Chapter 6: Numerical Studies of the Sensitivities of Penalized ML and SM (Log-)Density Estimators in $\mathcal{Q}_{\text{ker}}$

As we have seen from Theorem 5.2 in Chapter 5 that, even though the expressions of the influence functions of the penalized ML and SM (log-)density estimators in $\mathcal{Q}_{\text{ker}}$ exist, they are hard to be directly used to compare the sensitivities of these (log-)density estimators. Instead, we are going to compare their sensitivities numerically, which will be the focus of Section 6.1. Since we have seen the penalized and early stopping SM density estimators are qualitatively very similar through numerical examples in Chapters 3 and 4, we only consider the penalized SM density estimator in this chapter. Since one of the most popular methods of selecting the penalty parameter in the penalized SM density estimation approach is the K-fold cross-validation, we turn to studying the sensitivity of the cross-validated penalized SM density estimator in Section 6.2. Since we have both penalized ML and SM density estimators, we discuss in Section 6.3 which density estimator we should use and how to use it.

6.1 Comparison of the Sensitivities of Penalized ML and SM Density Estimators

We study the sensitivities of the penalized ML and SM density estimators numerically in this section.

The tool we use is the sample influence function defined in Chapter 5. We discuss how to compute the sample influence function of a log-density estimator and that of a density estimator in Section 6.1.1. Since we can use either the sample influence function of the log-density estimator or that of the density estimator, we compare them in Section 6.1.2 and show that the former one is a better choice for us. The main comparison of the sensitivities of penalized ML and SM density estimators will be given in Section 6.1.3.

In order to achieve the desired goal, we still define the following maps as we have done in Section 5.4 in Chapter 5

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

where

$$f^{(\lambda)}_{\mathrm{ML},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ A(f) - \int_{\mathcal{X}} f(x)\mathrm{d}F(x) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} \right\}, \tag{6.1}$$
$$f^{(\rho)}_{\mathrm{SM},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ \frac{1}{2}\langle f, C_F f\rangle_{\mathcal{H}} - \langle f, z_F\rangle_{\mathcal{H}} + \frac{\rho}{2}\|f\|^2_{\mathcal{H}} \right\}, \tag{6.2}$$

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 $z_F := -\int_{\mathcal{X}} \sum_{u=1}^{d} (\partial_u^2 k(x, \cdot) + \partial_u \log \mu(x) \partial_u k(x, \cdot))\mathrm{d}F(x) \in \mathcal{H}$.

In order to ensure the comparability of the penalized ML and SM density estimates, we minimize the objective functionals in (6.1) and (6.2) over the same finite-dimensional approximating subspace of $\mathcal{H}$ that is found by the procedures described in Chapter 4, and denote this finalized subspace by $\widetilde{\mathcal{H}}$.

6.1.1 Computation of the Sample Influence Function

We describe how to compute the sample influence function in practice.

We choose ε in the definition of the sample influence function to be $\varepsilon_0 := \frac{1}{n+1}$. Then, $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$ and $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$ assess the sensitivity of the penalized ML and SM log-density estimators evaluated at x to the additional observation y, respectively; and, similarly, $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F_n, y)$ and $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F_n, y)$ assess the sensitivity of the penalized ML and SM density estimators evaluated at x to the additional observation y, respectively.

Algorithm 6.1 describes how to compute $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$; if one would like to compute $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F_n, y)$, simply do not take the logarithm in Steps 2 and 3.

Algorithm 6.1 Computation of $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F_n, y)$

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- $\lambda > 0$, penalty parameter;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized NLL loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of evaluation points.

1: Compute $f^{(\lambda)}_{\mathrm{ML},F_n}$ and $f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}$ using Algorithm 4.3 in Chapter 4;
2: Compute $\log q_{f^{(\lambda)}_{\mathrm{ML},F_n}}(x_\ell)$ and $\log q_{f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell)$ for all $\ell = 1, \cdots, L$;
3: Compute

$$\left(\log q_{f^{(\lambda)}_{\mathrm{ML},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell) - \log q_{f^{(\lambda)}_{\mathrm{ML},F_n}}(x_\ell)\right) \times (n+1), \qquad \text{for all } \ell = 1, \cdots, L;$$

4: **return** the results from Step 3.

Similarly, Algorithm 6.2 describes how to compute $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$; if one would like to compute $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F_n, y)$, do not take the logarithm in Steps 2 and 3.

Algorithm 6.2 Computation of $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y)$

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- $\rho > 0$, penalty parameter;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized SM loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of evaluation points.

1: Compute $f^{(\rho)}_{\mathrm{SM},F_n}$ and $f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}$ using Proposition 4.3 in Chapter 4;
2: Compute $\log q_{f^{(\rho)}_{\mathrm{SM},F_n}}(x_\ell)$ and $\log q_{f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell)$ for all $\ell = 1, \cdots, L$;
3: Compute

$$\left(\log q_{f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}}(x_\ell) - \log q_{f^{(\rho)}_{\mathrm{SM},F_n}}(x_\ell)\right) \times (n+1), \qquad \text{for all } \ell = 1, \cdots, L;$$

4: **return** the results from Step 3.

6.1.2 Comparison of the Sample Influence Functions of Log-density and Density Estimators

Our goal of the current section is to show, between the sample influence function of the log-density estimator and that of the density estimator in $\mathcal{Q}_{\mathrm{ker}}$, the former is the better choice for us.

We first use numerical examples to demonstrate this. We still use the `waiting` variable in the Old Faithful Geyser dataset but remove the original isolated observation 108 therein. This is to avoid the interaction between 108 and the additional observation y. We use the same $\mathcal{X}$, k, μ, and $\widetilde{\mathcal{H}}$ as those in Chapter 4.

We focus on the penalized SM density estimator for now. Fix $\rho = e^{-11}$. Figure 6.1 shows the penalized SM (log-)density estimates with and without $y = 120$ and the corresponding sample influence functions, and Figure 6.2 shows the penalized SM (log-)density estimates with and without $y = 180$ and the corresponding sample

influence functions. Recall that we choose μ to be the pdf of the Gamma distribution with the shape and scale parameters to be 36 and 2, respectively, and $\mu(x) \to 0$ and $\log \mu(x) \to -\infty$ as $x \to \infty$. Then, comparing Panel [D] in Figure 6.1 and that in Figure 6.2, we see that, when y is large, the spike in the penalized SM density estimate may disappear due to this particular choice of μ. The sample influence function of the penalized SM density estimator may fail to capture the bump or the spike in the penalized SM density estimate when an isolated observation is present (see Panel [F] in Figure 6.2). In other words, understanding the sensitivity of the penalized SM density estimator via the sample influence function of the density estimator can be misleading.

However, the issue with the sample influence function of the density estimator described above does *not* occur to the sample influence function of the log-density estimator. Comparing Panel [C] in Figure 6.1 and that in Figure 6.2, we see the sample influence function of the penalized SM log-density estimator succeeds in capturing the spike at the isolated observation, no matter how far y is from the bulk of the data. Numerical examples of the penalized ML density estimator also confirm this; see Figures 6.3 and Figure 6.4 for evidence.

Analytically, note that

$$
\begin{aligned}
& S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)(x) - S_\rho(F_n)(x) \\
&= \big[\cancel{\log \mu(x)} + f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big] \\
&\qquad - \big[\cancel{\log \mu(x)} + f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big] \\
&= \big[f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big] - \big[f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big],
\end{aligned}
$$

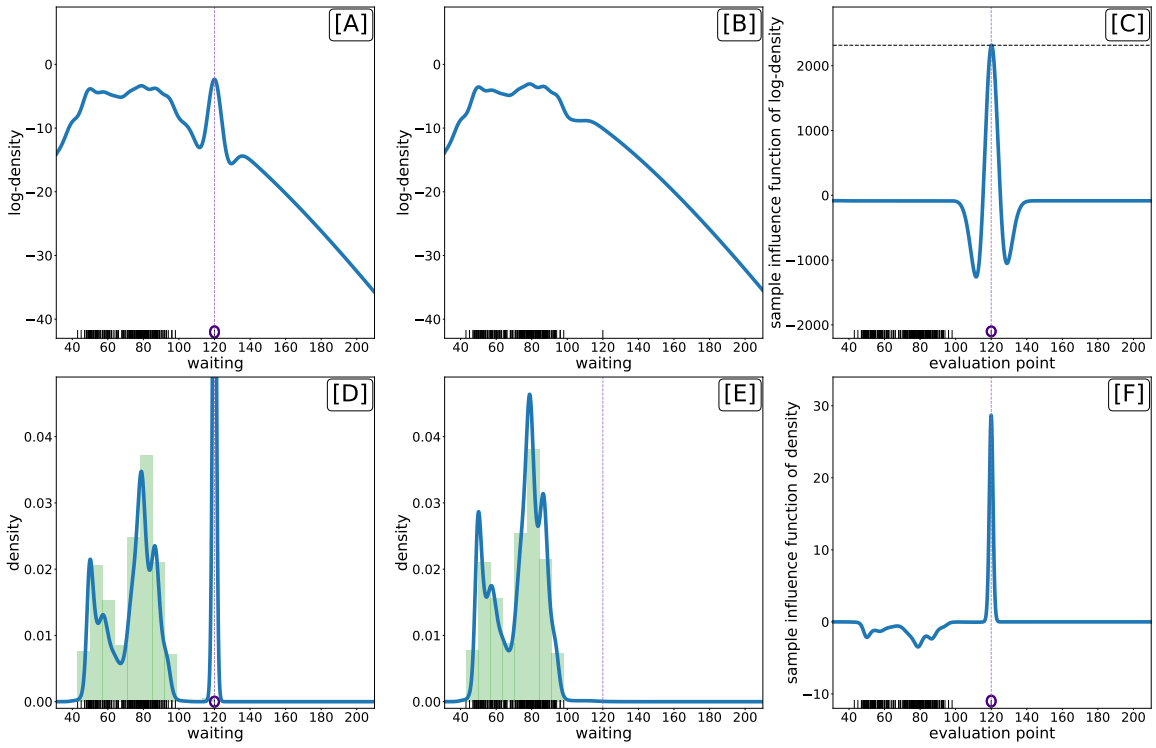

Figure 6.1: Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$.

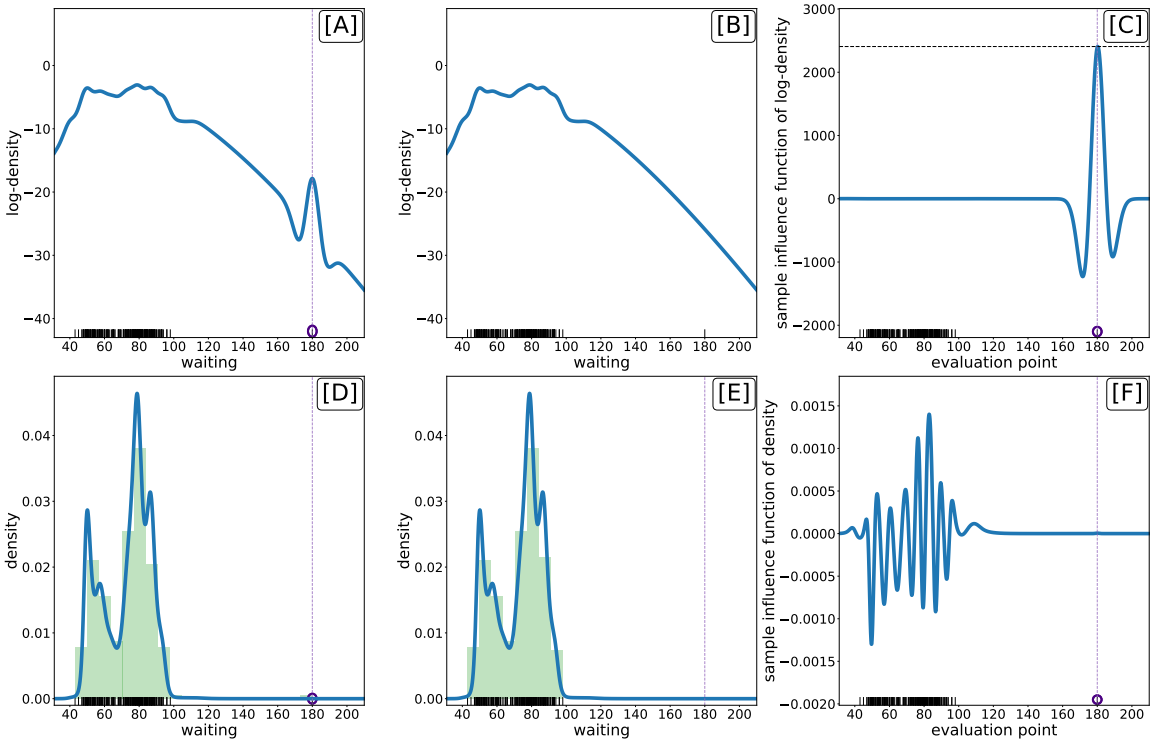

Figure 6.2: Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$.

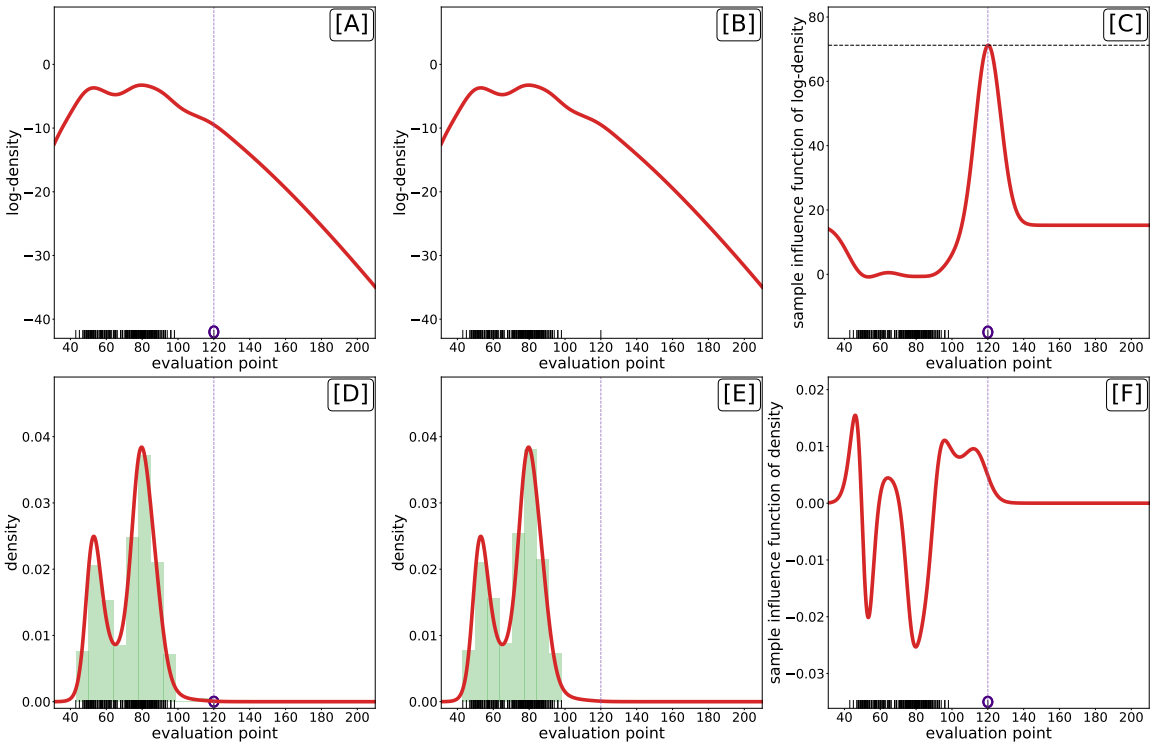

Figure 6.3: Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$.

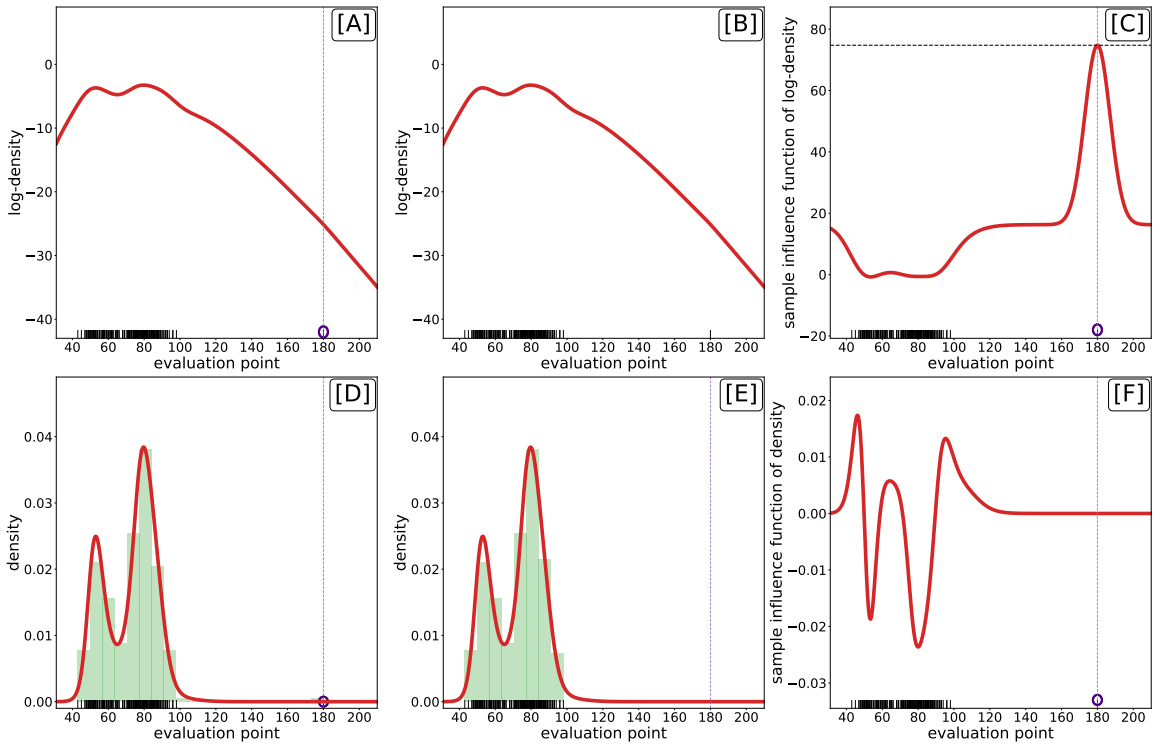

Figure 6.4: Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$.

from which we see $\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F, y)$ does *not* depend on $\mu(x)$, but only on the natural parameter part. If we work with $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F, y)$, however, we have

$$\begin{aligned}&\widetilde{S}_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)(x) - \widetilde{S}_\rho(F_n)(x)\\ =&\mu(x)\Big[\exp\big(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y}(x) - A(f^{(\rho)}_{\mathrm{SM},(1-\varepsilon_0)F_n+\varepsilon_0\delta_y})\big) - \exp\big(f^{(\rho)}_{\mathrm{SM},F_n}(x) - A(f^{(\rho)}_{\mathrm{SM},F_n})\big)\Big],\end{aligned}$$

and we cannot get rid of $\mu(x)$ in the front. The resulting $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{S}_\rho, F, y)$ inevitably depends on $\mu(x)$. By a similar approach as above, we can also see $\mathrm{SIF}_{x,\varepsilon_0}(T_\lambda, F, y)$ does *not* depend on $\mu(x)$ but $\mathrm{SIF}_{x,\varepsilon_0}(\widetilde{T}_\lambda, F, y)$ does.

Thus, from both numerical examples and analytic analysis, we see the sample influence function of the *log*-density estimator is the better choice than that of the density estimator. Hence, we will only consider the former in the sequel.

6.1.3 Comparison of the Sensitivities

Our goal of this section is to compare the sensitivities of the penalized ML and SM density estimators in $\mathcal{Q}_{\mathrm{ker}}$.

Let us still fix $\rho = e^{-11}$ and $y = 120$, and return to the first row of Figure 6.1 where we show $S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)$, $S_\rho(F_n)$, and the resulting sample influence function evaluated at different points. It is apparent that $y = 120$ has different effects on S_ρ at different evaluation points. The overall influence of y on S_ρ, $\widehat{M}_{\varepsilon_0}(S_\rho, F_n, y)$, is approximately equal to 2315.48, which is achieved roughly at 120.

Let us still fix $\rho = e^{-11}$ but vary y. The left panel in Figure 6.5 shows the overall influence of y on S_ρ against different choices of y. It is obvious that different locations of y have different overall influences on S_ρ. When y is below 40 or above 100 (low-density region), it has a larger overall influence on S_ρ; and when y is between 40 and 100 (high-density region), it has a smaller overall influence.

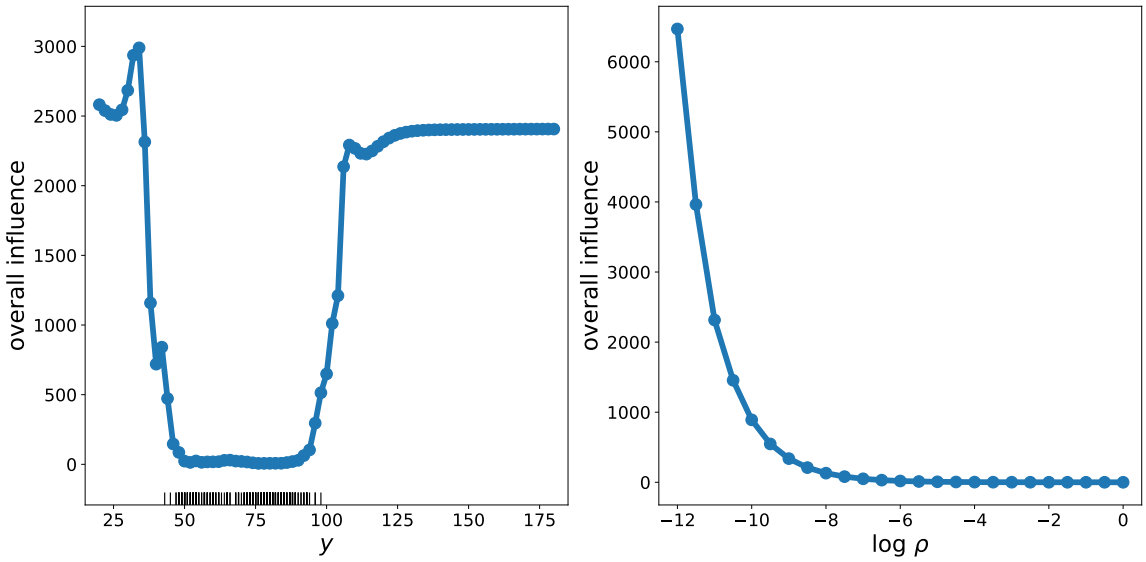

Figure 6.5: Left panel shows the overall influence versus different choices of y, where we fix $\rho = e^{-11}$ and the rugs indicate the location of the `waiting` data. Right panel shows the overall influence against different choices of ρ (shown in log scale), where we fix $y = 120$.

Now, fix $y = 120$ but vary the values of ρ. The right panel in Figure 6.5 shows the overall influence of $y = 120$ on S_ρ versus different choices of ρ. It is obvious that the overall influence of y keeps increasing as the value of ρ keeps decreasing.

Thus, both the location of y and the value of ρ impact the overall influence on log-density estimators. To exhibit the sensitivity of the penalized SM log-density estimators subject to these two factors, a natural idea is to plot the heat map of the overall influence against them, which is shown in Figure 6.6. However, recall our ultimate goal is to compare the sensitivities of both penalized ML and SM density estimators, and their respective penalty parameters, λ and ρ, are on different scales. Thus, plotting the penalty parameter on the horizontal axis is not conducive to comparison. Instead, we plot the RKHS norm of natural parameter under F_n, i.e.,

$\|f^{(\lambda)}_{\mathrm{ML},F_n}\|_{\mathcal{H}}$ and $\|f^{(\rho)}_{\mathrm{SM},F_n}\|_{\mathcal{H}}$, on the horizontal axis. The smaller the penalty parameter is, the larger the RKHS norm of the natural parameter is.

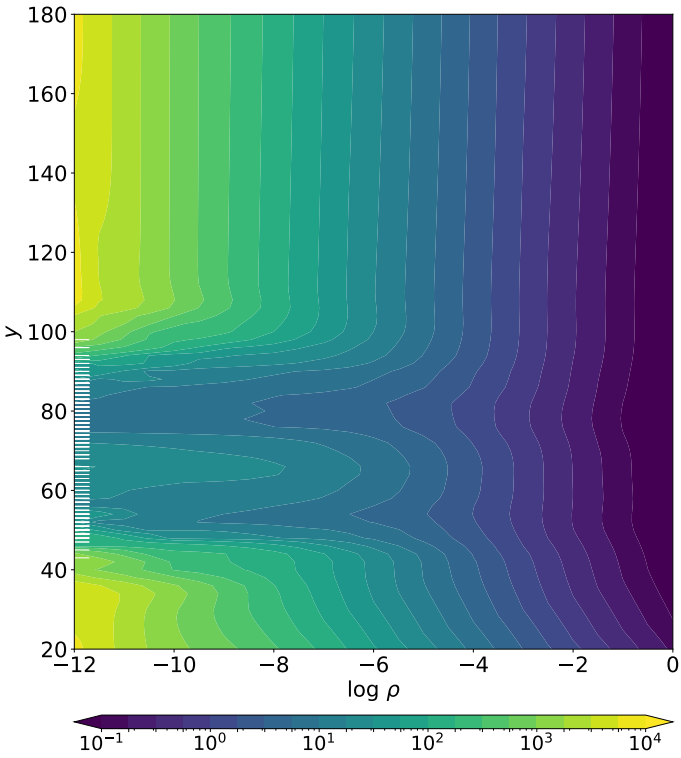

Figure 6.6: Heat map of the overall influence on the penalized SM log-density estimates against y and ρ (shown in log scale). White rugs indicate locations of the `waiting` data.

The resulting heat maps for the penalized ML and SM density estimates are shown in Figure 6.7, where we choose $y = 20, 22, \cdots, 180$, $\lambda = 0, e^{-15}, e^{-14.5}, \cdots, e^{0.5}, e^{1}$, and $\rho = e^{-12}, e^{-11.5}, \cdots, e^{0}$. If we look at each panel individually, findings are consistent as before: with a fixed value of the penalty parameter, y in the low-density region has a larger overall influence on log-density estimates than that in the high-density region; with a fixed y, a larger RKHS norm of the natural parameter (corresponding to a smaller penalty parameter value) implies a larger overall influence of y. In particular,

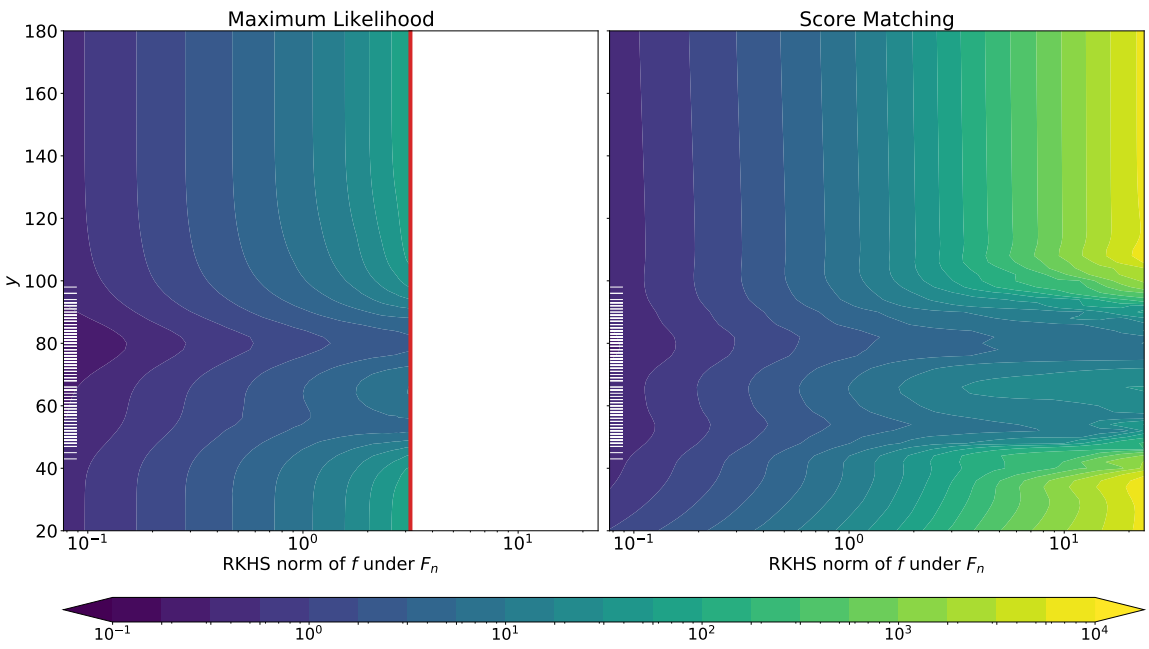

Figure 6.7: Heat maps of the overall influence on penalized ML (left) and SM (right) log-density estimates against y and the RKHS norm of the natural parameter under F_n (shown in log scale). Red vertical line in left panel indicates the case $\lambda = 0$. White rugs indicate locations of `waiting` data.

if we fix a y value, say y_0, the maximal possible overall influence of y_0 on the penalized ML density estimates for all values of $\lambda \geq 0$ is the intersection of $y = y_0$ and the red vertical line, which corresponds to the overall influence of y_0 on the unpenalized ML density estimate (i.e., $\lambda = 0$).

If we compare two panels in Figure 6.7, we observe that, for each choice of y, as we decrease the values of penalty parameters so that the RKHS norm of f under F_n keeps increasing, the overall influences on the penalized ML density estimates stop increasing at the red vertical line, but that on the penalized SM density estimates can continue increasing. In other words, for each choice of y, the overall influences on the penalized SM log-density estimates with sufficiently small values of $\rho > 0$ are larger than the overall influence on the *unpenalized* ML log-density estimate, implying that,

when there is a small amount of penalization, the penalized SM density estimator is more sensitive to an additional observation (not only to the isolated observation) than the unpenalized and penalized ML density estimator. Additional numerical examples also confirm our observations here.

6.2 The Sensitivity of K-fold Cross-validated Penalized SM Density Estimator

K-fold cross-validation (CV) is perhaps the most popular method to choose the penalty parameter in practice. In this section, we are going to investigate whether the K-fold cross-validated penalized SM density estimator is sensitive to the presence of an additional observation or not. The procedure of computing the overall influence of y on the K-fold cross-validated penalized SM density estimator is shown in Algorithm 6.3.

We still use the `waiting` variable in the Old Faithful Geyser dataset and choose the number of folds to be $K = 3, 5, 10$. For each value of K, we replicate the procedures outlined in Algorithm 6.3 for 30 times. The additional observations y we choose are the same as those in the preceding section, $y = 20, 22, \cdots, 180$. Figure 6.8 shows the results.

For each number of folds, similar to our earlier observations, when y is in the high-density region (between 40 and 100), the overall influences of y on cross-validated penalized SM density estimates tend to be small; and when y is in the low-density region (< 40 or > 100), the overall influences of y tend to be much larger.

In addition, we see as the number of folds increases, the cross-validated penalized SM log-density estimates become less sensitive in general. This suggests that when one uses the K-fold CV to select the penalty parameter, it is better to use a relatively

Algorithm 6.3 Computation of the overall influence of K-fold cross-validated penalized SM density estimator

Require:

- $X_1, \cdots, X_n$, data;
- $y \in \mathcal{X}$, contaminant;
- K, the number of folds in CV;
- $\{\rho_j\}_{j=1}^M$, a list of penalty parameter candidates;
- $\widetilde{\mathcal{H}}$, the finite-dimensional approximating subspace over which we minimize the penalized NLL loss functional;
- $\{x_\ell\}_{\ell=1}^L \subset \mathcal{X}$, a set of dense evaluation points in $\mathcal{X}$.

1: Compute the best density estimate using $X_1, \cdots, X_n$ with the penalty parameter selected by CV; denote the resulting log-density estimate by $S_{\mathrm{CV}}(F_n)$;
2: Compute the best density estimate using $X_1, \cdots, X_n$ and y with the penalty parameter selected by CV; denoted the resulting log-density estimate by $S_{\mathrm{CV}}(F_{n+1})$;
3: Compute

$$\mathrm{SIF}_{x_\ell, \varepsilon_0}(S_{\mathrm{CV}}, F_n, y) := \Big(S_{\mathrm{CV}}(F_{n+1})(x_\ell) - S_{\mathrm{CV}}(F_n)(x_\ell) \Big) \times (n+1),$$

for all $\ell = 1, \cdots, L$;
4: **return** $\max_{\ell=1,\cdots,L} \big| \mathrm{SIF}_{x_\ell, \varepsilon_0}(S_{\mathrm{CV}}, F_n, y) \big|$.

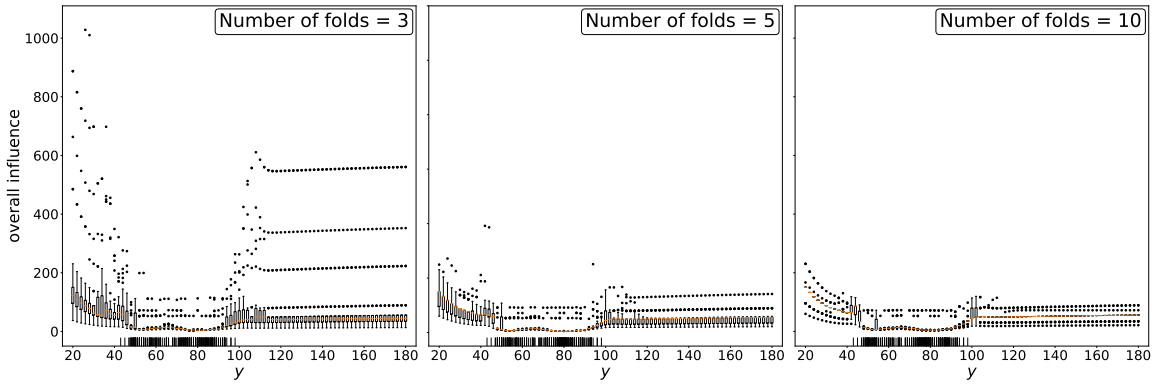

Figure 6.8: Overall influence of y on the K-fold cross-validated penalized SM density estimates against the values of y. We choose $K = 3$ (left panel), 5 (middle panel), and 10 (right panel).

large number of folds, e.g., $K = 5$ or 10, in order to obtain a density estimate that is *not* sensitive to the presence of an isolated observation.

6.3 Which One to Use: Penalized ML or Regularized SM Density Estimators?

Now, we can compute penalized ML and regularized SM density estimators within $\mathcal{Q}_{\mathrm{ker}}$. A natural question arises: which density estimator should one use? Regularized SM density estimators are easy to compute, and can be obtained by solving a linear system or adding and multiplying certain matrices, but are sensitive to the presence of isolated observations, especially when the amount of regularization is small. Penalized ML density estimator, on the other hand, is hard to compute as one has to handle A and its derivative, but is *not* very sensitive to the presence of isolated observations. We recommend using regularized SM density estimators, primarily due to their computational advantage. When using them, one needs to impose an appropriate amount of regularization to ensure the resulting density estimate is not just a spike at the isolated observation; otherwise, such a density estimate can be of no use

and be misleading. One can use the K-fold cross-validation to select the appropriate amount of regularization as we did in this chapter and the earlier ones.