# Chapter 5: Influence Function of a (Log-)Density Function and Its Properties

From numerical examples and discussions in Chapter 4, we have seen that the regularized SM density estimates with and without the presence of an isolated observation can be very different. This motivates us to study the sensitivity of the density estimators to the input data. The tool we use an extension of the classic influence function.

We will give a review of the influence function and its applications in Section 5.1, and then introduce our extension of this classic notion to the studies of density estimators in Section 5.2. We will focus on the finite-dimensional exponential family and the infinite-dimensional kernel exponential family in Sections 5.3 and 5.4, respectively, and discuss various properties of the influence functions of ML and SM (log-)density projections (to be defined) in them.

## 5.1 Influence Function and Its Applications in Statistics

Influence function, a classical notion from robust statistics, was first introduced by Hampel (1968) to investigate the infinitesimal behavior of real- and vector-valued statistical functionals and has become one of the most important tools in robust statistics (Hampel et al., 1986).

Let $F$ be a distribution over the sample space $\mathcal{X}$, and $T$ to be a statistical functional, any function of $F$. In this section, we assume that $T$ is vector-valued so that $T(F) \in \mathbb{R}^m$. The *influence function* of $T$ at $F$ is defined to be

$$\mathrm{IF}(T, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big) - T(F)\Big) \tag{5.1}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big|_{\varepsilon=0}, \tag{5.2}$$

where $\delta_y$ denotes the point mass 1 at $y \in \mathcal{X}$. Inspecting (5.1) and (5.2), we see $\mathrm{IF}(T, F, y)$ is nothing but the Gâteaux derivative of the statistical functional $T$ at the distribution $F$ in the direction of the point mass $\delta_y$. Various properties of $\mathrm{IF}(T, F, y)$ have been discussed by Hampel (1974), Hampel et al. (1986), and Wasserman (2006).

The influence function has a nice interpretation. It measures the impact of an infinitesimally small amount of contamination of the original distribution $F$ at $y$ on the quantity of interest $T(F)$.

The *empirical influence function* (Definition 2.18 in Wasserman, 2006) is defined by letting $F$ in $\mathrm{IF}(T, F, y)$ be $F_n$. In addition, the *sample influence function* (Hampel, 1974; Cook and Weisberg, 1982) is defined to be

$$\mathrm{SIF}_\varepsilon(T, F_n, y) := \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F_n + \varepsilon\delta_y\big) - T(F_n)\Big), \tag{5.3}$$

where $\varepsilon > 0$ is tiny. Essentially, $\mathrm{SIF}_\varepsilon(T, F_n, y)$ is a finite-difference approximation of $\mathrm{IF}(T, F_n, y)$, and can be very useful when $T((1-\varepsilon)F_n + \varepsilon\delta_y)$ and $T(F_n)$ are analytically intractable or the limit by letting $\varepsilon \to 0^+$ is hard to handle. If, in particular, we let $\varepsilon = \frac{1}{n+1}$ in (5.3), we obtain Tukey's *sensitivity curve* which is used to assess the sensitivity of an estimator to the position of an additional observation *not* present in the sample (Hampel et al., 1986); and, if we let $\varepsilon = -\frac{1}{n-1}$ in (5.3), the resulting sample influence function has a close connection with the leave-one-out cross validation and the jackknife (see Example 3.18 in Wasserman, 2006, for a discussion).

We provide three examples below to illustrate the concept of the influence function.

*Example* 5.1 (Mean). Let $\mathcal{X} = \mathbb{R}$ and consider the statistical function defined by $T_1(F) = \int_{\mathbb{R}} x \mathrm{d}F(x) =: \mathbb{E}_F[X]$, where we assume $\mathbb{E}_F[X]$ exists. Since

$$\begin{aligned}\frac{1}{\varepsilon}\Big(T_1\big((1-\varepsilon)F + \varepsilon\delta_y\big) - T_1(F)\Big) &= \frac{1}{\varepsilon}\left(\int_{\mathcal{X}} x \mathrm{d}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \int_{\mathcal{X}} x \mathrm{d}F(x)\right) \\ &= \frac{1}{\varepsilon}\int_{\mathcal{X}} x \mathrm{d}(-\varepsilon F + \varepsilon\delta_y)(x) \\ &= \int_{\mathcal{X}} x \mathrm{d}(\delta_y - F)(x) \\ &= y - \mathbb{E}_F[X],\end{aligned}$$

we have

$$\mathrm{IF}(T_1, F, y) = y - \mathbb{E}_F[X]. \tag{5.4}$$

▶

*Example* 5.2 (Median). Let $\mathcal{X} = \mathbb{R}$ and consider the statistical function defined by $T_2(F) = F^{-1}(\frac{1}{2})$, the median of the distribution $F$. Here, we assume $F$ has a density function $p_0$ that is symmetric around 0 and $p_0(0) \neq 0$. Then, we have $F^{-1}(\frac{1}{2}) = 0$. Let $F_{\varepsilon,y} := (1-\varepsilon)F + \varepsilon\delta_y$.

First consider the case $y = 0$. Then, for any $\varepsilon \in (0,1)$, we have $F_{\varepsilon,0}(x) < \frac{1}{2}(1-\varepsilon)$ for all $x < 0$ and $F_{\varepsilon,0}(x) > \frac{1}{2}(1+\varepsilon)$ for all $x > 0$. It follows that $F_{\varepsilon,0}^{-1}(\frac{1}{2}) = 0$ and $\mathrm{IF}(T_2, F, 0) = 0$.

Now, suppose $y \neq 0$. We must have $\frac{1}{2} = F_{\varepsilon,y}\big(F_{\varepsilon,y}^{-1}(\frac{1}{2})\big)$, which is equivalent to

$$\frac{1}{2} = (1-\varepsilon)F\left(F_{\varepsilon,y}^{-1}\left(\frac{1}{2}\right)\right) + \varepsilon\delta_y\left(F_{\varepsilon,y}^{-1}\left(\frac{1}{2}\right)\right).$$

Differentiating both sides of the preceding equation with respect to $\varepsilon$ and evaluating at $\varepsilon = 0$, we have

$$0 = -F\left(F^{-1}\left(\frac{1}{2}\right)\right) + p_0\left(F^{-1}\left(\frac{1}{2}\right)\right)\left[\frac{\mathrm{d}F_{\varepsilon,y}^{-1}(\frac{1}{2})}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0}\right] + \delta_y\left(F^{-1}\left(\frac{1}{2}\right)\right)$$

$$= -\frac{1}{2} + p_0(0)\mathrm{IF}(T_2, F, y) + \delta_y(0).$$

Rearranging the preceding equation yields $\mathrm{IF}(T_2, F, y) = \frac{1-2\delta_y(0)}{2p_0(0)}$.

Now, if $y > 0$, $\delta_y(0) = 0$ and $\mathrm{IF}(T_2, F, y) = \frac{1}{2f(0)}$; if $y < 0$, $\delta_y(0) = 1$ and $\mathrm{IF}(T_2, F, y) = \frac{1-2}{2f(0)} = \frac{-1}{2f(0)}$. Summarizing all cases above, we have

$$\mathrm{IF}(T_2, F, y) = \frac{\mathrm{sign}(y)}{2f(0)}. \tag{5.5}$$

▶

*Example* 5.3 ($M$-estimator). The $M$-estimator, first proposed by Huber (1964), is a generalization of the maximum likelihood estimator and aims at designing new robust estimators. Here, we define an *$M$-estimator*, denoted by $T(F)$, to be the one that satisfies

$$\int_{\mathcal{X}} \psi(x, T(F))\mathrm{d}F(x) = 0,$$

where $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ and $\Theta \subseteq \mathbb{R}^m$ is the parameter space. If, in particular, we let $\psi(x, \theta) = \frac{\partial}{\partial \eta} \log f_\eta(x)|_{\eta=\theta}$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta$, where $\big\{f_\theta : \mathcal{X} \to [0, \infty) \mid \theta \in \Theta\big\}$ is the statistical model we assume, we obtain the maximum likelihood estimator.

We now derive the influence function of the $M$-estimator. Let $\varepsilon > 0$ be fixed. Under the distribution $(1-\varepsilon)F + \varepsilon\delta_y$, the corresponding $M$-estimator must satisfy

$$\int_{\mathcal{X}} \psi\big(x, T((1-\varepsilon)F + \varepsilon\delta_y)\big)\mathrm{d}\big((1-\varepsilon)F(x) + \varepsilon\delta_y(x)\big) = 0.$$

Differentiating both sides of the preceding equation with respect to $\varepsilon$ and using the chain rule yield

$$\int_{\mathcal{X}} \psi\big(x, T((1-\varepsilon)F + \varepsilon\delta_y)\big)\mathrm{d}(\delta_y(x) - F(x)) + \left[\int_{\mathcal{X}} \frac{\partial}{\partial t}\psi(x, t)\Big|_{t=T((1-\varepsilon)F+\varepsilon\delta_y)} \mathrm{d}\big((1-\varepsilon)F(x) + \varepsilon\delta_y(x)\big)\right] \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big) = 0.$$

Evaluating the preceding equation at $\varepsilon = 0$, we obtain

$$\int_{\mathcal{X}} \psi(x, \theta(F)) \mathrm{d}(\delta_y(x) - F(x)) + \left[ \int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right] \left[ \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T\big((1-\varepsilon)F + \varepsilon\delta_y\big) \Big|_{\varepsilon=0} \right] = 0.$$

Since $\int_{\mathcal{X}} \psi(x, T(F)) \mathrm{d}F(x) = 0$ by definition and $\int_{\mathcal{X}} \psi(x, T(F)) \mathrm{d}\delta_y(x) = \psi(y, T(F))$, we can simplify the preceding equation as

$$\psi(y, T(F)) + \left[ \int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(y,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right] \mathrm{IF}(T, F, x) = 0.$$

Finally, if we assume the $m \times m$ matrix $-\int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t)\big|_{t=T(F)} \mathrm{d}F(x)$ is invertible, we obtain

$$\mathrm{IF}(T, F, y) = \left( - \int_{\mathcal{X}} \frac{\partial}{\partial t} \psi(x,t) \Big|_{t=T(F)} \mathrm{d}F(x) \right)^{-1} \psi(y, \theta(F)).$$

▶

In the influence function approach to robustness, an estimator is regarded to be robust if its *gross-error sensitivity*, $\sup_{y \in \mathcal{X}} \|\mathrm{IF}(T, F, y)\|_2$, is finite. Here, the gross-error sensitivity measures the worst influence which a small amount of contamination can have on the value of the estimator (Hampel et al., 1986).

Let us go back to Examples 5.1 and 5.2 discussed earlier and look their gross-error sensitivities. From (5.4) and (5.5), we have $\sup_{y \in \mathcal{X}} |\mathrm{IF}(T_1, F, y)| = \infty$ and $\sup_{y \in \mathcal{X}} |\mathrm{IF}(T_2, F, y)| = \frac{1}{2f(0)} < \infty$, respectively, from which we conclude that the median is more robust than the mean.

Since being introduced into the world of statistics in the 1960s, the influence function has found a wide range of applications. They can be used to understand the robustness properties of various estimators, to design new estimators with certain robustness properties (Hampel et al., 1986), to identify influential observations in

model fitting (Cook, 1977; Cook and Weisberg, 1982), to perform model validation (Debruyne, Hubert, and Suykens, 2008; Koh and Liang, 2017), and to design efficient subsampling algorithms to reduce the computational load in training machine learning models (Ting and Brochu, 2018; Raj et al., 2020).

## 5.2 Extension of the Influence Function in Density Estimation Problem

Even though the influence function has been used in various statistical applications, it has not been used much in the density estimation problem.

The main difficulty is that the influence function was traditionally defined for real- or vector-valued statistical functionals. But, the object of primary interest in the density estimation problem is a function, or more precisely, a pdf. We need to extend the definition of the influence function by allowing the statistical functional therein to be function-valued.

We now present our approach. Let $T$ be a map from the collection of distribution functions over $\mathcal{X}$ to the class of log-density functions over $\mathcal{X}$, and $\widetilde{T}$ be a map from the collection of distribution functions over $\mathcal{X}$ to the class of density functions over $\mathcal{X}$; that is, if $F$ is a distribution function over $\mathcal{X}$, $T(F)$ is a log-density function and $\widetilde{T}(F)$ is a density function over $\mathcal{X}$. Then, we define the *influence functions of* $T(F)$ *and* $\widetilde{T}(F)$ *evaluated at* $x \in \mathcal{X}$ to be

$$\mathrm{IF}_x(T, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big), \tag{5.6}$$

$$\mathrm{IF}_x(\widetilde{T}, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \widetilde{T}(F)(x)\Big), \tag{5.7}$$

respectively. Note that both $T$ and $\widetilde{T}$ are function-valued statistical functionals, and assign to a distribution over $\mathcal{X}$ a log-density function and density function over $\mathcal{X}$, respectively. In addition, $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ are both real-valued as they

only depend on the evaluations of $T$ and $\widetilde{T}$, respectively, at different distribution functions.

In (5.6) (*resp.* (5.7)), $\mathrm{IF}_x(T, F, y)$ (*resp.* $\mathrm{IF}_x(\widetilde{T}, F, y)$) is the Gâteaux derivative of $T$ (*resp.* $\widetilde{T}$) at $F$ in the direction of $\delta_y$ evaluated at $x$, and describes how the value of $T(F)$ (*resp.* $\widetilde{T}(F)$) at $x$ is affected by $y$. In particular, $\mathrm{IF}_x(T, F, y) > 0$ means the value of the log-density function at $x$ increases with the presence of $y$, and $\mathrm{IF}_x(T, F, y) < 0$ means the value of the log-density function at $x$ decreases with the presence of $y$. A similar interpretation can be extended to $\mathrm{IF}_x(\widetilde{T}, F, y)$, simply by replacing "log-density function" with "density function".

The following proposition establishes the relationship between $\mathrm{IF}_x(\widetilde{T}, F, y)$ and $\mathrm{IF}_x(T, F, y)$.

**Proposition 5.1.** *Suppose* $\widetilde{T}(F)(x) = \exp(T(F)(x))$ *for all* $x \in \mathcal{X}$. *Then, we have*

$$\mathrm{IF}_x(\widetilde{T}, F, y) = \widetilde{T}(F)(x)\mathrm{IF}_x(\widetilde{T}, F, y), \qquad \text{for all } x \in \mathcal{X}. \tag{5.8}$$

The proof of Proposition 5.1 can be found in Section 5.5.1.

From (5.8), we can view $\mathrm{IF}_x(\widetilde{T}, F, y)$ as a weighted version of $\mathrm{IF}_x(T, F, y)$, where the weight is $\widetilde{T}(x)$, the value of the unperturbed density function at $x$.

In addition, the following proposition establishes the relationship among the KL-divergence, $\mathrm{IF}_x(T, F, y)$, and $\mathrm{IF}_x(\widetilde{T}, F, y)$.

**Proposition 5.2.** *Suppose* $\widetilde{T}(F)(x) = \exp(T(F)(x))$ *for all* $x \in \mathcal{X}$, $\int_{\mathcal{X}} \widetilde{T}(F)(x)|T((1-\varepsilon)F + \varepsilon\delta_y)(x)|\mathrm{d}x < \infty$ *for all* $\varepsilon \in [0, 1]$, *and* $\int_{\mathcal{X}} |\mathrm{IF}_x(\widetilde{T}, F, y)|\mathrm{d}x < \infty$. *Then, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}((1-\varepsilon)F + \varepsilon\delta_y)\Big)\Big|_{\varepsilon=0} &= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y)\mathrm{d}x \\ &= -\int_{\mathcal{X}} \mathrm{IF}_x(\widetilde{T}, F, y)\mathrm{d}x.\end{aligned}$$

We next use a simple example to illustrate the definitions of the influence functions of $T(F)$ and $\widetilde{T}(F)$ evaluated at $x \in \mathcal{X}$.

*Example* 5.4 (Normal location model). Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Q}$ contain all pdfs of the form

$$q_\theta(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \text{ for all } x \in \mathcal{X}, \qquad \theta \in \mathbb{R}.$$

Let $T(F) = \log q_{\theta(F)}$ and $\widetilde{T}(F) = q_{\theta(F)}$, where

$$\theta(F) := \arg\max_{\theta \in \mathbb{R}} \left\{ \int_{\mathcal{X}} \log q_\theta(x) \mathrm{d}F(x) \right\} = \mathbb{E}_F[X],$$

and we assume $\mathbb{E}_F[X]$ exists. If we suppose $F$ has a pdf, say $p_0$, then $q_{\theta(F)}$ is the ML density projection of $p_0$ onto the family $\mathcal{Q}$. To see this is a *projection*, suppose $p_0$ satisfies $\int_{\mathcal{X}} p_0(x)|\log p_0(x)|\mathrm{d}x < \infty$ and $\int_{\mathcal{X}} p_0(x)|\log q_\theta(x)|\mathrm{d}x < \infty$ for all $\theta \in \mathbb{R}$, and notice that $q_{\theta(F)}$ minimizes $\mathrm{KL}(p_0 \| q_\theta)$ over all $q_\theta \in \mathcal{Q}$; that is, $q_{\theta(F)}$ has the smallest KL-divergence to $p_0$ among all pdfs in $\mathcal{Q}$.

By simple algebra, we have, for all $x \in \mathcal{X}$,

$$\begin{aligned} \mathrm{IF}_x(T, F, y) &= (y - \mathbb{E}_F[X])(x - \mathbb{E}_F[X]), \\ \mathrm{IF}_x(\widetilde{T}, F, y) &= q_{\theta(F)}(x)(y - \mathbb{E}_F[X])(x - \mathbb{E}_F[X]). \end{aligned}$$

If we let $\mathbb{E}_F[X] = 0$ and $y = 2$, $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ evaluated at different values of $x$ are shown in Figure 5.1. ▶

If we fix $y$ and $F$ and view $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(\widetilde{T}, F, y)$ as functions of the evaluation point $x$, both can vary with $x$, which is obvious from Figure 5.1. In other words, with a fixed $F$, a fixed $y$ can have different effects on each of $T(F)$ and $\widetilde{T}(F)$ at different evaluation points. In order to have a summarizing quantity to describe the maximal possible effect of $y$ on $T(F)$ and $\widetilde{T}(F)$, we use

$$M(T, F, y) := \sup_{x \in \mathcal{X}} |\mathrm{IF}_x(T, F, y)|, \qquad \text{and} \qquad M(\widetilde{T}, F, y) := \sup_{x \in \mathcal{X}} |\mathrm{IF}_x(\widetilde{T}, F, y)|,$$

and call them the *overall influences* of $y$ on $T(F)$ and $\widetilde{T}(F)$, respectively. They describe the maximal possible effects of $y$ on $T(F)$ and $\widetilde{T}(F)$, respectively.

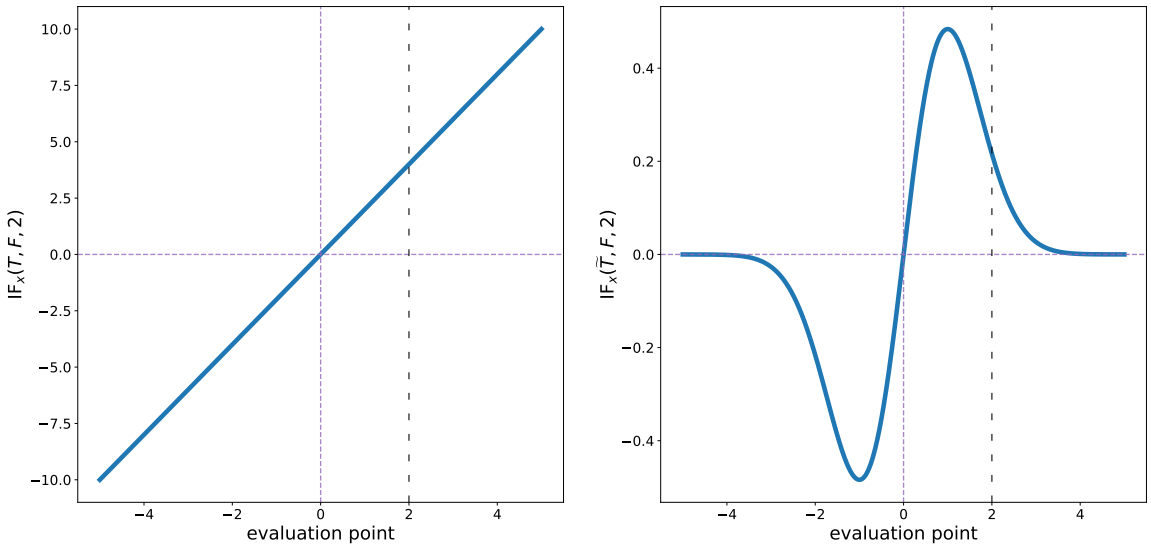


Figure 5.1: $\mathrm{IF}_x(T, F, y)$ (left panel) and $\mathrm{IF}_x(\widetilde{T}, F, y)$ (right panel) evaluated at different $x \in \mathcal{X}$ with $\mathbb{E}_F[X] = 0$ and $y = 2$. The black dashed vertical line indicates the location of the contaminant $y$.

*Example* 5.4 (Normal location model, continued). The overall influences of $y$ on $T(F)$ and $\widetilde{T}(F)$ are

$$M(T, F, y) = \begin{cases} 0, & \text{if } y = \mathbb{E}_F[X] \\ \infty, & \text{otherwise} \end{cases}, \quad \text{and} \quad M(\widetilde{T}, F, y) = \frac{1}{\sqrt{2e\pi}}\big|y - \mathbb{E}_F[X]\big|,$$

respectively. ▶

Finally, we define the *sample influence functions of* $T(F)$ *and* $\widetilde{T}(F)$ *evaluated at* $x \in \mathcal{X}$ to be

$$\mathrm{SIF}_{x,\varepsilon}(T, F, y) := \frac{1}{\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big),$$
$$\mathrm{SIF}_{x,\varepsilon}(\widetilde{T}, F, y) := \frac{1}{\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - \widetilde{T}(F)(x)\Big),$$

respectively, and the corresponding *sample overall influences* to be

$$\widehat{M}_\varepsilon(T, F, y) := \sup_{x\in\mathcal{X}}|\mathrm{SIF}_{x,\varepsilon}(T, F, y)|, \qquad \text{and} \qquad \widehat{M}_\varepsilon(\widetilde{T}, F, y) := \sup_{x\in\mathcal{X}}|\mathrm{SIF}_{x,\varepsilon}(\widetilde{T}, F, y)|$$

respectively. Similarly to (5.3), $\mathrm{SIF}_{x,\varepsilon}(T,F,y)$ (*resp.* $\mathrm{SIF}_{x,\varepsilon}(\widetilde{T},F,y)$) is the finite-difference approximation of $\mathrm{IF}_x(T,F,y)$ (*resp.* $\mathrm{IF}_{x,\varepsilon}(\widetilde{T},F,y)$).

## 5.3 Influence Function of (Log-)Density Projection in a Finite-dimensional Exponential Family

With the introduction of $\mathrm{IF}_x(T,F,y)$ and $\mathrm{IF}_x(\widetilde{T},F,y)$ in the preceding section, we focus on the influence functions of the ML and SM (log-)density projections (to be defined later) in an $m$-dimensional exponential family $\mathcal{Q}_{\mathrm{fin}}$ (introduced in Chapter 2) in this section.

Recall $\mathcal{Q}_{\mathrm{fin}}$ contains all pdfs of the form

$$\tilde{q}_\theta(x) := \mu(x)\exp\big(\langle\theta,\varphi(x)\rangle - B(\theta)\big) \text{ for all } x\in\mathcal{X}, \qquad \theta\in\Theta,$$

where we assume each component of $\varphi$, $\varphi_j : \mathcal{X}\to\mathbb{R}$, is twice continuously differentiable. We define the following maps

$$\begin{aligned} T(F) &= \log\tilde{q}_{\theta_{\mathrm{ML},F}}, && \text{and} && S(F) = \log\tilde{q}_{\theta_{\mathrm{SM},F}},\\ \widetilde{T}(F) &= \tilde{q}_{\theta_{\mathrm{ML},F}}, && \text{and} && \widetilde{S}(F) = \tilde{q}_{\theta_{\mathrm{SM},F}}, \end{aligned}$$

where $F$ is a distribution function over $\mathcal{X}$,

$$\begin{aligned} \theta_{\mathrm{ML},F} &:= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\int_{\mathcal{X}} -\log\tilde{q}_\theta(x)\mathrm{d}F(x)\right\}\\ &= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{B(\theta) - \Big\langle\theta,\mathbb{E}_F[\varphi(X)]\Big\rangle\right\},\\ \theta_{\mathrm{SM},F} &:= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\int_{\mathcal{X}}\sum_{u=1}^{d}\left(\frac{1}{2}\big(\partial_u\log\tilde{q}_\theta(x)\big)^2 + \partial_u\log\tilde{q}_\theta(x)\right)\mathrm{d}F(x)\right\}\\ &= \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\frac{1}{2}\theta^\top\mathbb{E}_F[D_1(X)D_1(X)^\top]\theta - \theta^\top\mathbb{E}_F[W(X)]\right\}, \end{aligned}$$

$D_1$ is a map from $\mathcal{X}$ to $\mathbb{R}^{m\times d}$ with $[D_1(x)]_{j,u} = \partial_u\varphi_j(x)$ for all $x\in\mathcal{X}$, $D_2$ is a map from $\mathcal{X}$ to $\mathbb{R}^{m\times d}$ with $[D_2(x)]_{j,u} = \partial_u^2\varphi_j(x)$ for all $x\in\mathcal{X}$, $W$ is a map from $\mathcal{X}$ to $\mathbb{R}^m$ with

$W(x) = -(D_1(x)\nabla \log \mu(x) + D_2(x)\mathbf{1}_d)$ for all $x \in \mathcal{X}$, and $\mathbf{1}_d := (1, \cdots, 1)^\top \in \mathbb{R}^d$. Supposing that the pdf of $F$, denoted by $p_0$, exists, we observe that $\widetilde{T}(F)$ and $\widetilde{S}(F)$ are the ML and SM density projections of $p_0$ onto the family $\mathcal{Q}_{\text{fin}}$, as they have the smallest KL- and H-divergences to $p_0$ among all $\tilde{q}_\theta \in \mathcal{Q}_{\text{fin}}$, respectively.

Then, the following theorem provides explicit expressions for $\text{IF}_x(T, F, y)$ and $\text{IF}_x(S, F, y)$. The expressions for $\text{IF}_x(\widetilde{T}, F, y)$ and $\text{IF}_x(\widetilde{S}, F, y)$ are easy to obtain by multiplying $\text{IF}_x(T, F, y)$ and $\text{IF}_x(S, F, y)$ with $\widetilde{T}(F)(x)$ and $\widetilde{S}(F)(x)$, respectively, using Proposition 5.1.

**Theorem 5.1** (Influence functions of ML and SM log-density projections in $\mathcal{Q}_{\text{fin}}$ evaluated at $x \in \mathcal{X}$)**.**

*(a) Assume* $\mathbb{E}_F[\varphi(X)]$ *exists and belongs to* $\text{int}(\Theta)$, $\int_{\mathcal{X}} \tilde{q}_{\theta_{\text{ML},F}}(x)\|\varphi(x)\|_2^2 \text{d}x < \infty$, *and* $\nabla^2 B(\theta_{\text{ML},F})$ *is invertible. Then, for all* $x \in \mathcal{X}$,

$$\text{IF}_x(T, F, y) = \left\langle \widetilde{G}_{\text{ML}}(F, y), \varphi(x) - \int_{\mathcal{X}} \tilde{q}_{\theta_{\text{ML},F}}(w)\varphi(w)\text{d}w \right\rangle, \tag{5.9}$$

*where*

$$\widetilde{G}_{\text{ML}}(F, y) := \left[\nabla^2 B(\theta_{\text{ML},F})\right]^{-1}\left(\varphi(y) - \mathbb{E}_F[\varphi(X)]\right)$$

*(b) Assume* $\mathbb{E}_F[W(X)]$ *exists,* $\int_{\mathcal{X}} \tilde{q}_{\theta_{\text{SM},F}}(x)\|\varphi(x)\|_2^2 \text{d}x < \infty$, *and* $\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]$ *is invertible. Then, for all* $x \in \mathcal{X}$,

$$\text{IF}_x(T, F, y) = \left\langle \widetilde{G}_{\text{SM}}(F, y), \varphi(x) - \int_{\mathcal{X}} \tilde{q}_{\theta_{\text{SM},F}}(w)\varphi(w)\text{d}w \right\rangle, \tag{5.10}$$

*where*

$$\widetilde{G}_{\text{SM}}(F, y) := \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \times \left\{W(y) - D_1(y)D_1(y)^\top \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \mathbb{E}_F[W(X)]\right\}.$$

The proof of Theorem 5.1 will be provided in Section 5.5.3.

From the proof, we can see $\widetilde{G}_{\mathrm{ML}}(F, y)$ is the influence function of the $M$-estimator $\theta_{\mathrm{ML},F}$ that satisfies

$$0 = \int_{\mathcal{X}} \Big( \nabla B(\theta_{\mathrm{ML},F}) - \varphi(x) \Big) \mathrm{d}F(x),$$

and $\widetilde{G}_{\mathrm{SM}}(F, y)$ is the influence function of the $M$-estimator $\theta_{\mathrm{SM},F}$ that satisfies

$$0 = \mathbb{E}_F \big[ D_1(X) D_1(X)^\top \big] \theta_{\mathrm{SM},F} - \mathbb{E}_F[W(X)];$$

in other words, they are the influence functions of their respective natural parameters.

Comparing (5.9) and (5.10), we see that $\mathrm{IF}_x(T, F, y)$ depends on $y$ *only* through the canonical statistics $\varphi$, but $\mathrm{IF}_x(S, F, y)$ depends on $y$ through the first two derivatives of $\varphi$ and the first derivative of $\log \mu$. This is the key difference between $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(S, F, y)$. In the examples we will see later, their differences will become more apparent.

From Theorem 5.1, we can obtain some properties of $\mathrm{IF}_x(T, F, y)$ and $\mathrm{IF}_x(S, F, y)$ which are given in the corollaries below.

**Corollary 5.1.**

*(a) Under the assumptions in Theorem 5.1(a),* $\mathrm{IF}_x(T, F, y) = 0$ *for all* $x \in \mathcal{X}$ *if and only if* $F$ *and* $y$ *satisfy*

$$\varphi(y) - \mathbb{E}_F[\varphi(X)] = 0. \tag{5.11}$$

*(b) Under the assumptions in Theorem 5.1(b),* $\mathrm{IF}_x(S, F, y) = 0$ *for all* $x \in \mathcal{X}$ *if and only if* $F$ *and* $y$ *satisfy*

$$W(y) - D_1(y) D_1(y)^\top \Big[ \mathbb{E}_F \big[ D_1(X) D_1(X)^\top \big] \Big]^{-1} \mathbb{E}_F[W(X)] = 0. \tag{5.12}$$

**Corollary 5.2.**

*(a) Under the assumptions in Theorem 5.1(a), suppose $F$ and $y$ satisfy $\varphi(y) - \mathbb{E}_F[\varphi(X)] \neq 0$.*

*(i) If $\sup_{j=1,\cdots,m} \sup_{x\in\mathcal{X}}|\varphi_j(x)| < \infty$, then $M(T,F,y) < \infty$.*

*(ii) If there exists $j^* \in \{1,\cdots,m\}$ and $x_0 \in \overline{\mathcal{X}}$ such that $\lim_{x\to x_0}|\varphi_{j^*}(x)| = \infty$ so that $\sup_{x\in\mathcal{X}}|\varphi_j(x)| = \infty$, $\varphi_{j^*}(y) - \mathbb{E}_F[\varphi_{j^*}(X)] \neq 0$, and $\lim_{x\to x_0} \frac{\varphi_j(x)}{\varphi_{j^*}(x)} = 0$ for all $j \neq j^*$, then $M(T,F,y) = \infty$.*

*(b) Under the assumptions in Theorem 5.1(b), suppose $F$ and $y$ satisfy*

$$W(y) - D_1(y)D_1(y)^\top \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \mathbb{E}_F[W(X)] \neq 0.$$

*(i) If $\sup_{j=1,\cdots,m} \sup_{x\in\mathcal{X}}|\varphi_j(x)| < \infty$, then $M(S,F,y) < \infty$.*

*(ii) If there exists $j^* \in \{1,\cdots,m\}$ and $x_0 \in \overline{\mathcal{X}}$ such that $\lim_{x\to x_0}|\varphi_{j^*}(x)| = \infty$ so that $\sup_{x\in\mathcal{X}}|\varphi_j(x)| = \infty$, the $j^*$-th element of*

$$W(y) - D_1(y)D_1(y)^\top \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \mathbb{E}_F[W(X)]$$

*is nonzero, and $\lim_{x\to x_0} \frac{\varphi_j(x)}{\varphi_{j^*}(x)} = 0$ for all $j \neq j^*$, then $M(S,F,y) = \infty$.*

These corollaries are obvious from Theorem 5.1, in particular, from (5.9) and (5.10), and their proofs are omitted here. In the rest of this section, we provide several examples to illustrate Theorem 5.1 and Corollaries 5.1 and 5.2.

*Example* 5.5 (Normal location-scale model). Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\omega)^2}{2\sigma^2}\right) \text{ for all } x \in \mathcal{X}, \qquad \theta := (\omega, \sigma^2) \in \Theta,$$

where $\Theta := \mathbb{R} \times (0, \infty)$. In this example, we have

$$\mu(x) = \frac{1}{\sqrt{2\pi}}, \qquad \varphi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \qquad B(\theta) = \frac{\omega^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2,$$

and

$$D_1(x) = \begin{pmatrix} 1 \\ 2x \end{pmatrix}, \qquad D_2(x) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \qquad (\log\mu)'(x) = 0, \qquad W(x) = \begin{pmatrix} 0 \\ -2 \end{pmatrix}.$$

Assume $m_1 := \mathbb{E}_F[X]$ and $m_2 := \mathbb{E}_F[X^2]$ both exist and $m_2 - m_1^2 > 0$. Then, using Theorem 5.1, we have

$$\mathrm{IF}_x(T, F, y) = \mathrm{IF}_x(S, F, y) = a_1 x^2 + a_2 x + a_3, \qquad \text{for all } x \in \mathcal{X},$$

where

$$\begin{aligned} a_1 &:= \frac{-m_1}{(m_2 - m_1^2)^2}(y - m_1) + \frac{1}{2(m_2 - m_1^2)^2}(y^2 - m_2), \\ a_2 &:= \frac{m_2 + m_1^2}{(m_2 - m_1^2)^2}(y - m_1) - \frac{m_1}{(m_2 - m_1^2)^2}(y^2 - m_2), \\ a_3 &:= \frac{m_1^3 - 2m_1 m_2}{(m_2 - m_1^2)^2}(y - m_1) + \frac{m_2}{2(m_2 - m_1^2)^2}(y^2 - m_2). \end{aligned}$$

Since, in this example, we have $\varphi_1(x) = x$ and $\varphi_2(x) = x^2$ for all $x \in \mathcal{X}$, and $\sup_{x\in\mathcal{X}} |\varphi_2(x)| = \infty$ with $\lim_{x\to\pm\infty} \varphi_2(x) = \infty$, and $\lim_{x\to\pm\infty} \frac{\varphi_1(x)}{\varphi_2(x)} = 0$, we have $M(T, F, y) = M(S, F, y) = \infty$, which verifies Corollary 5.2. ▶

*Example* 5.6 (Lognormal location model). Let $\mathcal{X} = (0, \infty)$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x - \theta)^2}{2}\right) \text{ for all } x \in \mathcal{X}, \qquad \theta \in \mathbb{R}.$$

In this example, we have

$$\mu(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\log x)^2}{2}\right), \qquad \varphi(x) = \log x, \qquad B(\theta) = \frac{\theta^2}{2},$$

and

$$D_1(x) = \frac{1}{x}, \quad D_2(x) = -\frac{1}{x^2}, \quad (\log \mu)'(x) = -\frac{1}{x} - \frac{\log x}{x}, \quad W(x) = \frac{2}{x^2} + \frac{\log x}{x^2}.$$

Then, using Theorem 5.1, we obtain

$$\begin{aligned} \mathrm{IF}_x(T, F, y) &= \big(\log y - m_1\big)\big(\log x - m_1\big), \\ \mathrm{IF}_x(S, F, y) &= \frac{1}{m_3 y^2}\left(\log y - \frac{m_2}{m_3}\right)\left(\log x - 2 - \frac{m_2}{m_3}\right), \end{aligned}$$

for all $x \in \mathcal{X}$, where we assume $m_1 := \mathbb{E}_F[\log X]$, $m_2 := \mathbb{E}_F[X^{-2} \log X]$ and $m_3 := \mathbb{E}_F[X^{-2}]$ exist.

In particular, $\mathrm{IF}_x(T, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if $F$ and $y$ satisfy $\log y = m_1$, which is exactly (5.11); and, $\mathrm{IF}_x(S, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if $F$ and $y$ satisfy $\log y - \frac{m_2}{m_3} = 0$, which is exactly (5.12). These illustrate Corollary 5.1.

Also, note that, in this example, $\sup_{x \in \mathcal{X}} |\varphi(x)| = \sup_{x \in \mathcal{X}} |\log x| = \infty$. Supposing $\log y - \mathbb{E}_F[\log X] \neq 0$, we have $M(T, F, y) = \infty$, which illustrates Corollary 5.2(a); supposing $\log y - \frac{m_2}{m_3} \neq 0$, we have $M(S, F, y) = \infty$, which illustrates Corollary 5.2(b). ▶

*Example* 5.7 (Gamma rate model). Let $\mathcal{X} = (0, \infty)$ and $\mathcal{Q}$ contain all pdfs of the form

$$\tilde{q}_\theta(x) := \frac{x^{\alpha-1}}{\Gamma(\alpha)} \exp\Big(-\theta x + \alpha \log \theta\Big) \text{ for all } x \in \mathcal{X}, \qquad \theta \in (0, \infty).$$

We assume $\alpha > 1$ is known. In this example

$$\mu(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}, \qquad \varphi(x) = -x, \qquad B(\theta) = -\alpha \log \theta,$$

and

$$D_1(x) = -1, \quad D_2(x) = 0, \quad (\log \mu)'(x) = \frac{\alpha - 1}{x}, \quad W(x) = \frac{\alpha - 1}{x}.$$

Then, using Theorem 5.1, we obtain

$$\begin{aligned}\mathrm{IF}_x(T, F, y) &= \frac{\alpha}{m_1^2}(y - m_1)(x - m_1),\\ \mathrm{IF}_x(S, F, y) &= (\alpha - 1)(m_2 - y^{-1})\left(x - \frac{\alpha}{\alpha - 1}\frac{1}{m_2}\right),\end{aligned}$$

for all $x \in \mathcal{X}$, where we assume $m_1 := \mathbb{E}_F[X]$ and $m_2 := \mathbb{E}_F[X^{-1}]$ exist.

In particular, $\mathrm{IF}_x(T, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if $F$ and $y$ satisfy $y = m_1$, which is exactly (5.11); and, $\mathrm{IF}_x(S, F, y) = 0$ for all $x \in \mathcal{X}$ if and only if $F$ and $y$ satisfy $y^{-1} - m_2 = 0$, which is exactly (5.12). These illustrate Corollary 5.1.

Note that, in this example, $\sup_{x\in\mathcal{X}}|\varphi(x)| = \sup_{x\in\mathcal{X}}|-x| = \infty$. Supposing $\mathbb{E}_F[X] - y \neq 0$, we have $M(T, F, y) = \infty$, which illustrates Corollary 5.2(a); supposing $m_2 - y^{-1} \neq 0$, we have $M(S, F, y) = \infty$, which illustrates Corollary 5.2(b). ▶

## 5.4 Influence Function of (Log-)Density Projection in a Kernel Exponential Family

We now turn to the influence functions of the ML and SM (log-)density projections in a kernel exponential family $\mathcal{Q}_{\mathrm{ker}}$ (introduced in Chapter 2) in this section. Recall $\mathcal{Q}_{\mathrm{ker}}$ contains all pdfs of the form

$$q_f(x) := \mu(x)\exp\big(f(x) - A(f)\big) \text{ for all } x \in \mathcal{X}, \qquad f \in \mathcal{F} \subseteq \mathcal{H}.$$

We still let $F$ be a distribution function over $\mathcal{X}$, and define the following maps

$$\begin{aligned} T_\lambda(F) &= \log q_{f^{(\lambda)}_{\mathrm{ML},F}}, && \text{and} && S_\rho(F) = \log q_{f^{(\rho)}_{\mathrm{SM},F}},\\ \widetilde{T}_\lambda(F) &= q_{f^{(\lambda)}_{\mathrm{ML},F}}, && \text{and} && \widetilde{S}_\rho(F) = q_{f^{(\rho)}_{\mathrm{SM},F}},\end{aligned}$$

where

$$f^{(\lambda)}_{\mathrm{ML},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}}\left\{A(f) - \int_{\mathcal{X}} f(x)\mathrm{d}F(x) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}\right\}, \tag{5.13}$$

$$f_{\mathrm{SM},F}^{(\rho)} := \arg\min_{f\in\mathcal{F}}\left\{\frac{1}{2}\langle f, C_F f\rangle_{\mathcal{H}} - \langle f, z_F\rangle_{\mathcal{H}} + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2\right\}, \tag{5.14}$$

with $C_F := \int_{\mathcal{X}} \sum_{u=1}^{d} \partial_u k(x,\cdot) \otimes \partial_u k(x,\cdot) \mathrm{d}F(x)$ mapping from $\mathcal{H}$ to $\mathcal{H}$, and $z_F := -\int_{\mathcal{X}} \sum_{u=1}^{d} (\partial_u \log \mu(x) \partial_u k(x,\cdot) + \partial_u^2 k(x,\cdot)) \mathrm{d}F(x) \in \mathcal{H}$. If the distribution function $F$ has a pdf $p_0$, the resulting $C_F$ is (3.1) defined in Chapter 3; if $F = F_n$, the empirical distribution function of data $X_1, \cdots, X_n$, the resulting $C_{F_n}$ is $\widehat{C}$ defined in (2.14) in Chapter 2.

Note that both objective functionals in (5.13) and (5.14) are strongly convex with constants $\lambda > 0$ and $\rho > 0$, respectively. It follows that each of $f_{\mathrm{ML},F}^{(\lambda)}$ and $f_{\mathrm{SM},F}^{(\rho)}$ exists and is unique (Corollary 11.17 in Bauschke and Combettes, 2011).

We then have the following results on the influence functions of the penalized ML and SM log-density projections onto $\mathcal{Q}_{\mathrm{ker}}$ evaluated at $x \in \mathcal{X}$.

**Theorem 5.2** (Influence functions of ML and SM log-density projections in $\mathcal{Q}_{\mathrm{ker}}$ evaluated at $x \in \mathcal{X}$)**.**

*(a) Under* ***(A1)*** *-* ***(A4)*** *in Chapter 2, we have, for all* $x \in \mathcal{X}$*,*

$$\mathrm{IF}_x(T_\lambda, F, y) = \left\langle G_{\mathrm{ML}}(F,y), k(x,\cdot) - \int_{\mathcal{X}} k(w,\cdot) q_{f_{\mathrm{ML},F}^{(\lambda)}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}, \tag{5.15}$$

*where*

$$G_{\mathrm{ML}}(F,y) := \left\{ \mathbb{E}_{q_{f_{\mathrm{ML},F}^{(\lambda)}}}\left[\left(k(X,\cdot) - \Upsilon\right) \otimes \left(k(X,\cdot) - \Upsilon\right)\right] + \lambda I \right\}^{-1} \left( k(y,\cdot) - \int_{\mathcal{X}} k(x,\cdot)\mathrm{d}F(x) \right) \in \mathcal{H},$$

*and* $\Upsilon := \int_{\mathcal{X}} k(x,\cdot) q_{f_{\mathrm{ML},F}^{(\lambda)}}(x)\mathrm{d}x \in \mathcal{H}$*.*

*(b) Under* ***(A1)*** *-* ***(A4)*** *in Chapter 2 and* ***(B1)*** *-* ***(B5)*** *in Chapter 3, we have, for all* $x \in \mathcal{X}$*,*

$$\mathrm{IF}_x(S_\rho, F, y) = \left\langle G_{\mathrm{SM}}(F,y), k(x,\cdot) - \int_{\mathcal{X}} k(w,\cdot) q_{f_{\mathrm{SM},F}^{(\rho)}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}, \tag{5.16}$$

*where*

$$G_{\mathrm{SM}}(F, y) := (C_F + \rho I)^{-1}\Big(z_{\delta_y} - (C_{\delta_y} + \rho I)(C_F + \rho I)^{-1} z_F\Big) \in \mathcal{H}.$$

The proof of Theorem 5.2 can be found in Section 5.5.4.

Comparing Theorems 5.1 and 5.2, we see they share many similarities. In particular, $\mathrm{IF}_x(T_\lambda, F, y)$ depends on $y$ through $k$ *only*, but $\mathrm{IF}_x(S_\rho, F, y)$ depends on $y$ through the first two partial derivatives of $k$ and the first derivative of $\log \mu$. Moreover, $G_{\mathrm{ML}}(F, y)$ is the influence function of the $M$-estimator $f^{(\lambda)}_{\mathrm{ML},F}$ that satisfies

$$0 = \int_{\mathcal{X}} \Big(\nabla A(f^{(\lambda)}_{\mathrm{ML},F}) - k(x, \cdot) + \lambda f^{(\lambda)}_{\mathrm{ML},F}\Big) \mathrm{d}F(x),$$

and $G_{\mathrm{SM}}(F, y)$ is the influence function of the $M$-estimator $f^{(\rho)}_{\mathrm{SM},F}$ that satisfies

$$0 = C_F f^{(\rho)}_{\mathrm{SM},F} - z_F + \rho f^{(\rho)}_{\mathrm{ML},F}.$$

The covariance operator

$$\mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\Big[\big(k(X, \cdot) - \Upsilon\big) \otimes \big(k(X, \cdot) - \Upsilon\big)\Big]$$

appearing in $G_{\mathrm{ML}}(F, y)$ plays the role of $\nabla^2 B(\theta_{\mathrm{ML},F})$ in $\widetilde{G}_{\mathrm{ML}}(F, y)$, where $\nabla^2 B(\theta_{\mathrm{ML},F})$ is the covariance matrix of $\varphi$ under $\tilde{q}_{\theta_{\mathrm{ML},F}}$. The operator $C_F$ appearing in $G_{\mathrm{SM}}(F, y)$ plays the role of $\mathbb{E}_F[D_1(X)D_1(X)^\top]$ in $\widetilde{G}_{\mathrm{SM}}(F, y)$, and $z_F$ in $G_{\mathrm{SM}}(F, y)$ plays the role of $\mathbb{E}_F[W(X)]$ in $\widetilde{G}_{\mathrm{SM}}(F, y)$.

Even though Theorem 5.2 provides explicit expressions of $\mathrm{IF}_x(T_\lambda, F, y)$ and $\mathrm{IF}_x(S_\rho, F, y)$, they are hard to work with directly to compare the sensitivities of the penalized ML and SM density estimators. We will work with the sample influence function in the next chapter and compare their sensitivities numerically.

## 5.5 Proofs

### 5.5.1 Proof of Proposition 5.1

*Proof of Proposition 5.1.* Under the assumption in the proposition and with an application of the chain rule, we have

$$\begin{aligned}
\mathrm{IF}_x(\widetilde{T}, F, y) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x)\Big)\Big|_{\varepsilon=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(\exp\big(T((1-\varepsilon)F + \varepsilon\delta_y)(x)\big)\Big)\Big|_{\varepsilon=0} \\
&= \exp\big(T(F)(x)\big)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x)\Big)\Big|_{\varepsilon=0} \\
&= \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y).
\end{aligned}$$

■

### 5.5.2 Proof of Proposition 5.2

*Proof of Proposition 5.2.* Let $\varepsilon \in (0, 1)$. By the definition of the KL-divergence, we have

$$\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big) = -\int_{\mathcal{X}} \widetilde{T}(F)(x)\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big)\mathrm{d}x.$$

Differentiating both sides wrt $\varepsilon$, interchanging differentiation and integral (which is allowed by the assumptions), and evaluating at $\varepsilon = 0$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{KL}\Big(\widetilde{T}(F)\Big\|\widetilde{T}\big((1-\varepsilon)F + \varepsilon\delta_y\big)\Big)\Big|_{\varepsilon=0} \\
&= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(T\big((1-\varepsilon)F + \varepsilon\delta_y\big)(x) - T(F)(x)\Big)\Big|_{\varepsilon=0}\mathrm{d}x \\
&= -\int_{\mathcal{X}} \widetilde{T}(F)(x)\mathrm{IF}_x(T, F, y)\mathrm{d}x \\
&= -\int_{\mathcal{X}} \mathrm{IF}_x(\widetilde{T}, F, y)\mathrm{d}x,
\end{aligned}$$

where the last equality follows from Proposition 5.1. ■

### 5.5.3 Proof of Theorem 5.1

*Proof of Theorem 5.1.* For simplicity, we let $F_\varepsilon := (1-\varepsilon)F + \varepsilon\delta_y$ for $\varepsilon \in (0,1)$.

(a) First note

$$\frac{1}{\varepsilon}\Big(T(F_\varepsilon)(x) - T(F)(x)\Big) = \left\langle \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(x)\right\rangle - \frac{1}{\varepsilon}\Big(B(\theta_{\mathrm{ML},F_\varepsilon}) - B(\theta_{\mathrm{ML},F})\Big).$$

If we let $\varepsilon \to 0^+$ on both sides and use the chain rule, we have

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\Big(T(F_\varepsilon)(x) - T(F)(x)\Big) \\
&= \left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(x)\right\rangle - \int_{\mathcal{X}} \tilde{q}_{\theta_{\mathrm{ML},F}}(w)\left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big), \varphi(w)\right\rangle \mathrm{d}w,
\end{aligned}$$

which exists if the limit $\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F})$ exists. The desired result follows if we can show

$$\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{ML},F_\varepsilon} - \theta_{\mathrm{ML},F}\big) = \big[\nabla^2 B(\theta_{\mathrm{ML},F})\big]^{-1}\big(\varphi(y) - \mathbb{E}_F[\varphi(X)]\big). \tag{5.17}$$

Note the LHS of (5.17) is the influence function of the $M$-estimator defined by

$$0 = \int_{\mathcal{X}} \Big(\nabla B(\theta_{\mathrm{ML},F}) - \varphi(x)\Big)\mathrm{d}F(x).$$

Using the result in Example 5.3, we can see (5.17) follows, which completes the proof.

(b) Similar to (a), we have

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\Big(S(F_\varepsilon)(x) - S(F)(x)\Big) \\
&= \left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big), \varphi(x)\right\rangle - \int_{\mathcal{X}} \tilde{q}_{\theta_{\mathrm{SM},F}}(w)\left\langle \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big), \varphi(w)\right\rangle \mathrm{d}w,
\end{aligned}$$

which exists if the limit $\lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F})$ exists. The desired result follows if we can show

$$\lim_{\varepsilon\to 0^+} \big(\theta_{\mathrm{SM},F_\varepsilon} - \theta_{\mathrm{SM},F}\big) = \Big[\mathbb{E}_F\big[D_1(X)D_1(X)^\top\big]\Big]^{-1} \times$$

$$\left\{W(y) - D_1(y)D_1(y)^\top \left[\mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\right]^{-1} \mathbb{E}_F[W(X)]\right\}. \tag{5.18}$$

Note the LHS of (5.18) is the influence function of the $M$-estimator defined by

$$0 = \mathbb{E}_F\left[D_1(X)D_1(X)^\top\right]\theta_{\mathrm{SM},F} - \mathbb{E}_F[W(X)].$$

Using the result in Example 5.3, we can see (5.18) follows, which completes the proof.

■

### 5.5.4 Proof of Theorem 5.2

*Proof of Theorem 5.2.* We let $F_\varepsilon := (1-\varepsilon)F + \varepsilon\delta_y$ for some $\varepsilon \in (0,1)$ throughout this proof.

(a) Using the chain rule, we obtain

$$\mathrm{IF}_x(T_\lambda, F, y) = \left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}, k(x,\,\cdot\,) - \int_{\mathcal{X}} k(w,\,\cdot\,) q_{f^{(\lambda)}_{\mathrm{ML},F}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}.$$

What remains to show is $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\big|_{\varepsilon=0} = G_{\mathrm{ML}}(F,y)$.

Observe that $f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}$ is an $M$-estimator and must satisfy

$$0 = \nabla A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}) - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F_\varepsilon(x) + \lambda f^{(\lambda)}_{\mathrm{ML},F_\varepsilon},$$

which is equivalent to the following

$$0 = \int_{\mathcal{X}} \mu(x)\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)k(x,\,\cdot\,)\mathrm{d}x - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F_\varepsilon(x) + \lambda f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}.$$

Differentiating both sides of the preceding equation with respect to $\varepsilon$ yields

$$\begin{aligned} 0 = \int_{\mathcal{X}} \mu(x)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]k(x,\,\cdot\,)\mathrm{d}x \\ - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}. \end{aligned} \tag{5.19}$$

We next work out $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]$ part. By the chain rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left[\exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\right]$$

$$= \exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big[f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\Big]$$
$$= \exp\big(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}(x) - A(f^{(\lambda)}_{\mathrm{ML},F_\varepsilon})\big)\left[\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, k(x,\,\cdot\,) - \int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right\rangle_{\mathcal{H}}\right].$$

Plugging the preceding equation back to (5.19), we have

$$\begin{aligned}
0 = &\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, k(x,\,\cdot\,)\right\rangle_{\mathcal{H}} k(x,\,\cdot\,)\mathrm{d}x \\
&\quad - \left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}, \int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right\rangle_{\mathcal{H}} \\
&\qquad - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon} \\
= &\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\right) \\
&\quad - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right]\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\right) \\
&\qquad - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}(-F(x) + \delta_y(x)) + \lambda\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon} \\
= &\left\{\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\right. \\
&\quad \left. - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] + \lambda I\right\}\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}\right) \\
&\qquad + \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F(x) - k(y,\,\cdot\,).
\end{aligned}$$

Evaluating at $\varepsilon = 0$ and rearranging terms, we obtain

$$\begin{aligned}
&\left\{\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]\right. \\
&\quad \left. - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)k(x,\,\cdot\,)\mathrm{d}x\right)\otimes\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] + \lambda I\right\}\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\Big|_{\varepsilon=0}\right) \\
&\qquad = k(y,\,\cdot\,) - \int_{\mathcal{X}} k(x,\,\cdot\,)\mathrm{d}F(x).
\end{aligned}$$

Note that

$$\left[\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)\big(k(x,\,\cdot\,)\otimes k(x,\,\cdot\,)\big)\mathrm{d}x\right]$$

$$
\begin{aligned}
&\quad - \left[\left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(x)k(x,\,\cdot\,)\mathrm{d}x\right) \otimes \left(\int_{\mathcal{X}} q_{f^{(\lambda)}_{\mathrm{ML},F}}(v)k(v,\,\cdot\,)\mathrm{d}v\right)\right] \\
&= \mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[k(X,\,\cdot\,)\otimes k(X,\,\cdot\,)\big] - \Upsilon\otimes\Upsilon \\
&= \mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[(k(X,\,\cdot\,)-\Upsilon)\otimes(k(X,\,\cdot\,)-\Upsilon)\big],
\end{aligned}
$$

which is the covariance operator we have discussed in Chapter 2, and is positive semi-definite. Then, with $\lambda > 0$, the operator

$$
\mathbb{E}_{q_{f^{(\lambda)}_{\mathrm{ML},F}}}\big[(k(X,\,\cdot\,)-\Upsilon)\otimes(k(X,\,\cdot\,)-\Upsilon)\big] + \lambda I
$$

is invertible, and $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\lambda)}_{\mathrm{ML},F_\varepsilon}\big|_{\varepsilon=0}$ is equal to $G_{\mathrm{ML}}(F,y)$ given in the theorem, which completes the proof.

(b) Similar to (a), using the chain rule, we obtain

$$
\mathrm{IF}_x(S_\rho,F,y) = \left\langle \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}, k(x,\,\cdot\,) - \int_{\mathcal{X}} k(w,\,\cdot\,) q_{f^{(\rho)}_{\mathrm{SM},F}}(w)\mathrm{d}w \right\rangle_{\mathcal{H}}.
$$

It remains to show $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\big|_{\varepsilon=0}$ is equal to $G_{\mathrm{SM}}(F,y)$.

Then, observe that $f^{\rho}_{\mathrm{SM},F_\varepsilon}$ is an $M$-estimator and must satisfy

$$
0 = C_{F_\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon} - z_{F_\varepsilon} + \rho f^{(\rho)}_{\mathrm{SM},F_\varepsilon},
$$

which is equivalent to

$$
0 = \big((1-\varepsilon)C_F + \varepsilon C_{\delta_y}\big) f^{(\rho)}_{\mathrm{SM},F_\varepsilon} - \big((1-\varepsilon)z_F + \varepsilon z_{\delta_y}\big) + \rho f^{(\rho)}_{\mathrm{SM},F_\varepsilon}.
$$

Now, differentiating both sides of the preceding equation with respect to $\varepsilon$ yields

$$
0 = \big(C_{\delta_y} - C_F\big) f^{(\rho)}_{\mathrm{SM},F_\varepsilon} + \big((1-\varepsilon)C_F + \varepsilon C_{\delta_y}\big)\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\right) - \big(z_{\delta_y} - z_F\big) + \rho\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\right).
$$

Letting $\varepsilon \to 0^+$ on both sides, we obtain

$$
0 = \big(C_{\delta_y} - C_F\big) f^{(\rho)}_{\mathrm{SM},F} + C_F\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}\right) - \big(z_{\delta_y} - z_F\big) + \rho\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f^{(\rho)}_{\mathrm{SM},F_\varepsilon}\bigg|_{\varepsilon=0}\right).
$$

In addition, since $f_{\mathrm{SM},F}^{(\rho)}$ satisfies $C_F f_{\mathrm{SM},F}^{(\rho)} - z_F + \rho f_{\mathrm{SM},F}^{(\rho)} = 0$, we can use it to simplify the preceding equation as

$$0 = C_{\delta_y} f_{\mathrm{SM},F}^{(\rho)} + C_F \left( \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f_{\mathrm{SM},F_\varepsilon}^{(\rho)} \bigg|_{\varepsilon=0} \right) - z_{\delta_y} + \rho \left( \frac{\mathrm{d}}{\mathrm{d}\varepsilon} f_{\mathrm{SM},F_\varepsilon}^{(\rho)} \bigg|_{\varepsilon=0} \right) + \rho f_{\mathrm{SM},F}^{(\rho)}. \tag{5.20}$$

Rearranging terms and using $f_{\mathrm{SM},F}^{(\rho)} = (C_F + \rho I)^{-1} z_F$ yield the desired result. ■