th. 9.1.6]. To complete the proof, we need to show that the assumptions in (S-115) and (S-116) are satisfied. By

Eq. (S-112) the inequality in (S-115) holds when $\mathbb{E}[\|\boldsymbol{\xi}(\mathbf{x},\mathbf{x}',\mathbf{x}'')\|^2; G\times G\times G]<\infty$. Using Eqs. (S-113), (S-114), the Cauchy-Schwartz and the triangle inequalities, the boundedness of the kernel function $K_h(\cdot)$ and Assumption (B-5), one can verify that the latter inequality indeed holds. The inequality in (S-116) is satisfied when $\mathbf{D}(\boldsymbol{\theta}_h^*,h)$ is positive definite. Again, we note that $0<\eta(\boldsymbol{\theta}_h^*,h)<\infty$ since the kernel function $K_h(\cdot)$ is strictly positive and bounded. □

Lemma 9. *The equality in (S-121) holds.*

Proof. Using (S-108)-(S-112), (S-118) and (S-119), one can verify that $\mathbb{E}[w_1(\mathbf{x});G]=0$. Hence, by (S-117)-(S-119), it follows that

$$\xi_1=\boldsymbol{\beta}^T\mathbb{E}\left[\boldsymbol{\zeta}(\mathbf{x})\boldsymbol{\zeta}^T(\mathbf{x});G\right]\boldsymbol{\beta}. \tag{S-123}$$

Furthermore, using (S-68), (S-113), (S-114), (S-119) and the symmetry property of the kernel function $K_h(\cdot)$, one can verify that

$$\begin{aligned}\boldsymbol{\zeta}(\mathbf{r}) &= 2\Big(\mathbb{E}\left[K_h(\mathbf{x}'-\mathbf{r});G\right]\times\mathbb{E}\left[\mathbf{h}(\mathbf{r},\mathbf{x}';\boldsymbol{\theta}_h^*);G\right] \\ &+ \mathbb{E}\left[K_h(\mathbf{x}'-\mathbf{r})\mathbf{h}(\mathbf{x}',\mathbf{x}'';\boldsymbol{\theta}_h^*);G\times G\right] \\ &+ \mathbb{E}\left[K_h(\mathbf{x}'-\mathbf{x}'')\mathbf{h}(\mathbf{x}',\mathbf{r};\boldsymbol{\theta}_h^*);G\times G\right]\Big)=2\eta(\boldsymbol{\theta}_h^*,h)\mathbf{v}(\mathbf{r},\boldsymbol{\theta}_h^*,h),\end{aligned} \tag{S-124}$$

where

$$\mathbf{v}(\mathbf{r},\boldsymbol{\theta},h)\triangleq\psi_G(\mathbf{r},h)\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h)+\mathbf{d}(\mathbf{r},\boldsymbol{\theta},h)-\mathbf{z}(\mathbf{r},\boldsymbol{\theta},h). \tag{S-125}$$

The vector function

$$\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h)\triangleq\mathbf{q}(\mathbf{r};\boldsymbol{\theta})-\mathbb{E}[\psi_F(\mathbf{x},\boldsymbol{\theta},h)\mathbf{q}(\mathbf{x};\boldsymbol{\theta});F_{\boldsymbol{\theta}}]=\nabla_{\boldsymbol{\theta}}\log f(\mathbf{r};\boldsymbol{\theta})-\nabla_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h), \tag{S-126}$$

where the second equality in (S-126) follows from the definitions of $\mathbf{q}(\cdot;\cdot)$, $\psi_F(\cdot,\cdot,\cdot)$ $u(\cdot,\cdot)$ stated in (B-5) and below (S-98) and (S-63), respectively, and from Remark 1. The vector function

$$\mathbf{d}(\mathbf{r},\boldsymbol{\theta},h)\triangleq\mathbb{E}[\mathbf{c}(\mathbf{x},\boldsymbol{\theta},h)\varphi_h(\mathbf{r}-\mathbf{x});G], \tag{S-127}$$

where $\varphi_h(\cdot)$ is defined below (S-63). Lastly, the vector function

$$\begin{aligned}\mathbf{z}(\mathbf{r},\boldsymbol{\theta},h) &\triangleq \frac{\mathbb{E}[\mathbf{c}(\mathbf{x},\boldsymbol{\theta},h)K_h(\mathbf{r}-\mathbf{x});F_{\boldsymbol{\theta}}]}{\mathbb{E}[(K_h*g(\mathbf{x}));F_{\boldsymbol{\theta}}]} \\ &\overset{(a)}{=} \frac{\nabla_{\boldsymbol{\theta}}\mathbb{E}[K_h(\mathbf{r}-\mathbf{x});F_{\boldsymbol{\theta}}]}{\mathbb{E}[(K_h*g(\mathbf{x}));F_{\boldsymbol{\theta}}]}-\frac{\mathbb{E}[K_h(\mathbf{r}-\mathbf{x});F_{\boldsymbol{\theta}}]}{\mathbb{E}[(K_h*g(\mathbf{x}));F_{\boldsymbol{\theta}}]}\nabla_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h) \\ &\overset{(b)}{=} \frac{\nabla_{\boldsymbol{\theta}}(K_h*f)(\mathbf{r};\boldsymbol{\theta})}{\mathbb{E}[(K_h*g(\mathbf{x}));F_{\boldsymbol{\theta}}]}-\frac{(K_h*f)(\mathbf{r};\boldsymbol{\theta})}{\mathbb{E}[(K_h*g(\mathbf{x}));F_{\boldsymbol{\theta}}]}\nabla_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h) \\ &\overset{(c)}{=} \frac{\nabla_{\boldsymbol{\theta}}v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)}-\frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)}\nabla_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h)=\frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)}\nabla_{\boldsymbol{\theta}}\log\frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)},\end{aligned} \tag{S-128}$$

where (a) follows from (S-126) and Remark 1, (b) is a direct consequence of the definition of the convolution operator, and (c) follows from the definitions of $v(\cdot,\cdot,\cdot)$ and $v(\cdot,\cdot)$ below (S-63). Hence, the relation in (S-121) follows directly from the definition of $\mathbf{D}(\cdot,\cdot)$ below (S-63), (S-123) and (S-124). □

IV. Fisher consistency

In this section, we show that, under some mild regularity assumptions, the M$\mathcal{K}$CDE is Fisher-consistent [s13].

Proposition 4. *Under Assumption (A-2), $\hat{\boldsymbol{\theta}}_h$ (6) is a Fisher consistent estimator [s13] of $\boldsymbol{\theta}_h^*$ (7), i.e., it can be represented as a statistical functional of the empirical probability distribution $\mathbf{S}[\hat{G}]$ that satisfies $\mathbf{S}[G] = \boldsymbol{\theta}_h^*$.*

Proof. First, we show that $\hat{\boldsymbol{\theta}}_h$ can be represented as a statistical functional of the empirical probability distribution $\hat{G} \triangleq \frac{1}{N}\sum_{n=1}^{N} \delta_{\mathbf{x}_n}$, where $\delta_{\mathbf{r}}$ denotes a Dirac measure [s3] concentrated at $\mathbf{r}$. Using Eq. (3), one can verify that the objective function

$$\mathcal{J}_h(\boldsymbol{\theta}) = \frac{\mathbb{E}[\tilde{K}_h(\mathbf{x}-\mathbf{x}')\log f(\mathbf{x};\boldsymbol{\theta}); \hat{G}\times\hat{G}]}{\mathbb{E}[\tilde{K}_h(\mathbf{x}-\mathbf{x}'); \hat{G}\times\hat{G}]} - \log \mathbb{E}[\tilde{K}_h(\mathbf{x}-\mathbf{x}'); \hat{G}\times F_{\boldsymbol{\theta}}] \triangleq H[\hat{G};\boldsymbol{\theta}]. \tag{S-129}$$

where $\tilde{K}_h(\mathbf{r}) \triangleq K_h(\mathbf{r}) - K_h(0)\mathbb{1}_0(\mathbf{r})$ and $\mathbb{1}_0(\cdot)$ denotes the indicator function of 0. Therefore, by (6) it follows that

$$\hat{\boldsymbol{\theta}}_h = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} H[\hat{G};\boldsymbol{\theta}] \triangleq \mathbf{S}[\hat{G}]. \tag{S-130}$$

Next, we prove Fisher consistency of $\hat{\boldsymbol{\theta}}_h$. By (S-28) and (S-129) it follows that

$$H[G;\boldsymbol{\theta}] = \bar{\mathcal{J}}_h(\boldsymbol{\theta}), \tag{S-131}$$

which as shown below (S-28) is uniquely maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_h^*$ when Assumption (A-2) is satisfied. Therefore, by (S-130) we conclude that

$$\boldsymbol{\theta}_h^* = \mathbf{S}[G], \tag{S-132}$$

implying that $\hat{\boldsymbol{\theta}}_h$ is a Fisher consistent estimator of $\boldsymbol{\theta}_h^*$. □

V. Influence function

In this section, we derive the influence function [s14] of the proposed M$\mathcal{K}$CDE. To that sake, we define the contaminated probability distribution

$$G_\epsilon \triangleq (1-\epsilon)F_{\boldsymbol{\theta}_0} + \epsilon\delta_{\mathbf{r}}, \tag{S-133}$$

where $0 \le \epsilon \le 1$, $\mathbf{r} \in \mathbb{R}^p$ and $\delta_{\mathbf{r}}$ is the Dirac probability measure at $\mathbf{r}$. The influence function of a Fisher consistent estimator with statistical functional $\mathbf{S}[\cdot]$ at $F_{\boldsymbol{\theta}_0}$ is defined as [s14]:

$$\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0) \triangleq \lim_{\epsilon\to 0} \frac{\mathbf{S}[G_\epsilon] - \mathbf{S}[F_{\boldsymbol{\theta}_0}]}{\epsilon} = \left.\frac{d\mathbf{S}[G_\epsilon]}{d\epsilon}\right|_{\epsilon=0}. \tag{S-134}$$

Proposition 5. *Assume that $\hat{\boldsymbol{\theta}}_h$ is Fisher consistent. Furthermore, assume that conditions (B-3) and (B-4), stated in Theorem 3, are satisfied. Then, the influence function of $\hat{\boldsymbol{\theta}}_h$ takes the form:*

$$\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h) = \bar{\mathbf{D}}^{-1}(\boldsymbol{\theta}_0,h)\bar{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta}_0,h)\bar{\psi}_F(\mathbf{r},\boldsymbol{\theta}_0,h), \tag{S-135}$$

where $\bar{\mathbf{D}}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\bar{\psi}_F(\mathbf{x},\boldsymbol{\theta},h)\bar{\mathbf{c}}(\mathbf{x},\boldsymbol{\theta},h)\bar{\mathbf{c}}^T(\mathbf{x},\boldsymbol{\theta},h); F_{\boldsymbol{\theta}}]$, $\bar{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \mathbf{q}(\mathbf{r},\boldsymbol{\theta}) - \mathbb{E}[\bar{\psi}_F(\mathbf{x},\boldsymbol{\theta},h)\mathbf{q}(\mathbf{x},\boldsymbol{\theta}); F_{\boldsymbol{\theta}}]$, $\mathbf{q}(\mathbf{r},\boldsymbol{\theta})$ is the score-function defined in (B-5), $\bar{\psi}_F(\mathbf{r},\boldsymbol{\theta},h) \triangleq (K_h * f)(\mathbf{r};\boldsymbol{\theta})/\mathbb{E}[(K_h * f)(\mathbf{x};\boldsymbol{\theta}); F_{\boldsymbol{\theta}}]$ and it is assumed that $\bar{\mathbf{D}}(\boldsymbol{\theta},h)$ is non-singular.

Proof. Throughout the proof we shall assume that integration and differentiation operations can be interchanged. Using [s11, Th. 2.40] it can be shown that this assumption is justified under conditions (B-3), (B-4) and the boundedness of the kernel function $K_h(\cdot)$.

By Eq. (S-97), the definition of $\psi_G(\cdot,\cdot)$ in Eq. (2), the definition of $\psi_F(\cdot,\cdot,\cdot)$ stated below Eq. (S-4) and the Fisher consistency of $\hat{\boldsymbol{\theta}}_h$ it follows that

$$\frac{\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);G_\epsilon\times G_\epsilon]}{\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}');G_\epsilon\times G_\epsilon]}-\frac{\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});G_\epsilon]\tilde{\mathbf{q}}(\mathbf{x};\mathbf{S}[G_\epsilon]))d\lambda(\boldsymbol{\tau})}{\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});G_\epsilon]f(\mathbf{x};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})}=\mathbf{0}, \tag{S-136}$$

where $\tilde{\mathbf{q}}(\mathbf{x};\boldsymbol{\theta})\triangleq\nabla_{\boldsymbol{\theta}}f(\mathbf{x};\boldsymbol{\theta})$. Under the assumption of a symmetric kernel function, it follows from (S-133) that the terms comprising (S-136) take the forms:

$$\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);G_\epsilon\times G_\epsilon]=(1-\epsilon)^2\mathbf{a}(\epsilon)+\epsilon(1-\epsilon)\left(\mathbf{b}(\mathbf{r},\epsilon)+\mathbf{q}(\mathbf{r};\mathbf{S}[G_\epsilon])c(\mathbf{r})\right)+\epsilon^2\mathbf{q}(\mathbf{r};\mathbf{S}[G_\epsilon])K_h(\mathbf{0}), \tag{S-137}$$

$$\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}');G_\epsilon\times G_\epsilon]=(1-\epsilon)^2d+2\epsilon(1-\epsilon)c(\mathbf{r})+\epsilon^2K_h(\mathbf{0}), \tag{S-138}$$

$$\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});G_\epsilon]\tilde{\mathbf{q}}(\mathbf{x};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})=(1-\epsilon)\mathbf{w}(\epsilon)+\epsilon\mathbf{z}(\mathbf{r},\epsilon) \tag{S-139}$$

and

$$\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});G_\epsilon]f(\boldsymbol{\tau};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})=(1-\epsilon)u(\epsilon)+\epsilon v(\mathbf{r},\epsilon), \tag{S-140}$$

where

$$\mathbf{a}(\epsilon)\triangleq\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}],\quad \mathbf{b}(\mathbf{r},\epsilon)\triangleq\mathbb{E}[K_h(\mathbf{x}-\mathbf{r})\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);F_{\boldsymbol{\theta}_0}], \tag{S-141}$$

$$c(\mathbf{r})\triangleq\mathbb{E}[K_h(\mathbf{x}-\mathbf{r});F_{\boldsymbol{\theta}_0}],\quad d\triangleq\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}');F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}], \tag{S-142}$$

$$\mathbf{w}(\epsilon)\triangleq\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});F_{\boldsymbol{\theta}_0}]\tilde{\mathbf{q}}(\mathbf{x};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})],\quad \mathbf{z}(\mathbf{r},\epsilon)\triangleq\int_{\mathbb{R}^p}K_h(\mathbf{r}-\boldsymbol{\tau})\tilde{\mathbf{q}}(\mathbf{x};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})], \tag{S-143}$$

$$u(\epsilon)\triangleq\int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});F_{\boldsymbol{\theta}_0}]f(\boldsymbol{\tau};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau})\quad\text{and}\quad v(\mathbf{r},\epsilon)\triangleq\int_{\mathbb{R}^p}K_h(\mathbf{r}-\boldsymbol{\tau})f(\boldsymbol{\tau};\mathbf{S}[G_\epsilon])d\lambda(\boldsymbol{\tau}). \tag{S-144}$$

Hence, by (S-136)-(S-140) we obtain that

$$\boldsymbol{\alpha}(\mathbf{r},\epsilon)-\boldsymbol{\beta}(\mathbf{r},\epsilon)=\mathbf{0}, \tag{S-145}$$

where

$$\boldsymbol{\alpha}(\mathbf{r},\epsilon)\triangleq\frac{(1-\epsilon^2)\mathbf{a}(\epsilon)+\epsilon(1-\epsilon)(\mathbf{b}(\mathbf{r},\epsilon)+c(\mathbf{r})\mathbf{q}(\mathbf{r};\mathbf{S}[G_\epsilon]))+\epsilon^2\mathbf{q}(\mathbf{r};\mathbf{S}[G_\epsilon])K_h(\mathbf{0})}{(1-\epsilon)^2d+2\epsilon(1-\epsilon)c(\mathbf{r})+\epsilon^2K_h(0)} \tag{S-146}$$

and

$$\boldsymbol{\beta}(\mathbf{r},\epsilon)\triangleq\frac{(1-\epsilon)\mathbf{w}(\epsilon)+\epsilon\mathbf{z}(\mathbf{r},\epsilon)}{(1-\epsilon)u(\epsilon)+\epsilon v(\mathbf{r},\epsilon)}. \tag{S-147}$$

From (S-145) we obtain that

$$\left.\frac{\partial\boldsymbol{\alpha}(\mathbf{r},\epsilon)}{\partial\epsilon}\right|_{\epsilon=0}-\left.\frac{\boldsymbol{\beta}(\mathbf{r},\epsilon)}{\partial\epsilon}\right|_{\epsilon=0}=\mathbf{0}. \tag{S-148}$$

Using (S-146) and (S-147) one can verify that

$$\left.\frac{\partial\boldsymbol{\alpha}(\mathbf{r},\epsilon)}{\partial\epsilon}\right|_{\epsilon=0}=\frac{\dot{\mathbf{a}}(0)+\mathbf{b}(\mathbf{r},0)+\mathbf{q}(\mathbf{r};\boldsymbol{\theta}_0)c(\mathbf{r})}{d}-\frac{2\mathbf{a}(0)c(\mathbf{r})}{d^2} \tag{S-149}$$

and

$$\left.\frac{\partial \boldsymbol{\beta}(\mathbf{r},\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = \frac{\dot{\mathbf{w}}(0)+\mathbf{z}(\mathbf{r},0)}{u(0)} - \frac{\mathbf{w}(0)(\dot{u}(0)+v(0))}{u^2(0)}, \tag{S-150}$$

where $\dot{\mathbf{a}}(0) \triangleq \left.\frac{d\mathbf{a}(\epsilon)}{d\epsilon}\right|_{\epsilon=0}$, $\dot{\mathbf{w}}(0) \triangleq \left.\frac{d\mathbf{w}(\epsilon)}{d\epsilon}\right|_{\epsilon=0}$ and $\dot{u}(0) \triangleq \left.\frac{du(\epsilon)}{d\epsilon}\right|_{\epsilon=0}$. By the chain-rule for derivatives we obtain that

$$\dot{\mathbf{a}}(0) = \left.\frac{d\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}]}{d\mathbf{S}[G_\epsilon]}\right|_{\epsilon=0} \left.\frac{d\mathbf{S}[G_\epsilon]}{d\epsilon}\right|_{\epsilon=0}. \tag{S-151}$$

Note that

$$\frac{d\mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{q}(\mathbf{x};\mathbf{S}[G_\epsilon]);F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}]}{d\mathbf{S}[G_\epsilon]} = \mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{H}(\mathbf{x};\mathbf{S}[G_\epsilon]);F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}], \tag{S-152}$$

where $\mathbf{H}(\cdot,\cdot)$ is the Hessian defined in (B-6). Also note that the Fisher consistency of $\hat{\boldsymbol{\theta}}_h$ implies that

$$\mathbf{S}[G_\epsilon] = \boldsymbol{\theta}_0 \quad \text{for} \quad \epsilon = 0 \tag{S-153}$$

and therefore,

$$\mathbf{H}(\mathbf{x};\mathbf{S}[G_\epsilon])|_{\epsilon=0} = \mathbf{H}(\mathbf{x};\boldsymbol{\theta}_0). \tag{S-154}$$

Additionally, by (S-134) it follows that $\left.\frac{d\mathbf{S}[G_\epsilon]}{d\epsilon}\right|_{\epsilon=0} = \mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h)$. Hence, we conclude that

$$\dot{\mathbf{a}}(0) = \mathbf{A}\times\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h). \tag{S-155}$$

where $\mathbf{A} \triangleq \mathbb{E}[K_h(\mathbf{x}-\mathbf{x}')\mathbf{H}(\mathbf{x};\boldsymbol{\theta}_0);F_{\boldsymbol{\theta}_0}\times F_{\boldsymbol{\theta}_0}]$. Similarly, it can be shown that

$$\dot{\mathbf{w}}(0) = \mathbf{B}\times\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h) \tag{S-156}$$

and

$$\dot{u}(0) = \boldsymbol{\eta}^T\times\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h), \tag{S-157}$$

where the matrix $\mathbf{B} \triangleq \int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});F_{\boldsymbol{\theta}_0}]\tilde{\mathbf{H}}(\mathbf{r};\boldsymbol{\theta}_0)d\lambda(\boldsymbol{\tau})$, with $\tilde{\mathbf{H}}(\mathbf{x};\boldsymbol{\theta}) \triangleq \nabla^2_{\boldsymbol{\theta}}f(\mathbf{x};\boldsymbol{\theta})$ and the vector $\boldsymbol{\eta} \triangleq \int_{\mathbb{R}^p}\mathbb{E}[K_h(\mathbf{x}-\boldsymbol{\tau});F_{\boldsymbol{\theta}_0}]\tilde{\mathbf{q}}(\mathbf{r};\boldsymbol{\theta}_0)d\lambda(\boldsymbol{\tau})$. Therefore, by (S-148)-(S-150) and (S-155)-(S-157) we obtain that

$$\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h) = \mathbf{J}^{-1}(\boldsymbol{\theta}_0)\mathbf{K}(\mathbf{r},\boldsymbol{\theta}_0), \tag{S-158}$$

where

$$\mathbf{J}(\boldsymbol{\theta}) \triangleq \frac{\mathbf{A}}{d} - \frac{\mathbf{B}}{u(0)} + \frac{\mathbf{w}(0)\boldsymbol{\eta}^T}{u^2(0)} = \frac{\mathbf{A}-\mathbf{B}}{d} + \frac{\mathbf{w}(0)\boldsymbol{\eta}^T}{d^2} \tag{S-159}$$

and

$$\begin{aligned}\mathbf{K}(\mathbf{r},\boldsymbol{\theta}) &\triangleq -\frac{\mathbf{b}(\mathbf{r},0)+c(r)\mathbf{q}(\mathbf{r};\boldsymbol{\theta})}{d} + \frac{2c(\mathbf{r})\mathbf{a}(0)}{d^2} + \frac{\mathbf{z}(\mathbf{r},0)}{u(0)} - \frac{v(\mathbf{r},0)\mathbf{w}(0)}{u^2(0)} \\ &= -\frac{\mathbf{b}(\mathbf{r},0)+c(r)\mathbf{q}(\mathbf{r};\boldsymbol{\theta})-\mathbf{z}(\mathbf{r},0)}{d} + \frac{2c(\mathbf{r})\mathbf{a}(0)-v(\mathbf{r},0)\mathbf{w}(0)}{d^2},\end{aligned} \tag{S-160}$$

where the last equalities in (S-159) and (S-160) follow from the definitions of d and $u(\cdot)$ in (S-142) and (S-144), respectively, and relation (S-153) according to which $d = u(0)$. Using the definitions of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{w}(\cdot)$, $\boldsymbol{\eta}$, d and $u(\cdot)$ stated above, relation (S-153) and the identity

$$\tilde{\mathbf{H}}(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta})\left(\mathbf{H}(\mathbf{x};\boldsymbol{\theta})+\mathbf{q}(\mathbf{x};\boldsymbol{\theta})\mathbf{q}^T(\mathbf{x};\boldsymbol{\theta})\right),$$

that follows directly from the definition of $\tilde{\mathbf{H}}(\cdot;\cdot)$ below (S-157), one can verify that

$$\mathbf{J}(\boldsymbol{\theta}) = -\bar{\mathbf{D}}(\boldsymbol{\theta}, h), \tag{S-161}$$

where $\bar{\mathbf{D}}(\boldsymbol{\theta}, h)$ is defined below Eq. (S-135). Now, by (S-136), (S-153) and the definitions of $\mathbf{a}(\cdot)$, $\mathbf{w}(\cdot)$, d and $u(\cdot)$ it follows that

$$\mathbf{a}(0) = \mathbf{w}(0). \tag{S-162}$$

Therefore, by (S-160), (S-162), the definitions of $\mathbf{b}(\cdot,\cdot)$, $c(\cdot)$, $\mathbf{z}(\cdot,\cdot)$, $\mathbf{a}(\cdot)$, $v(\cdot,\cdot)$ and $\mathbf{w}(\cdot)$ stated above, it can be shown that

$$\mathbf{K}(\mathbf{r}, \boldsymbol{\theta}) = -\bar{\mathbf{c}}(\mathbf{r}, \boldsymbol{\theta}_0, h)\bar{\psi}_F(\mathbf{r}, \boldsymbol{\theta}_0, h), \tag{S-163}$$

where $\bar{\mathbf{c}}(\mathbf{r}, \boldsymbol{\theta}_0, h)$ and $\bar{\psi}_F(\mathbf{r}, \boldsymbol{\theta}_0, h)$ are defined below Eq. (S-135). Hence the equality in (S-135) follows directly from (S-158), (S-161) and (S-163). □

VI. Implementation details

In this section we derive the fixed-point algorithms used for implementation of the M$\mathcal{K}$DE and the other compared estimators in Section 5. The proposed M$\mathcal{K}$DE was initialized at $\hat{\boldsymbol{\eta}}_{\text{init}} = 2 \times \mathbf{1}$ and $\hat{\sigma}^2_{\text{init}} = 1$. For the other compared algorithms, initialization closer to $\boldsymbol{\theta}_0$ was required to obtain fast and reliable convergence. Hence, these algorithms were initialized at $\hat{\boldsymbol{\eta}}_{\text{init}} = 4 \times \mathbf{1}$ and $\hat{\sigma}^2_{\text{init}} = 2$. In all compared estimators, the maximum number of fixed-point iterations and stopping criterion were set to 100 and $\|\hat{\boldsymbol{\theta}}_l - \hat{\boldsymbol{\theta}}_{l-1}\|/\|\hat{\boldsymbol{\theta}}_{l-1}\| < 10^{-4}$, respectively, where $\hat{\boldsymbol{\theta}}_l$ denotes a parameter estimate at the l-th iteration.

A. M$\mathcal{K}$DE

Under the nominal density function in (20) and the Gaussian kernel function (22), the objective in (3) can be written as:

$$\mathcal{J}_h(\boldsymbol{\theta}) = \sum\nolimits_{n=1}^{N} w(\mathbf{x}_n; h) \log f(\mathbf{x}_n; \boldsymbol{\theta}) - \log \sum\nolimits_{n=1}^{N} \bar{f}_h(\mathbf{x}_n; \boldsymbol{\theta}), \tag{S-164}$$

where $w(\mathbf{r}; h)$ is defined below (3) and

$$\bar{f}_h(\mathbf{r}; \boldsymbol{\theta}) \triangleq \frac{1}{2}\phi(\mathbf{r}; \mathbf{0}, (\sigma^2 + h^2)\mathbf{I}) + \frac{1}{2}\phi(\mathbf{r}; \boldsymbol{\eta}, (\sigma^2 + h^2)\mathbf{I}).$$

Hence, by equating the partial gradients of $\mathcal{J}_h(\boldsymbol{\theta})$ (S-164) to zero, one can verify that the M$\mathcal{K}$DE $\hat{\boldsymbol{\theta}}_h \triangleq [\hat{\boldsymbol{\eta}}_h^T, \hat{\sigma}_h^2]^T$ is the solution of the following simple fixed-point iterations:

$$\hat{\boldsymbol{\eta}}_h = \frac{\sum_{n=1}^{N} \tau_h(\mathbf{x}_n, \hat{\boldsymbol{\theta}}_h)\mathbf{x}_n}{\sum_{n=1}^{N} \tau_h(\mathbf{x}_n, \hat{\boldsymbol{\theta}}_h)}$$

and

$$\hat{\sigma}_h^2 = \frac{(\hat{\sigma}_h^2 + h^2)^2 \sum_{n=1}^{N} s_h(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_h)}{ph^2(\hat{\sigma}_h^2 + h^2) + \hat{\sigma}_h^2 \sum_{n=1}^{N} t_h(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_h)},$$

where

$$\tau_h(\mathbf{r}; \boldsymbol{\theta}) \triangleq \frac{1}{\sigma^2}\alpha_h(\mathbf{r}; \boldsymbol{\eta}, \sigma^2) - \frac{1}{\sigma^2 + h^2}\beta_h(\mathbf{r}; \boldsymbol{\eta}, \sigma^2),$$
$$s_h(\mathbf{r}; \boldsymbol{\theta}) = \alpha_h(\mathbf{r}; \boldsymbol{\eta}, \sigma^2)\|\mathbf{r} - \boldsymbol{\eta}\|^2 + \alpha_h(\mathbf{r}; \mathbf{0}, \sigma^2)\|\mathbf{r}\|^2, \tag{S-165}$$

$$t_h(\mathbf{r};\boldsymbol{\theta}) = \beta_h(\mathbf{r};\boldsymbol{\eta},\sigma^2)\|\mathbf{r}-\boldsymbol{\eta}\|^2 + \beta_h(\mathbf{r};\mathbf{0},\sigma^2)\|\mathbf{r}\|^2,$$

$$\alpha_h(\mathbf{r};\boldsymbol{\eta},\sigma^2) \triangleq \frac{w(\mathbf{r};h)\phi(\mathbf{r};\boldsymbol{\eta},\sigma^2\mathbf{I})}{2f(\mathbf{r};\boldsymbol{\theta})}$$

and

$$\beta_h(\mathbf{r};\boldsymbol{\eta},\sigma^2) \triangleq \frac{\phi(\mathbf{r};\boldsymbol{\eta},(\sigma^2+h^2)\mathbf{I})}{2\sum_{n=1}^N \bar{f}_h(\mathbf{x}_n;\boldsymbol{\theta})}.$$

B. MLE implementation

The MLE is obtained via minimization of the log-likelihood function under the nominal density function in (20). Hence, by equating the partial gradients of the log-likelihood function to zero, one can verify that the MLE $\hat{\boldsymbol{\theta}}_{\text{ML}} \triangleq [\hat{\boldsymbol{\eta}}_{\text{ML}}^T, \hat{\sigma}^2_{\text{ML}}]^T$ is the solution of the following fixed-point iterations:

$$\hat{\boldsymbol{\eta}}_{\text{ML}} = \frac{\sum_{n=1}^N \tau(\mathbf{x}_n,\hat{\boldsymbol{\theta}}_{\text{ML}})\mathbf{x}_n}{\sum_{n=1}^N \tau(\mathbf{x}_n,\hat{\boldsymbol{\theta}}_{\text{ML}})}$$

and

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{2pN}\sum_{n=1}^N \frac{s(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_{\text{ML}})}{f(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_{\text{ML}})},$$

where

$$\tau(\mathbf{r};\boldsymbol{\theta}) \triangleq \frac{\phi(\mathbf{r};\boldsymbol{\eta},\sigma^2)}{2f(\mathbf{r};\boldsymbol{\theta})}$$

and

$$s(\mathbf{r};\boldsymbol{\theta}) \triangleq \phi(\mathbf{r};\boldsymbol{\eta},\sigma^2)\|\mathbf{r}-\boldsymbol{\eta}\|^2 + \phi(\mathbf{r};\mathbf{0},\sigma^2)\|\mathbf{r}\|^2. \tag{S-166}$$

C. MαDE

As stated in Section 5, the MαDE was implemented similarly to [s15], by minimizing the empirical α-divergence between the power-transformed density $g_\alpha(\mathbf{x}) \triangleq g^{\frac{1}{\alpha}}(\mathbf{x})/\int_{\mathbb{R}^p} g^{\frac{1}{\alpha}}(\mathbf{x})d\lambda(\mathbf{x})$ and $f(\mathbf{x};\boldsymbol{\theta})$ (20). Hence, by computing the partial gradients of the empirical α-divergence to zero, one can verify that the MαDE $\hat{\boldsymbol{\theta}}_\alpha \triangleq [\hat{\boldsymbol{\eta}}_\alpha^T, \hat{\sigma}^2_\alpha]^T$ is the solution of the following simple fixed-point iterations:

$$\hat{\boldsymbol{\eta}}_\alpha = \frac{\sum_{n=1}^N f^{-\alpha}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_\alpha)\phi(\mathbf{x}_n;\hat{\boldsymbol{\eta}}_\alpha,\hat{\sigma}^2_\alpha)\mathbf{x}_n}{\sum_{n=1}^N f^{-\alpha}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_\alpha)\phi(\mathbf{x}_n;\hat{\boldsymbol{\eta}}_\alpha,\hat{\sigma}^2_\alpha)}$$

and

$$\hat{\sigma}^2_\alpha = \frac{\sum_{n=1}^N f^{-\alpha}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_\alpha)s(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_\alpha)}{2p\sum_{n=1}^N f^{1-\alpha}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_\alpha)},$$

where $s(\cdot;\cdot)$ is defined in (S-166).

D. Mβ DE ($\beta = 1$)

Under the nominal density function in (20), it can be shown that the MβDE [s16] with $\beta = 1$ is the minimum of the following objective:

$$J_\beta(\boldsymbol{\theta}) \triangleq f(\mathbf{0};\boldsymbol{\psi}(\boldsymbol{\theta})) - \frac{2}{N}\sum_{n=1}^N f(\mathbf{x}_n;\boldsymbol{\theta}), \tag{S-167}$$

where $\boldsymbol{\psi}(\boldsymbol{\theta}) \triangleq [\boldsymbol{\eta}^T, 2\sigma^2]$. Hence, by equating the partial gradients of $J_\beta(\boldsymbol{\theta})$ (S-167) to zero, one can verify that the MβDE $\hat{\boldsymbol{\theta}}_\beta \triangleq [\hat{\boldsymbol{\eta}}_\beta^T, \hat{\sigma}_\beta^2]^T$ is the solution of the following fixed-point iterations:

$$\hat{\boldsymbol{\eta}}_\beta = \frac{\frac{1}{N}\sum_{n=1}^N \phi(\mathbf{x}_n; \hat{\boldsymbol{\eta}}_\beta, \hat{\sigma}_\beta^2)\mathbf{x}_n}{\frac{1}{N}\sum_{n=1}^N \phi(\mathbf{x}_n; \hat{\boldsymbol{\eta}}_\beta, \hat{\sigma}_\beta^2) - \frac{1}{4}\phi(\mathbf{0}; \hat{\boldsymbol{\eta}}_\beta, 2\hat{\sigma}_\beta^2)}$$

and

$$\hat{\sigma}_\beta^2 = \frac{\frac{1}{N}\sum_{n=1}^N s(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_\beta) + pf(\mathbf{0}, \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_\beta))}{\frac{2p}{N}\sum_{n=1}^N f(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_\beta) + \frac{1}{4\hat{\sigma}_\beta^2}\phi(\mathbf{0}; \hat{\boldsymbol{\eta}}_\beta, \hat{\sigma}_\beta^2)\|\hat{\boldsymbol{\theta}}_\beta\|^2},$$

where $s(\cdot;\cdot)$ is defined in (S-166).

E. MγDE ($\gamma = 1$)

Under the nominal density function in (20), it can be shown that the γ-divergence [s17] with tuning parameter $\gamma = 1$ is the minimum of the following objective:

$$J_\gamma(\boldsymbol{\theta}) \triangleq \frac{1}{2}\log f(\mathbf{0}; \tilde{\boldsymbol{\theta}}) - \log\frac{1}{N}\sum_{n=1}^N f(\mathbf{x}_n; \boldsymbol{\theta}), \tag{S-168}$$

where $\boldsymbol{\psi}(\boldsymbol{\theta})$ is defined below (S-167). Hence, by equating the partial gradients of $J_\gamma(\boldsymbol{\theta})$ (S-168) to zero, one can verify that the MγDE $\hat{\boldsymbol{\theta}}_\gamma \triangleq [\hat{\boldsymbol{\eta}}_\gamma^T, \hat{\sigma}_\gamma^2]^T$ is the solution of the following fixed-point iterations:

$$\hat{\boldsymbol{\eta}}_\gamma = \frac{\sum_{n=1}^N \tau(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_\gamma)\mathbf{x}_n}{\sum_{n=1}^N \tau(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_\gamma) - \frac{1}{4}V(\hat{\boldsymbol{\theta}}_\gamma)}$$

and

$$\hat{\sigma}_\gamma^2 = \frac{\sum_{n=1}^N t(\mathbf{x}_n; \hat{\boldsymbol{\theta}}_\gamma)}{\frac{1}{4\hat{\sigma}_\gamma^2}V(\hat{\boldsymbol{\theta}}_\gamma)\|\hat{\boldsymbol{\theta}}_\gamma\|^2 + \frac{p}{2}},$$

where

$$\tau(\mathbf{r}; \boldsymbol{\theta}) \triangleq \frac{\phi(\mathbf{r}; \boldsymbol{\eta}, \sigma^2)}{2\sum_{n=1}^N f(\mathbf{x}_n; \boldsymbol{\theta})}, \quad V(\boldsymbol{\theta}) \triangleq \frac{\phi(\mathbf{0}; \boldsymbol{\eta}, \sigma^2)}{\phi(\mathbf{0}; \boldsymbol{\eta}, \sigma^2) + \phi(\mathbf{0}; \mathbf{0}, \sigma^2)}, \quad t(\mathbf{r}; \boldsymbol{\theta}) \triangleq \frac{s(\mathbf{r}; \boldsymbol{\theta})}{2\sum_{n=1}^N f(\mathbf{x}_n; \boldsymbol{\theta})}$$

and $s(\cdot;\cdot)$ is defined in (S-166).

REFERENCES

[s1] T. M. Cover, *Elements of information theory,* John Wiley & Sons, 1999.

[s2] K. B. Athreya and S. N. Lahiri, *Measure theory and probability theory,* Springer-Verlag, 2006.

[s3] G. B. Folland, *Real Analysis,* John Wiley and Sons, 1984.

[s4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis,* third edition, John Wiley & Sons, 2003.

[s5] H. White, *Estimation, inference and specification analysis,* Cambridge university press, 1996.

[s6] R. J. Serfling, *Approximation theorems of mathematical statistics.* John Wiley & Sons, 1980.

[s7] W. K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica,* pp. 1161-1167, 1991.

[s8] C. H. Edwards, *Advanced calculus of several variables,* Courier Corporation, 2012.

[s9] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," *Handbook of econometrics,* vol. 4, pp. 2111-2245, 1994.

[s10] H. B. Mann and A. Wald, "On stochastic limit and order relationships," *Ann. Math. Stat.,* vol. 14, pp. 217-226, 1943.

[s11] M. Giaquinta and G. Modica, *Mathematical Analysis: An Introduction to Functions of Several Variables.* Birkhäuser, p. 88, Boston, 2009.

[s12] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses.* Springer Texts in Statistics, 2005.

[s13] D. R. Cox and D. V. Hinkley, *Theoretical Statistics,* Chapman & Hall, 1974.

[s14] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

[s15] A. Iqbal and A-K. Seghouane, "An α-divergence-based approach for robust dictionary learning," *IEEE Transactions on Image Processing,* vol. 28, no. 11, pp. 5729-5739, 2019.

[s16] A. Basu, I. R. Harris, N. L. Hjort and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika,* vol. 85, no. 3, pp. 549-559, 1998.

[s17] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis,* vol. 99, no. 9, pp. 2053-2081, 2008.