Appendix of "γ-ABC: Outlier-Robust Approximate Bayesian Computation Based on a Robust Divergence Estimator"

Contents

A Derivation of γ-divergence Estimator

We show how to derive the k-NN based γ-divergence estimator in (7). The k-NN based γ-divergence estimator and its derivation is as follows.

$$\widehat{D}_\gamma(X^n\|Y^m) = \frac{1}{\gamma(1+\gamma)} \times \left(\log \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{c}}{k}\hat{p}_k(x_i))^\gamma\right)\left(\frac{1}{m}\sum_{j=1}^{m}(\frac{\bar{c}}{k}\hat{q}_k(y_j))^\gamma\right)^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{c}}{k}\hat{q}_k(x_i))^\gamma\right)^{1+\gamma}} \right),$$

where $\gamma(\in \mathbb{R}) > 0$.

We rewrite Eq. (3) as

$$\begin{aligned}
&\frac{1}{\gamma(1+\gamma)}\log\int_{\mathcal{M}} p(x)p^\gamma(x)\mathrm{d}x - \frac{1}{\gamma}\log\int_{\mathcal{M}} p(x)q^\gamma(x)\mathrm{d}x \frac{1}{1+\gamma}\log\int_{\mathcal{M}'} q(y)q^\gamma(y)\mathrm{d}y \\
&= \frac{1}{\gamma(1+\gamma)}\left(\log\mathbb{E}_{p(x)}\left[p^\gamma(x)\right] - (1+\gamma)\log\mathbb{E}_{p(x)}\left[q^\gamma(x)\right] + \gamma\log\mathbb{E}_{q(y)}\left[q^\gamma(y)\right]\right), \qquad (10)
\end{aligned}$$

where $\mathcal{M}$ and $\mathcal{M}'$ are the supports of p and q. By simply plugging Eqs. (4) and (5) into Eq. (10), we estimate $D_\gamma(p\|q)$ with

$$\begin{aligned}
&\widehat{D}_\gamma(X^n\|Y^m) \\
&= \frac{1}{\gamma(1+\gamma)}\left[\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{(n-1)\bar{c}\rho_k^d(i)}\right)^\gamma\right) - (1+\gamma)\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{m\bar{c}\nu_k^d(i)}\right)^\gamma\right) + \gamma\log\left(\frac{1}{m}\sum_{j=1}^{m}\left(\frac{k}{(m-1)\bar{c}\bar{\rho}_k^d(j)}\right)^\gamma\right)\right] \\
&= \frac{1}{\gamma(1+\gamma)}\left(\log\frac{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(n-1)^\gamma\rho_k^{d\gamma}(i)}\right)\left(\frac{1}{m}\sum_{j=1}^{m}\frac{1}{(m-1)^\gamma\bar{\rho}_k^{d\gamma}(j)}\right)^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^\gamma\nu_k^{d\gamma}(i)}\right)^{1+\gamma}}\right) \\
&= \frac{1}{\gamma(1+\gamma)}\left(\log\frac{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{c}^\gamma}{k^\gamma}\cdot\frac{k^\gamma}{(n-1)^\gamma\bar{c}^\gamma\rho_k^{d\gamma}(i)}\right)\left(\frac{1}{m}\sum_{j=1}^{m}\frac{\bar{c}^\gamma}{k^\gamma}\cdot\frac{k^\gamma}{(m-1)^\gamma\bar{c}^\gamma\bar{\rho}_k^{d\gamma}(j)}\right)^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{c}^\gamma}{k^\gamma}\cdot\frac{k^\gamma}{m^\gamma\bar{c}^\gamma\nu_k^{d\gamma}(i)}\right)^{1+\gamma}}\right) \\
&= \frac{1}{\gamma(1+\gamma)}\left(\log\frac{\left(\frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{c}}{k}\hat{p}_k(x_i))^\gamma\right)\left(\frac{1}{m}\sum_{j=1}^{m}(\frac{\bar{c}}{k}\hat{q}_k(y_j))^\gamma\right)^\gamma}{\left(\frac{1}{n}\sum_{i=1}^{n}(\frac{\bar{c}}{k}\hat{q}_k(x_i))^\gamma\right)^{1+\gamma}}\right).
\end{aligned}$$

In second equation, because of logarithm, $k/\bar{c}$ in first term is vanished. The third equation holds because $k^\gamma/\bar{c}^\gamma \cdot \bar{c}^\gamma/k^\gamma = 1$.

Therefore, the definition holds.

B Robust properties on ABC with our method

We investigate the behavior of the *sensitivity curve* (SC), which is an empirical analogue of *influence function* (IF) both of which are used in quantifying the robustness of statistics. The analysis corresponds to a finite-sample analogue of what is called *redescending property* [55] in the context of influence function analysis. Note that we refer to the redescending property in the asymptotic sense, where some authors use the term *redescending* only when there exists a finite threshold $\rho > 0$ such that the influence function ψ satisfies $\forall|x| > \rho, \psi(x) = 0$ [36].

B.1 Notation

Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{R}_{\geq 0}$ denote the set of real numbers, positive integers, and non-negative real numbers, respectively. Let $\mathbb{1}\{\cdot\}$ denote the indicator function. For $m \in \mathbb{N}$, define $[m] \coloneqq \{1, \ldots, m\}$.

We fix $X^n := (X_1, \dots, X_n)$. For $Y^m = (Y_1, \dots, Y_m) \in \mathbb{R}^{m \times d}$, define $\|Y^m\|_{\mathrm{col},\infty} := \max_{j \in [m]} \|Y_j\|$. Let Θ be the parameter space, $\mathrm{d}G_\theta^m(Y^m) := \prod_{j=1}^m p_\theta(Y_j)\mathrm{d}Y_j$, and define $P_\theta(A) := \int \mathbb{1}\{Y^m \in A\}\mathrm{d}G_\theta^m(Y^m)$ for (Borel) measurable set $A \subset \mathbb{R}^{m \times d}$.

Definition 1 (Population pseudo-posterior)**.** *The population pseudo-posterior for $\widehat{D}, \epsilon, \pi$ is defined as*

$$\hat{\pi}(\theta|X^n) := \frac{\pi(\theta)P_\theta(\widehat{D}(X^n \parallel Y^m) < \epsilon)}{\int \pi(\theta')P_{\theta'}(\widehat{D}(X^n \parallel Y^m) < \epsilon)\mathrm{d}\theta'}.$$

For convenience of notation, we define $X^n_{[X_0]}$ as $X^n_{[X_0]} := (X_0, X_1, \dots, X_n)$, i.e., the data X^n combined with the contamination X_0. We consider the behavior of $\hat{\pi}$ under a contamination X_0, i.e., the properties of $\hat{\pi}(\theta|X^n_{[X_0]})$.

Definition 3 (Sensitivity curve [36, 2.1e])**.** *The sensitivity curve of $\hat{\pi}$ is defined as*

$$\mathrm{SC}^\theta_{n+1}(X_0) := (n+1)\left(\hat{\pi}(\theta|X^n_{[X_0]}) - \hat{\pi}(\theta|X^n)\right).$$

B.2 Theorem and Proof

In the following theorem, we will see how SC^θ_{n+1} behaves when the outlier X_0 goes far away from the origin.

Theorem 1 (Sensitivity curve analysis)**.** *Assume $k < \min\{n, m\}$. Also assume that $F_\theta(\epsilon) := P_\theta(\widehat{D}(X^n \parallel Y^m) < \epsilon)$ is β-Lipschitz continuous for all $\theta \in \Theta$. Let $\widehat{D}$ be the γ-divergence estimator in Eq. (7). Then we have*

$$\lim_{\|X_0\| \to \infty} \mathrm{SC}^\theta_{n+1}(X_0) \leq -\frac{\beta\pi(\theta)}{\Lambda_n(1+\gamma)} \log\left(1 - \frac{1}{n^2}\right)^{n+1},$$

where $\Lambda_n := \int \pi(\theta')F_{\theta'}(\epsilon)\mathrm{d}\theta'$. Furthermore, if Λ_n converges to $\Lambda \neq 0$ for $n \to \infty$, then the right-hand side expression converges to 0.

Proof. For simplicity, define $\widehat{D}^{n,m} := \widehat{D}(X^n \parallel Y^m)$ and $\widehat{D}^{n,m}_{[X_0]} := \widehat{D}(X^n_{[X_0]} \parallel Y^m)$. Let us start by considering $\lim_{\|X_0\| \to \infty} \int \mathbb{1}\{\widehat{D}^{n,m}_{[X_0]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m)$. To obtain this limit, observe that we only need to take an arbitrary sequence $\{X'_j\}_{j=1}^\infty$ satisfying $\|X'_j\| \to \infty$ and calculate $\lim_{j \to \infty} \int \mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m)$ (see Remark 3). Fix such a sequence $\{X'_j\}_{i=1}^\infty$.

We first consider the point-wise limit $\lim_{j \to \infty} \mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}$ for each value of Y^m because we later interchange the limit and the integration by applying the bounded convergence theorem [67, 11.32]: $\lim_{j \to \infty} \int \mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m) = \int \lim_{j \to \infty} \mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m)$ using the boundedness of $|\mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}|$ (bounded by 1) and the finiteness of the measure $\mathrm{d}G_\theta^m(Y^m)$.

Fix Y^m. Since $\{X'_j\}_{i=1}^\infty$ is diverging, if j is large enough, X'_j is never within the k-nearest neighbors of any of the points in X^n or Y^m (here, we used the assumption $k < n, m$), hence $\rho_k^d(i)$ and $\nu_k^d(i)$ $(i = 1, \dots, n)$ do not depend on X'_j if j is large enough. Let $A_1 := \sum_{i=1}^n \left(\frac{1}{\rho_k^d(i)}\right)^\gamma$ and $A_2 := \sum_{i=1}^n \left(\frac{1}{\nu_k^d(i)}\right)^\gamma$, and by abuse of notation, substitute $X_0 := X'_j$ so as to enable using the convenient notation $\rho_k(0)$ and $\nu_k(0)$. We can rewrite the event $\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}$ in terms of $\widehat{D}^{n,m}$ based on

the following calculation:

$$
\begin{aligned}
&\gamma(1+\gamma)\left(\widehat{D}^{n,m}_{[X'_j]} - \widehat{D}^{n,m}\right) \\
&= \left\{\log\left(\frac{k}{((n+1)-1)\bar{c}}\right)^{\gamma}\frac{1}{n+1}\sum_{i=0}^{n}\left(\frac{1}{\rho_k^d(i)}\right)^{\gamma} - \log\left(\frac{k}{(n-1)\bar{c}}\right)^{\gamma}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\rho_k^d(i)}\right)^{\gamma}\right\} \\
&\quad - (1+\gamma)\left\{\log\left(\frac{k}{m\bar{c}}^{\gamma}\frac{1}{n+1}\sum_{i=0}^{n}\left(\frac{1}{\nu_k^d(i)}\right)^{\gamma}\right) - \log\left(\frac{k}{m\bar{c}}^{\gamma}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\nu_k^d(i)}\right)^{\gamma}\right)\right\} \\
&= \log\left(\frac{n-1}{n}\right)^{\gamma}\left(\frac{1}{n+1}\rho_k^{-d\gamma}(0) + \frac{1}{n+1}A_1\right)\left(\frac{1}{n}A_1\right)^{-1} - (1+\gamma)\log\left(\frac{1}{n+1}\nu_k^{-d\gamma}(0) + \frac{1}{n+1}A_2\right)\left(\frac{1}{n}A_2\right)^{-1} \\
&= \left\{\gamma\log\frac{n-1}{n} + \log\frac{n}{n+1} - (1+\gamma)\log\frac{n}{n+1}\right\} + \left\{\log\left(A_1^{-1}\rho_k^{-d\gamma}(0)+1\right) - (1+\gamma)\log\left(A_2^{-1}\nu_k^{-d\gamma}(0)+1\right)\right\} \\
&= \gamma\log(1-\frac{1}{n^2}) + \left\{\log\left(A_1^{-1}\rho_k^{-d\gamma}(0)+1\right) - (1+\gamma)\log\left(A_2^{-1}\nu_k^{-d\gamma}(0)+1\right)\right\}.
\end{aligned}
$$

Therefore, $\widehat{D}^{n,m}_{[X'_j]} < \epsilon \Leftrightarrow \widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)$ if j is large enough, where

$$
\tilde{\epsilon} \coloneqq \epsilon - \frac{1}{1+\gamma}\log(1-\frac{1}{n^2}), \qquad \phi(X'_j) \coloneqq \log\left(A_1^{-1}\rho_k^{-d\gamma}(0)+1\right) - (1+\gamma)\log\left(A_2^{-1}\nu_k^{-d\gamma}(0)+1\right)
$$

and $\rho_k(i), \nu_k(i)$ are based on the temporary notation $X_0 = X'_j$. In terms of indicator functions, we have just shown that

$$
\mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\} = \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\} \tag{11}
$$

holds if j is large enough. We have $\lim_{j\to\infty}\phi(X'_j) = 0$ as well.

Now we show that, for each fixed distinct set of points $(Y_2, \ldots, Y_m)$, we have $\lim_{j\to\infty}\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\} = \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\}$ for almost all Y_1. Fix distinct points $Y_2, \ldots, Y_m$. First, we can show that

$$
\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\} \;\le\; \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\} \;\le\; \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\} + \left(\mathbb{1}\{\widehat{D}^{n,m} = \tilde{\epsilon}\} - \mathbb{1}\{\widehat{D}^{n,m} = \tilde{\epsilon} + \phi(X'_j)\}\right) \tag{12}
$$

holds if j is large enough. To see the first inequality, observe the following: if Y_1 is such that $\widehat{D}^{n,m} < \tilde{\epsilon}$, there exists J such that for all $j > J$ it holds that $|\phi(X'_j)| < \tilde{\epsilon} - \widehat{D}^{n,m}$, and hence $\widehat{D}^{n,m} < \tilde{\epsilon} - |\phi(X'_j)| \le \tilde{\epsilon} + \phi(X'_j)$. Therefore, if j is large enough, $\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\} \le \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\}$ as functions of Y_1. The second inequality can be shown by similarly obtaining $\mathbb{1}\{\widehat{D}^{n,m} > \tilde{\epsilon}\} \le \mathbb{1}\{\widehat{D}^{n,m} > \tilde{\epsilon} + \phi(X'_j)\}$ for large enough j and rearranging the terms. By Equation (12), defining $\mathcal{Z} \coloneqq \{Y_1 : \widehat{D}^{n,m} = \tilde{\epsilon}\} \cup \left(\bigcup_j \{Y_1 : \widehat{D}^{n,m} = \tilde{\epsilon} + \phi(X'_j)\}\right)$, we have $\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\} = \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\}$ if j is large enough, for each $Y_1 \notin \mathcal{Z}$. On the other hand, by Proposition 1, each of $(\widehat{D}^{n,m})^{-1}(\{\tilde{\epsilon}\})$ and $(\widehat{D}^{n,m})^{-1}(\{\tilde{\epsilon} + \phi(X'_j)\})$ has zero Lebesgue measure, hence their countable union $\mathcal{Z}$ also has zero Lebesgue measure. As a result,

$$
\lim_{j\to\infty}\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\} = \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\} \quad \text{a.e. } Y_1 \tag{13}
$$

holds for all $(Y_2, \ldots, Y_m)$.

Now, apply the bounded convergence theorem [67, 11.32], the Fubini-Tonelli theorem [9, Theorem 18.3], and Equation (13) to obtain

$$
\begin{aligned}
&\lim_{j\to\infty} P_\theta(\widehat{D}^{n,m}_{[X'_j]} < \epsilon) = \lim_{j\to\infty}\int \mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m) = \int \lim_{j\to\infty}\mathbb{1}\{\widehat{D}^{n,m}_{[X'_j]} < \epsilon\}\mathrm{d}G_\theta^m(Y^m) \\
&= \int \lim_{j\to\infty}\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\}\mathrm{d}G_\theta^m(Y^m) = \int\left(\int \lim_{j\to\infty}\mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon} + \phi(X'_j)\}\mathrm{d}G_\theta(Y_1)\right)\prod_{j=2}^{m}\mathrm{d}G_\theta(Y_j) \\
&= \int\left(\int \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\}\mathrm{d}G_\theta(Y_1)\right)\prod_{j=2}^{m}\mathrm{d}G_\theta(Y_j) = \int \mathbb{1}\{\widehat{D}^{n,m} < \tilde{\epsilon}\}\mathrm{d}G_\theta^m(Y^m) = P_\theta(\widehat{D}^{n,m} < \tilde{\epsilon}),
\end{aligned}
$$

where we also took into account that the points $Y_2, \ldots, Y_m$ are almost surely distinct. Since the choice of $\{X'_j\}_{i=1}^{\infty}$ was arbitrary, the above calculation implies

$$\lim_{\|X_0\|\to\infty} P_\theta(\widehat{D}^{n,m}_{[X_0]} < \epsilon) = P_\theta(\widehat{D}^{n,m} < \tilde{\epsilon}).$$

Therefore, defining $\eta_\theta(\epsilon) \coloneqq \pi(\theta) P_\theta(\widehat{D}^{n,m} < \epsilon)$,

$$\begin{aligned}
\lim_{\|X_0\|\to\infty} \hat{\pi}(\theta|X^n_{[X_0]}) &= \lim_{\|X_0\|\to\infty} \frac{\pi(\theta) P_\theta(\widehat{D}^{n,m}_{[X_0]} < \epsilon)}{\int \pi(\theta') P_{\theta'}(\widehat{D}^{n,m}_{[X_0]} < \epsilon) \mathrm{d}\theta'} \\
&= \left(\lim_{\|X_0\|\to\infty} \pi(\theta) P_\theta(\widehat{D}^{n,m}_{[X_0]} < \epsilon)\right) \left(\lim_{\|X_0\|\to\infty} \int \pi(\theta') P_{\theta'}(\widehat{D}^{n,m}_{[X_0]} < \epsilon) \mathrm{d}\theta'\right)^{-1} \\
&= (\eta_\theta(\tilde{\epsilon})) \left(\int \eta_{\theta'}(\tilde{\epsilon}) \mathrm{d}\theta'\right)^{-1}
\end{aligned}$$

where we applied the bounded convergence theorem [67, 11.32] to the integration in the denominator as $P_\theta \leq 1$. As a result, denoting $\Delta_{\theta,\tilde{\epsilon},\epsilon} \coloneqq \eta_\theta(\tilde{\epsilon}) - \eta_\theta(\epsilon)$ and noting that $\tilde{\epsilon} \geq \epsilon$ hence $\Delta_{\theta,\tilde{\epsilon},\epsilon} \geq 0$,

$$\begin{aligned}
&\lim_{\|X_0\|\to\infty} \mathrm{SC}^\theta_{n+1}(X_0) = (n+1)\left(\lim_{\|X_0\|\to\infty} \hat{\pi}(\theta|X^n_{[X_0]}) - \hat{\pi}(\theta|X^n)\right) \\
&= (n+1)\left(\frac{\eta_\theta(\tilde{\epsilon})}{\int \eta_{\theta'}(\tilde{\epsilon})\mathrm{d}\theta'} - \frac{\eta_\theta(\epsilon)}{\int \eta_{\theta'}(\epsilon)\mathrm{d}\theta'}\right) = (n+1)\frac{\Lambda_n(\eta_\theta(\epsilon) + \Delta_{\theta,\tilde{\epsilon},\epsilon}) - \eta_\theta(\epsilon)\left(\Lambda_n + \int \Delta_{\theta',\tilde{\epsilon},\epsilon}\mathrm{d}\theta'\right)}{\left(\Lambda_n + \int \Delta_{\theta',\tilde{\epsilon},\epsilon}\mathrm{d}\theta'\right)\Lambda_n} \\
&= (n+1)\frac{\Lambda_n \Delta_{\theta,\tilde{\epsilon},\epsilon} - \eta_\theta(\tilde{\epsilon})\int \Delta_{\theta',\tilde{\epsilon},\epsilon}\mathrm{d}\theta'}{\left(\Lambda_n + \int \Delta_{\theta',\tilde{\epsilon},\epsilon}\mathrm{d}\theta'\right)\Lambda_n} \leq (n+1)\frac{\Lambda_n \Delta_{\theta,\tilde{\epsilon},\epsilon}}{\left(\Lambda_n + \int \Delta_{\theta',\tilde{\epsilon},\epsilon}\mathrm{d}\theta'\right)\Lambda_n} \\
&\leq (n+1)\frac{\Lambda_n \Delta_{\theta,\tilde{\epsilon},\epsilon}}{\Lambda_n^2} = \frac{1}{\Lambda_n}(n+1)(\eta_\theta(\tilde{\epsilon}) - \eta_\theta(\epsilon)).
\end{aligned}$$

Finally, applying $\eta_\theta(\tilde{\epsilon}) - \eta_\theta(\epsilon) \leq \beta\pi(\theta)(\tilde{\epsilon} - \epsilon)$, we obtain

$$\lim_{\|X_0\|\to\infty} \mathrm{SC}^\theta_{n+1}(X_0) \leq -\frac{\beta\pi(\theta)}{\Lambda_n(1+\gamma)} \log\left(1 - \frac{1}{n^2}\right)^{n+1}$$

as desired.

If Λ_n converges to a nonzero value H, we have

$$\begin{aligned}
\lim_{n\to\infty} -\frac{\beta\pi(\theta)}{\Lambda_n(1+\gamma)} \log\left(1 - \frac{1}{n^2}\right)^{n+1} &= -\frac{\beta\pi(\theta)}{\Lambda(1+\gamma)}\left(\lim_{n\to\infty} \log\left(1 - \frac{1}{n^2}\right)\left(\left(1 - \frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}}\right) \\
&= -\frac{\beta\pi(\theta)}{\Lambda(1+\gamma)} \log\left((1-0)\left(\frac{1}{e}\right)^0\right) = 0.
\end{aligned}$$

□

The following Proposition 1 is used in the proof of Theorem 1. Proposition 1 reflects the smoothness of $\widehat{D}^{n,m}$ to show that the transformation $Y_1 \mapsto \widehat{D}^{n,m}$ results in a continuous random variable.

Proposition 1 ($\{\widehat{D}^{n,m} = c\}$ has zero measure.)**.** *Fix distinct points $(Y_2, \ldots, Y_m)$ and define $f(Y_1) \coloneqq \widehat{D}^{n,m}$. Then, for any $c \in \mathbb{R}$, the set $f^{-1}(\{c\})$ has Lebesgue measure zero.*

Proof. We start by observing that the space of Y_1, namely $\mathbb{R}^d$, can be split into a finite family of disjoint open sets $\{U_l\}_{l=1}^L$ such that $U^{\mathrm{c}} \coloneqq \mathbb{R}^d \setminus \left(\bigcup_{l=1}^L U_l\right)$ has measure zero and that for all $Y_1 \in U_l$, the k-NN (more precisely, the index of the k-NN point) of X_i ($i \in [n]$) among $\{Y_j\}_{j=1}^m$ and that of Y_j among $\{Y_{j'}\}_{j' \neq j}$ are identical. Such a partition makes the problem easier because within each partition cell, U_l, the k-NN distances $\nu_k(i)$ and $\mu_k(j)$ take the simple form as mere Euclidean distances between two predetermined points.

Such $\{U_l\}_{l=1}^L$ can be constructed as follows. Define $A_{ji} = \|Y_j - X_i\|$ and $B_{jj'} = \|Y_j - Y_{j'}\|$ and consider the distance matrices

$$A = \begin{pmatrix} A_{11} & \cdots & \cdots & A_{1n} \\ \vdots & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}, B = \begin{pmatrix} B_{11} & \cdots & \cdots & B_{1m} \\ \vdots & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{pmatrix}.$$

For each point X_i or $Y_{j'}$, the corresponding k-NN points are determined by the order of the elements in the corresponding columns $A_{\cdot,i}$ and $B_{\cdot,j'}$. In A, the only variables with respect to Y_1 are the first row. Similarly, the variables in B are the first row and the first column. In other words, the bottom-right blocks obtained by removing the first rows and first columns are constant with respect to Y_1.

Let us first consider A. The k-NN points for each X_i can be determined by finding where A_{1i} is ranked among the ranking of column i. Since the elements of column i except the first element, $(A_{2i}, \ldots, A_{mi})$, is constant with respect to Y_1, they can be sorted as $(A_{(2),i}, \ldots, A_{(m),i})$ in ascending order to define a partitioning of $\mathbb{R}^d$ in each of which A_{1i} has the same ranking among the elements in the column i: $V_j^i = \{Y_1 \in \mathbb{R}^d : \|Y_1 - X_i\| \in (A_{(j),i}, A_{(j+1),i})\}$ ($j \in [m]$), where $A_{(1),i} = 0$ and $A_{(m+1),i} = \infty$. By taking the intersections of such partitions, $V_{(j_1,\ldots,j_n)} = V_{j_1}^1 \cap \cdots \cap V_{j_n}^n$, we obtain a family of disjoint open sets $\mathcal{V} \coloneqq \{V_{(j_1,\ldots,j_n)}\}_{(j_1,\ldots,j_n)\in[m]^n}$ that covers almost everywhere $\mathbb{R}^d$ because each $V^{(i)\mathrm{c}} \coloneqq \mathbb{R}^d \setminus \bigcup_{j\in[m]} V_j^i$ has Lebesgue measure zero and

$$\mathbb{R}^d = \bigcap_i \mathbb{R}^d = \bigcap_i \left(V^{(i)\mathrm{c}} \cup \bigcup_{j_i} V_{j_i}^i \right) = V^{\mathrm{c}} \cup \bigcap_i \bigcup_{j_i} V_{j_i}^i$$

where V^{c} is a set with less Lebesgue measure than the sum of the measures of $V^{(i)\mathrm{c}}$ hence has zero measure.

Similarly, let us consider B. The second-to-last columns of B can be treated in the same way as A to obtain the almost-everywhere finite partition $\mathcal{W}_j$ of $\mathbb{R}^d$ for each column $j = 2, \ldots, m$ in which the ranking of B_{1j} remains invariant for each column (note that, although the diagonal elements of B are not used for determining the k-NN points, their existence does not affect the above construction). Now we consider the first column and construct an almost-everywhere partition of $\mathbb{R}^d$ in each of which the ordering of $\|Y_2 - Y_1\|, \ldots, \|Y_m - Y_1\|$ does not change. The existence of such a finite partition is guaranteed by the existence of l-th degree Voronoi diagrams for $l = 1, \ldots, m-1$ [3, 26]. In l-th degree Voronoi diagram $\{W_a^{(l)}\}_a$, each cell $W_a^{(l)}$ represents a region in which Y_1 has the same set of points as the l-nearest neighbors. Therefore, by taking the intersections $W_{(a_1,\ldots,a_{m-1})} = W_{a_1}^{(1)} \cap \cdots \cap W_{a_{m-1}}^{(m-1)}$, we obtain regions in each of which the ordering of the distances $\|Y_2 - Y_1\|, \ldots, \|Y_m - Y_1\|$ remain the same. There are only finite regions in the l-th degree Voronoi diagram for all $l = 1, \ldots, m$, hence the family of their intersections are also finite, and the boundaries of Voronoi cells have zero Lebesgue measure as they correspond to the sets where two of the sites are at an equal distance. Therefore, we have obtained the desired partition which we denote by $\mathcal{W}_1$.

By taking all intersections of the above partitions, $\mathcal{V}, \{\mathcal{W}_j\}_{j=1}^m$, we obtain the desired finite partition $\mathcal{U} = \{U_l\}_{l=1}^L$ that covers almost everywhere $\mathbb{R}^d$ and in each U_l, the indices of the k-NN points remain the same.

Let us define $f_l \coloneqq f|_{U_l}$. Now, we show that each $f^{-1}(\{c\}) \cap U_l$ has zero measure. In each U_l, the distances $\nu_k(i)$ and $\mu_k(j)$ are strictly positive as no two points overlap. Therefore, $f_l : U_l \to \mathbb{R}$ is a real analytic function since it is a composition of analytic functions:

$$\begin{aligned} f_l(Y_1) &= -\frac{1}{\gamma} \log \left(\sum_{i=1}^n (\nu_k(i))^{-\gamma d} \right) + \frac{1}{1+\gamma} \log \left(\sum_{j=1}^m (\mu_k(j))^{-\gamma d} \right) + \text{const.} \\ &= -\frac{1}{\gamma} \log \left(\sum_{i=1}^n \exp\left(-\gamma d \frac{1}{2} \log(\nu_k(i))^2 \right) \right) + \frac{1}{1+\gamma} \log \left(\sum_{j=1}^m \exp\left(-\gamma d \frac{1}{2} \log(\mu_k(j))^2 \right) \right) + \text{const.}, \end{aligned}$$

and $(\nu_k(i))^2, (\mu_k(j))^2$ are either quadratic forms of Y_1 or constants. As a result, $f_l^{-1}(\{c\}) = f^{-1}(\{c\}) \cap U_l$ is a zero set of a real analytic function $f_l - c$ that is not a constant function, hence has zero Lebesgue measure [23, Lemma 1.2], [56].

Finally, the assertion of the proposition follows immediately from

$$\lambda(f^{-1}(\{c\})) = \lambda\left(f^{-1}(\{c\}) \cap \left(U^{\mathrm{c}} \cup \bigcup_{l=1}^{L} U_l\right)\right) \leq \lambda\left(f^{-1}(\{c\}) \cap U^{\mathrm{c}}\right) + \sum_{l=1}^{L} \lambda\left(f^{-1}(\{c\}) \cap U_l\right),$$

where we denoted the Lebesgue measure by λ. □

B.3 Remarks

Remark 1 (Relation to redescending property of influence functions)**.** *It should be noted that the above theorem is a finite-sample analogue of the redescending property of influence functions. In the case of influence functions, redescending property is defined as convergence to zero under $\|X_0\| \to \infty$ [55]. The discrepancy that the limit in our case is nonzero (only converges to zero with $n \to \infty$) stems from the fact that we consider the finite sample analogue, namely, the sensitivity curve. This is intuitively comprehensible since the influence function reflects the response to contamination in the underlying distribution, i.e., a population quantity.*

Remark 2 (The reason to consider sensitivity curve instead of influence functions.)**.** *The reason we consider SC instead of IF is two-fold: (1) we are interested in the pseudo-posterior distribution $\hat{\pi}(\theta|X^n)$ with respect to a finite sample X^n, hence the SC can more precisely provide the information of our interest, and (2) the IF of the quantities based on the considered divergence estimator may not even exist. The definition of the considered divergence estimator is based on k-NN density estimators, and it does not have a straightforward representation as a statistical functional (i.e., a functional of the underlying data distribution). Furthermore, even if we consider the divergence estimator as a functional of the underlying probability density function of the data, the k-NN density estimator is not square-integrable if $k = 1$ [8, Proposition 3.1], hence the standard definition of influence functions as a dual point in the Hilbert space L^2 is not applicable. Therefore, we consider the sensitivity curve for the theoretical analysis, which can directly reflect the detailed procedure to construct the estimate from given data points.*

Remark 3 (Diverging limit and diverging sequence limit)**.** *In the proof, we used the fact that if $\lim_{i\to\infty} f(X'_j) = L$ for any diverging sequence $\{X'_j\}_{i=1}^{\infty}$ (i.e., $\|X'_j\| \to \infty$), we have $\lim_{\|X_0\|\to\infty} f(X_0) = L$. We show a proof by contradiction. First recall that $\lim_{\|X_0\|\to\infty} f(X_0) = L$ means that for any $\epsilon > 0$, there exists $B > 0$ such that for any X_0 satisfying $\|X_0\| > B$ it holds that $|f(X_0) - L| < \epsilon$. To show this by contradiction, assume that there exists $\epsilon > 0$ such that for any $B > 0$ there exists X_0 satisfying $\|X_0\| > B$ and $|f(X_0) - L| \geq \epsilon$. Now fix such an ϵ and define $B_i \coloneqq 2^i$ for $i \in \mathbb{N}$. By assumption, there exist a sequence $\{x_i\}_{i=1}^{\infty}$ such that $\|x_i\| > B_i$ and $|f(x_i) - L| \geq \epsilon$. Because $\{x_i\}_{i=1}^{\infty}$ is a diverging sequence, it has to hold that $\lim_{i\to\infty} f(x_i) = L$. This is a contradiction.*

Remark 4 (Exchanging the limits)**.** *The current statement of the theorem takes the limit of $\lim_{\|X_0\|\to\infty}$ for each fixed n. One should note that A_1 and A_2 in the proof depend on n and the sample X^n. Similarly, $\rho_k^{-d\gamma}(0)$ and $\nu_k^{-d\gamma}(0)$ depend on the sample. Therefore, care should be taken if one wants to merge the two limit operations $\lim_{n\to\infty}$ and $\lim_{\|X_0\|\to\infty}$.*

C Preliminaries for Asymptotic Analysis

In this section, we summarize several specific lemmas and theorems to show the asymptotic properties of the proposed discrepancy in Eq. (7). Here, we denote $\to_w$, $\to_d$ and $\to_p$ as the *weak convergence* of distribution functions, the convergence of random variables *in distribution* and the convergence of random variables *in probability*, respectively.

Remembering the fact that $\rho_k(i)$ is a random variable, which is the measure of discrepancy between X_i and its k-th nearest neighbor in $X^n \backslash X_i$, the following lemmas and theorems hold.

Lemma 1. *Let $\zeta_{n,k,1} \coloneqq \log(n-1)\rho_k^d(1)$ be a random variable, and let $F_{n,k,x}(u) \coloneqq \Pr(\zeta_{n,k,1} < u | X_1 = x)$ denotes its conditional distribution function. Then,*

$$F_{n,k,x}(u) = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j},$$

where $P_{n,u,x} \coloneqq \int_{\mathcal{M}\cap\mathcal{B}(x,R_n(u))} p(t)\mathrm{d}t$ and $R_n(u) \coloneqq (e^u/(n-1))^{1/d}$.

Proof. We can obtain

$$\begin{aligned}
&F_{n,k,x}(u) = \Pr(\zeta_{n,k,1} < u | X_1 = x) \\
&= \Pr(\log(n-1)\rho_k^d(1) < u | X_1 = x) = \Pr\left(\rho_k(1) < \left(\frac{e^u}{n-1}\right)^{1/d} | X_1 = x\right) \\
&= \Pr\left(\rho_k(1) < R_n(u) | X_1 = x\right) \text{ (because } R_n(u) \coloneqq (e^u/(n-1))^{1/d}).
\end{aligned}$$

The last expression can be interpreted as the probability of k or more elements from $\{X_2 \dots X_n\}$ being contained in $\mathcal{M} \cap \mathcal{B}(x, R_n(u))$ given $X_1 = x$. Since we have i.i.d. observations, this condition can be ignored. Therefore, we can see this probability as binomial distribution and obtain

$$\begin{aligned}
F_{n,k,x}(u) &= \Pr\left(\rho_k(1) < R_n(u) | X_1 = x\right) \\
&= \sum_{j=k}^{n-1} \begin{pmatrix} n-1 \\ j \end{pmatrix} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1} \begin{pmatrix} n-1 \\ j \end{pmatrix} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j},
\end{aligned}$$

and the claim holds. □

Lemma 2 (Log-Erlang distribution)**.** *Let u be a random variable from the Erlang distribution as*

$$f_{x,k}(u) = \frac{1}{\Gamma(k)} \lambda(x)^k u^{k-1} \exp(-\lambda(x)u),$$

where $\lambda(x) > 0$ and $k \in \mathbb{Z}^+$. Here, $\mathbb{Z}^+$ denotes the set of positive integer. Then, $l = \log u$ is a random variable from the log-Erlang distribution as

$$g_{n,k}(l) = \frac{1}{\Gamma(k)} \lambda(x)^k (\exp(l))^k \exp(-\lambda(x)\exp(l)).$$

Proof. If we set $l = \log u$, we obtain $u = \exp(l)$ and $\frac{\mathrm{d}l}{\mathrm{d}u} = \frac{1}{u} = \frac{1}{\exp(l)}$. When we denote the distribution of l as $g_{n,k}(l)$,

$$\begin{aligned}
g_{n,k}(l) &= f_{n,k}(u)\left|\frac{\mathrm{d}u}{\mathrm{d}l}\right| = \frac{1}{\Gamma(k)} \lambda(x)^k u^{k-1} \exp(-\lambda(x)u) \cdot \exp(l) \\
&= \frac{1}{\Gamma(k)} \lambda(x)^k (\exp(l))^{k-1} \exp(-\lambda(x)\exp(l)) \cdot \exp(l) = \frac{1}{\Gamma(k)} \lambda(x)^k (\exp(l))^k \exp(-\lambda(x)\exp(l)).
\end{aligned}$$

This is the same as the definition of the log-Gamma distribution. Because of $k \in \mathbb{Z}^+$, we can see that $g_{n,k}(l)$ is the log-Erlang distribution.

The claim is proved. □

Lemma 3 (Expectation of log-Erlang distribution)**.** *Let $f_{x,k}(u) \coloneqq \frac{1}{\Gamma(k)} \lambda(x)^k (\exp(l))^k \exp(-\lambda(x)\exp(l))$ be the density of the log-Erlang distribution with parameters $\lambda(x) > 0$ and $k \in \mathbb{Z}^+$. Then, the 1-th moments of the log-Erlang distribution can be calculated as*

$$\int_0^\infty u f_{x,k}(u) \mathrm{d}u = \psi(k) - \log(\lambda(x)),$$

where $\psi(\cdot)$ is a digamma function.

Proof. Because the function $f_{x,k}(u)$ is the density of the log-Erlang distribution, we obtain

$$\int_{\mathbb{R}} (\exp(u))^k \exp(-\lambda(x)\exp(u)) \mathrm{d}u = \int_{\mathbb{R}} \exp(ku - \lambda(x)\exp(u)) \mathrm{d}u = \Gamma(k)\lambda(x)^{-k}.$$

Differentiating the inside of the above integration by k, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}k}\exp(ku-\lambda(x)\exp(u)) = u\exp(ku-\lambda(x)\exp(u)) = u\cdot\Gamma(k)\lambda(x)^{-k}f_{x,k}(u).$$

Therefore, the expectation of u is written as

$$\begin{aligned}
\mathbb{E}[u] &= \int_0^\infty u f_{x,k}(u)\mathrm{d}u = \int_0^\infty u\frac{1}{\Gamma(k)}\lambda(x)^k\exp(ku-\lambda(x)\exp(u))\mathrm{d}u \\
&= \frac{\lambda(x)^k}{\Gamma(k)}\int_0^\infty u\exp(ku-\lambda(x)\exp(u))\mathrm{d}u = \frac{\lambda(x)^k}{\Gamma(k)}\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}k}\exp(ku-\lambda(x)\exp(u))\mathrm{d}u \\
&= \frac{\lambda(x)^k}{\Gamma(k)}\frac{\mathrm{d}}{\mathrm{d}k}\int_0^\infty \exp(ku-\lambda(x)\exp(u))\mathrm{d}u = \frac{\lambda(x)^k}{\Gamma(k)}\frac{\mathrm{d}}{\mathrm{d}k}\Gamma(k)\lambda(x)^{-k} \\
&= \frac{\lambda(x)^k}{\Gamma(k)}\left(\frac{\mathrm{d}}{\mathrm{d}k}\Gamma(k)\cdot\lambda(x)^{-k} - \Gamma(k)\cdot\lambda(x)^{-k}\log(\lambda(x))\right) \\
&= \frac{1}{\Gamma(k)}\frac{\mathrm{d}}{\mathrm{d}k}\Gamma(k) - \log(\lambda(x)) = \psi(k) - \log(\lambda(x)).
\end{aligned}$$

The claim is hold. □

We show the following properties on the log-Erlang distribution according to standard proof techniques in [48].

Lemma 4. *Suppose that Lebesgue-approximable function on p in Assumptions 2 and 3 holds. Let u be fixed. Then, $F_{n,k,x}(u) \to_w F_{k,x}(u)$ for almost all $x \in \mathcal{M}$, where*

$$F_{k,x}(u) \coloneqq 1 - \exp(-\lambda(x)\exp(u))\sum_{j=0}^{k-1}\frac{1}{j!}(\lambda(x)\exp(u))^j$$

is the log-Erlang distribution with $\lambda(x) = \bar{c}p(x)$.

Proof. According to Assumptions 2 and 3, we can see that for all $\delta > 0$ and almost all $x \in \mathcal{M}$ there exists $n_0(x,\delta,u) \in \mathbb{Z}_+$ such that if $n > n_0(x,\delta,u)$, then $\mathcal{B}(x,R_n) = \mathcal{B}(x,R_n)\cap\mathcal{M}$, and

$$p(x)-\delta < \frac{\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} p(t)\mathrm{d}t}{\frac{\exp(u)\bar{c}}{n-1}} < p(x)+\delta \quad \left(\mathcal{V}(\mathcal{B}(x,R_n)\cap\mathcal{M}) = \frac{\exp(u)\bar{c}}{n-1}\right).$$

Therefore, if $n > n_0(x,\delta,u)$,

$$\begin{aligned}
F_{n,k,u}(u) &= 1 - \sum_{j=0}^{k-1}\binom{n-1}{j}(P_{n,u,x})^j(1-P_{n,u,x})^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1}\binom{n-1}{j}\left(\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} p(t)\mathrm{d}t\right)^j\left(1-\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} p(t)\mathrm{d}t\right)^{n-1-j} \\
&\geq 1 - \sum_{j=0}^{k-1}\binom{n-1}{j}\left(\frac{\exp(u)}{n-1}\bar{c}(p(x)+\delta)\right)^j\left(1-\frac{\exp(u)}{n-1}\bar{c}(p(x)-\delta)\right)^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1}\frac{(n-1)!}{j!(n-1-j)!}\left(\frac{\exp(u)}{n-1}\bar{c}(p(x)+\delta)\right)^j\left(1-\frac{\exp(u)}{n-1}\bar{c}(p(x)-\delta)\right)^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1}\frac{1}{j!}\frac{(n-1)!}{(n-1-j)!(n-1)^j}\left(\exp(u)\bar{c}(p(x)+\delta)\right)^j\left(1-\frac{\exp(u)}{n-1}\bar{c}(p(x)-\delta)\right)^{n-1-j}.
\end{aligned}$$

Because of the fact that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{(n-1)!}{(n-1-j)!(n-1)^j} = 1, \\
&\lim_{n\to\infty}\left(1-\frac{\exp(u)}{n-1}\bar{c}(p(x)-\delta)\right)^{n-1-j} = \exp(-\exp(u)\bar{c}(p(x)-\delta)),
\end{aligned}$$

we obtain for all $\delta > 0$ and for almost all $x \in \mathcal{M}$,

$$\liminf_{n\to\infty} F_{n,k,u}(u) \geq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} \Big(\exp(u)\bar{c}(p(x)+\delta) \Big)^j \exp(-\exp(u)\bar{c}(p(x)-\delta)).$$

By choosing $\delta \to 0$, we can see that

$$\liminf_{n\to\infty} F_{n,k,u}(u) \geq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (\exp(u)\lambda(x))^j \exp(-\exp(u)\lambda(x)),$$

where $\lambda(x) \coloneqq \bar{c}p(x)$.

In the same way, we can show that for almost all $x \in \mathcal{M}$

$$\limsup_{n\to\infty} F_{n,k,u}(u) \leq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (\exp(u)\lambda(x))^j \exp(-\exp(u)\lambda(x)).$$

When we define $F_{k,x}(u) \coloneqq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (\exp(u)\lambda(x))^j \exp(-\exp(u)\lambda(x))$, the claim is proved. □

Lemma 5. *Let $\xi_{n,k,x}$ and $\xi_{k,x}$ be random variables with $F_{n,k,x}$ and $F_{k,x}$ distribution functions, and let $\kappa \in \mathbb{R}$ be arbitrary. Then for almost all $x \in \mathcal{M}$ we have that $\xi^\kappa_{n,k,x} \to_d \xi^\kappa_{k,x}$, where $f_n \to_d f$ indicates convergence of random variable f_n in distribution.*

Proof. According to Lemma 4, we obtain $F_{n,k,x}(u) \to_w F_{k,x}(u)$ for almost all of $x \in \mathcal{M}$. This is equal to the fact that $F_{n,k,x}(u) \to_d F_{k,x}(u)$ for almost all of $x \in \mathcal{M}$. Since the function of $(\cdot)^\kappa$ is continuous on $(0,\infty)$ and $X_i \in (0,\infty)$ almost surely, by using the continuous mapping theorem ([78]), the claim is proved. □

For proving Corollary 1, we introduce the Lévy's Upward Theorem as follow.

Theorem 4 (Lévy's Upward Theorem)**.** *Let $\{Z_n\}_{n\geq 0}$ be a collection of random variables, and let $\mathcal{F}_n$ be a filtration on the same probability space. If $\sup_{n\geq 0}|Z_n|$ is integrable, $Z_n \to Z_\infty$ almost surely as $n \to \infty$ and $\mathcal{F}_n \uparrow F_\infty$, then $\mathbb{E}[Z_n|\mathcal{F}_n] \to \mathbb{E}[Z_\infty|\mathcal{F}_\infty]$ both almost surely and in mean.*

To show Theorem 2, we analyze the following asymptotic behavior of the logarithm of random variable.

Theorem 5 (Theorem 21 in Poczos and Schneider [64])**.** *Suppose that the boundedness of an expectation on p in Assumptions 2 and 3 holds. If $0 \leq \kappa$ and $\xi^\kappa_{n,k,x} \to_d \xi^\kappa_{k,x}$, or $-k < \kappa < 0$ and $\xi^\kappa_{n,k,x} \to_d \xi^\kappa_{k,x}$, then $\lim_{n\to\infty} \mathbb{E}[\xi^\kappa_{n,k,x}] = \mathbb{E}[\xi^\kappa_{k,x}]$.*

Theorem 6 (The asymptotic expectation)**.** *Suppose that the boundedness of an expectation on p in Assumptions 2 and 3 holds. If $-k < \kappa < 0$, or $0 \leq \kappa$, then we obtain*

$$\lim_{n\to\infty} \mathbb{E}\Big[\log(n-1)^\kappa \rho^{d\kappa}_k(1) | X_1 = x \Big] = \kappa(\psi(k) - \log(\bar{c}p(x))),$$
$$\lim_{m\to\infty} \mathbb{E}\Big[\log m^\kappa \nu^{d\kappa}_k(1) | X_1 = x \Big] = \kappa(\psi(k) - \log(\bar{c}q(x))).$$

Proof. It is enough to show the first equation because the second equation can be showed in the same way. According to Lemma 5, we obtain $\xi^\kappa_{n,k,x} \to_d \xi^\kappa_{k,x}$ for almost all $x \in \mathcal{M}$. Then,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\Big[\log(n-1)^\kappa \rho^{d\kappa}_k(1) | X_1 = x \Big] = \kappa \lim_{n\to\infty} \mathbb{E}\Big[\log(n-1) \rho^{d}_k(1) | X_1 = x \Big] \\
&= \kappa \lim_{n\to\infty} \mathbb{E}\Big[\zeta_{n,k,1} | X_1 = x \Big] = \kappa \lim_{n\to\infty} \mathbb{E}\Big[\xi_{n,k,x} \Big] = \kappa \mathbb{E}[\lim_{n\to\infty} \xi_{n,k,x}] \ \ (\xi^\kappa_{n,k,x} \to_d \xi^\kappa_{k,x} \text{ by Lemma 5}) \\
&= \kappa \mathbb{E}[\xi_{k,x}] \text{ (by Theorem 5)} \\
&= \kappa \int_0^\infty u f_{x,k}(u) \mathrm{d}u = \kappa(\psi(k) - \log(\lambda(x))) \text{ (by Lemma 3)} \\
&= \kappa(\psi(k) - \log(\bar{c}p(x))).
\end{aligned}$$

Thus, the claim is proved. □

Theorem 7. *Suppose that the boundedness of an expectation on q in Assumption 2-4 holds. If $-k < \kappa < 0$, or $0 \leq \kappa$, then we obtain*

$$\lim_{m\to\infty} \mathbb{E}\left[\log(m-1)^{\kappa}\bar{\rho}_k^{d\kappa}(1)|Y_1 = y\right] = \kappa(\psi(k) - \log(\bar{c}q(y))).$$

Proof. We can show this in the same way of Theorem 6 by substituting n, ρ_k^d, x to $m, \bar{\rho}_k^d, y$. □

To show Theorem 3, we focus on $\hat{p}_{k(n)}^{\gamma}(x)$, $\hat{q}_{k(n)}^{\gamma}(x)$ and $\hat{q}_{k(m)}^{\gamma}(y)$ and guarantee the convergence in probability of each estimators.

Lemma 6 (Moments of inverse Erlang distribution)**.** *Let $f_{x,k} = \frac{1}{\Gamma(k)}\lambda^k(x)u^{-1-k}\exp(-\lambda(x)/u)$ be the density of inverse Erlang distribution with parameters $\lambda(x) > 0$ and $k \in \mathbb{Z}^+$. Let $\kappa \in \mathbb{R}$ such that $\kappa < k$. Then, the κ-th moments of inverse Erlang distribution can be calcurated as*

$$\int_0^{\infty} u^{\kappa} f_{x,k}(u)\mathrm{d}u = \lambda^{\kappa}(x)\frac{\Gamma(k-\kappa)}{\Gamma(k)}.$$

Proof. The κ-th moments of $f_{x,k}$ is

$$\begin{aligned}\int_0^{\infty} u^{\kappa} f_{x,k}(u)\mathrm{d}u &= \int_0^{\infty} u^{\kappa}\frac{1}{\Gamma(k)}\lambda^k(x)u^{-1-k}\exp(-\lambda(x)/u)\mathrm{d}u \\ &= \frac{\lambda^k(x)}{\Gamma(k)}\int_0^{\infty} u^{-1-(k-\kappa)}\exp(-\lambda(x)/u)\mathrm{d}u.\end{aligned}$$

If $k > \kappa$, the integral term in the above equals to the marginalization of inverse gamma distribution. Thus,

$$\begin{aligned}\int_0^{\infty} u^{\kappa} f_{x,k}(u)\mathrm{d}u &= \frac{\lambda^k(x)}{\Gamma(k)}\int_0^{\infty} u^{-1-(k-\kappa)}\exp(-\lambda(x)/u)\mathrm{d}u \\ &= \frac{\lambda^k(x)}{\Gamma(k)}\frac{\Gamma(k-\kappa)}{\lambda^{k-\kappa}(x)} = \lambda^{\kappa}(x)\frac{\Gamma(k-\kappa)}{\Gamma(k)}.\end{aligned}$$

The claim is proved. □

Lemma 7 ($\hat{p}_{k(n)}^{\gamma}(x)$ converges to $p^{\gamma}(x)$ in probability)**.** *Suppose that Assumptions 2 and 3 are satisfied. Let $\kappa = \gamma < k$. If $k(n)$ denotes the number of neighbors applied at sample size n, $\lim_{n\to\infty} k(n) = \infty$ and $\lim_{n\to\infty} n/k(n) = \infty$, then $\hat{p}_{k(n)}^{\gamma}(x) \to_p p_{k(n)}^{\gamma}(x)$ for almost all x.*

Proof. According to the Chebyshev's inequality, if we set $X_i = x$, $k(n) = k$ and $\epsilon > 0$, we obtain

$$\begin{aligned}\mathbb{P}(|\hat{p}_k^{\gamma}(x) - p_k^{\gamma}(x)| > \epsilon) &\leq \frac{1}{\epsilon^2}\mathbb{V}[\hat{p}_k^{\gamma}(x)] = \frac{1}{\epsilon^2}\mathbb{V}\left[\left(\frac{k}{(n-1)\bar{c}\rho_k^d(i)}\right)^{\gamma}\right] \\ &= \frac{1}{\epsilon^2}\left(\frac{k}{(n-1)\bar{c}}\right)^{2\gamma}\mathbb{V}\left[\frac{1}{\rho_k^{d\gamma}(i)}\right] = \frac{1}{\epsilon^2}\left(\frac{1}{\bar{c}}\right)^{2\gamma}\left(\frac{k}{n-1}\right)^{2\gamma}\mathbb{V}\left[\frac{1}{\rho_k^{d\gamma}(i)}\right].\end{aligned}$$

According to Corollary 1 of Pérez-Cruz [61], the random variable $\rho_k^d(i)$ measures the waiting time between the origin and the k-th event of a uniformly spaced distribution, and this waiting time is distributed as an Erlang distribution or a unit-mean and $1/k$ variance gamma distribution. Therefore, the random variable $1/\rho_k^d(i)$ is distributed as an inverse Erlang distribution.

According to Lemma 6 and $\gamma < k$, the moments of $1/\rho_k^{d\gamma}(i)$ can be calculated. Therefore, we can see $\mathbb{V}\left[\frac{1}{\rho_k^{d\gamma}(i)}\right] < \infty$. According to the assumption that $\lim_{n\to\infty} n/k(n) = \infty$, we obtain $\lim_{n\to\infty} k(n)/n = 0$ and therefore

$$\lim_{n\to\infty}\mathbb{P}(|\hat{p}_k^{\gamma}(x) - p_k^{\gamma}(x)| > \epsilon) \leq \lim_{n\to\infty}\frac{1}{\epsilon^2}\left(\frac{1}{\bar{c}}\right)^{2\gamma}\left(\frac{k}{n-1}\right)^{2\gamma}\mathbb{V}\left[\frac{1}{\rho_k^{d\gamma}(i)}\right] = 0,$$

for any x in the support of $p(x)$ and any ϵ. The claim is proved. □

Lemma 8 ($\hat{q}^{\gamma}_{k(n)}(x)$ converges to $q^{\gamma}(x)$ in probability)**.** *Suppose that Assumptions 2 and 3 are satisfied. Let* $0 < \kappa = \gamma < k$. *If* $k(n)$ *denotes the number of neighbors applied at sample size* n, $\lim_{n\to\infty} k(n) = \infty$ *and* $\lim_{n\to\infty} n/k(n) = \infty$, *then* $\hat{q}^{\gamma}_{k(n)}(x) \to_p q^{\gamma}_{k(n)}(x)$ *for almost all* x.

Proof. It can be shown in the same way of Lemma 7. □

Lemma 9 ($\hat{q}^{\gamma}_{k(m)}(y)$ converges to $q^{\gamma}(y)$ in probability)**.** *Suppose that Assumptions 2-4 are satisfied. Let* $\kappa = \gamma < k$. *If* $k(n)$ *denotes the number of neighbors applied at sample size* m, $\lim_{m\to\infty} k(m) = \infty$ *and* $\lim_{m\to\infty} n/k(m) = \infty$, *then* $\hat{q}^{\gamma}_{k(m)}(y) \to_p q^{\gamma}_{k(m)}(y)$ *for almost all* y.

Proof. It can be shown in the same way of Lemma 7 by substituting n, ρ^d_k, x to $m, \bar{\rho}^d_k, y$. □

D Proofs for Asymptotic Analysis

In this section, we summarize the essential theoretical analysis for our estimator to guarantee the main characteristics.

D.1 Proof of Theorem 2

The following lemma is necessary to show Theorem 2.

Lemma 10 (Switching limit and expectation)**.** *Let* $\kappa > 0$ *or* $-k < \kappa < 0$.*Then, the following equality holds.*

$$\lim_{n\to\infty}\int_{\mathcal{M}} f_n(x)p(x)\mathrm{d}x = \int_{\mathcal{M}}\lim_{n\to\infty} f_n(x)p(x)\mathrm{d}x,$$
$$\lim_{m\to\infty}\int_{\mathcal{M}} g_m(x)p(x)\mathrm{d}x = \int_{\mathcal{M}}\lim_{m\to\infty} g_m(x)p(x)\mathrm{d}x,$$
$$\lim_{m\to\infty}\int_{\mathcal{M}'} \bar{g}_m(y)q(y)\mathrm{d}y = \int_{\mathcal{M}'}\lim_{m\to\infty} \bar{g}_m(y)q(y)\mathrm{d}y,$$

where

$$f_n(x) \coloneqq \mathbb{E}\left[\log(n-1)^{\kappa}\rho^{d\kappa}_k(1)|X_1 = x\right],\ g_m(x) \coloneqq \mathbb{E}\left[\log m^{\kappa}\nu^{d\kappa}_k|X_1 = x\right],\ \bar{g}_m(y) \coloneqq \mathbb{E}\left[\log(m-1)^{\kappa}\bar{\rho}^{d\kappa}_k|Y_1 = y\right].$$

Proof. Poczos and Schneider [64] proved in Theorem 37

$$f'_n(x) \coloneqq \int_0^{\infty} u^{\kappa}F'_{n,k,x_1}\mathrm{d}u \le \kappa L(x,1,\kappa,k,p,\delta,\delta_1) < \infty\ (\kappa > 0),$$
$$f'_n(x) \coloneqq\le \kappa\left[\frac{\hat{L}(\bar{p},1)}{k+\kappa} - \frac{1}{\kappa}\right] < \infty\ (-k < \kappa < 0),$$

where

$$L(x,\omega,\kappa,k,p,\delta,\delta_1) \coloneqq \delta_1 + \delta_1\int \|x-y\|^{\kappa}p(y)\mathrm{d}y + (\bar{c}r(x))^{-\kappa}H(x,p,\delta,\omega),$$

and

$$f'_n(x) \coloneqq \mathbb{E}\left[(n-1)^{\kappa}\rho^{d\kappa}_k(1)|X_1 = x\right],$$

and F'_{n,k,x_1} is the conditional density function for $\zeta^{'\kappa}_{n,k,x_1} = (n-1)\rho^d_k(1)$. According to the fact that if $a(x) \le b(x)$ then $\mathbb{E}[a(x)] \le \mathbb{E}[b(x)]$, we can obtain

$$f_n(x) \le f'_n(x) < \infty.$$

We can also obtain

$$g'_m(x) < \infty,\ \bar{g}'_m(y) < \infty,$$

where

$$g'_m(x) \coloneqq \mathbb{E}\left[m^\kappa \nu_k^{d\kappa}(1)|X_1 = x\right],\ \bar{g}'_m(y) \coloneqq \mathbb{E}\left[(m-1)^\kappa \bar{\rho}_k^{d\kappa}(1)|Y_1 = y\right],$$

In the same way as Theorem 37 of Poczos and Schneider [64]. Therefore, the following inequality holds:

$$g_m(x) \leq g'_m(x) < \infty,\ \bar{g}_m(y) \leq \bar{g}'_m(y) < \infty.$$

From these, for $0 < \kappa < k$ or $-k < \kappa < 0$, we can see that under the conditions in Theorem 2, there exist some functions J_1, J_2, J_3 and threshold numbers $N_{p,q,1}, N_{p,q,2}, N_{p,q,3}$ such that if $n, m > N_{p,q,1}$, $n, m > N_{p,q,2}$ and $n, m > N_{p,q,3}$, then for almost all $x \in \mathcal{M}$ and $y \in \mathcal{M}'$, $f_n(x) \leq J_1(x)$, $g_m(x) \leq J_2(x)$ and $\bar{g}_m(y) \leq J_3(y)$ and $\int_M J_1(x)p(x)\mathrm{d}x < \infty$, $\int_M J_2(x)p(x)\mathrm{d}x < \infty$ and $\int_{M'} J_3(x)q(y)\mathrm{d}y < \infty$. By applying the Lebesgue dominated convergence theorem, the claim is proved. □

By using these lemmas and theorem in Appendix C and Lemma 10, we show asymptotic unbiasedness of our estimator claimed in Theorem 8 and 9.

Theorem 8 (Asymptotic unbiasedness)**.** *Let $\kappa \coloneqq \gamma$ and suppose $0 < \gamma < k$. Suppose that Assumptions 2-4 are satisfied, and that q is bounded from above. Then, $\widehat{D}_\gamma(X^n\|Y^m)$ is asymptotically unbiased, i.e.,*

$$\lim_{n,m\to\infty} \mathbb{E}\left[\widehat{D}_\gamma(X^n\|Y^m)\right] = D_\gamma(p\|q),$$

where $\widehat{D}_\gamma(X^n\|Y^m)$ is defined in Eq. (7).

Proof. Now, we want to show that

$$D_\gamma(p\|q) = \lim_{n,m\to\infty} \mathbb{E}\left[\widehat{D}_\gamma(p(X^n)\|q(Y^m))\right].$$

If we use Eq. (7) as the γ-divergence estimator, it can be rewritten as

$$\begin{aligned}
&\widehat{D}_\gamma(p(X^n)\|q(Y^m)) \\
&= \frac{1}{\gamma(1+\gamma)}\left[\log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{k}{(n-1)\bar{c}\rho_k^d(i)}\right)^\gamma\right) - (1+\gamma)\log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{k}{m\bar{c}\nu_k^d(i)}\right)^\gamma\right)\right. \\
&\qquad\left. + \gamma\log\left(\frac{1}{m}\sum_{j=1}^m\left(\frac{k}{(m-1)\bar{c}\bar{\rho}_k^d(j)}\right)^\gamma\right)\right] \\
&= \frac{1}{\gamma(1+\gamma)}\left[\log\left(\frac{k}{\bar{c}}\right)^\gamma + \log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{(n-1)\rho_k^d(i)}\right)^\gamma\right) - (1+\gamma)\log\left(\frac{k}{\bar{c}}\right)^\gamma\right. \\
&\qquad - (1+\gamma)\log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{m\nu_k^d(i)}\right)^\gamma\right) + \gamma\log\left(\frac{k}{\bar{c}}\right)^\gamma + \gamma\log\left(\frac{1}{m}\sum_{j=1}^m\left(\frac{1}{(m-1)\bar{\rho}_k^d(j)}\right)^\gamma\right)\Bigg] \\
&= \frac{1}{\gamma(1+\gamma)}\left[\log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{(n-1)\rho_k^d(i)}\right)^\gamma\right) - (1+\gamma)\log\left(\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{m\nu_k^d(i)}\right)^\gamma\right)\right. \\
&\qquad\left. + \gamma\log\left(\frac{1}{m}\sum_{j=1}^m\left(\frac{1}{(m-1)\bar{\rho}_k^d(j)}\right)^\gamma\right)\right]. \qquad (14)
\end{aligned}$$

Taking expectation and a limit and switching the limit and expectation by using Lemma 10, we can obtain

$$\begin{aligned}
&\lim_{n,m\to\infty}\mathbb{E}\left[\widehat{D}_\gamma(p(X^n)\|q(Y^m))\right]\\
&=\lim_{n,m\to\infty}\frac{1}{\gamma(1+\gamma)}\left[\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{(n-1)\rho_k^d(i)}\right)^\gamma\right)-(1+\gamma)\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m\nu_k^d(i)}\right)^\gamma\right)\right.\\
&\qquad\left.+\gamma\log\left(\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{(m-1)\bar{\rho}_k^d(j)}\right)^\gamma\right)\right]\\
&=\lim_{n,m\to\infty}\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[\mathbb{E}\left[\log\left(\frac{1}{(n-1)^\gamma\rho_k^{d\gamma}(1)}\right)\middle|X_1=x\right]-(1+\gamma)\mathbb{E}\left[\log\left(\frac{1}{m^\gamma\nu_k^{d\gamma}(1)}\right)\middle|X_1=x\right]\right]\\
&\qquad+\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\mathbb{E}\left[\log\left(\frac{1}{(m-1)^\gamma\bar{\rho}_k^{d\gamma}(j)}\right)\middle|Y_1=y\right]\right]\\
&=\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[\lim_{n\to\infty}\mathbb{E}\left[\log\left(\frac{1}{(n-1)^\gamma\rho_k^{d\gamma}(1)}\right)\middle|X_1=x\right]-(1+\gamma)\lim_{m\to\infty}\mathbb{E}\left[\log\left(\frac{1}{m^\gamma\nu_k^{d\gamma}(1)}\right)\middle|X_1=x\right]\right]\\
&\qquad+\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\lim_{m\to\infty}\mathbb{E}\left[\log\left(\frac{1}{(m-1)^\gamma\bar{\rho}_k^{d\gamma}(j)}\right)\middle|Y_1=y\right]\right].
\end{aligned}$$

According to Theorem 7, we obtain

$$\begin{aligned}
&\lim_{n,m\to\infty}\mathbb{E}\left[\widehat{D}_\gamma(p(X^n)\|q(Y^m))\right]\\
&\quad=\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[-\gamma(\psi(k)-\log(\bar{c}p(X_1)))+\gamma(1+\gamma)(\psi(k)-\log(\bar{c}q(X_1)))\right]\\
&\qquad\qquad-\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\gamma(\psi(k)-\log(\bar{c}q(Y_1)))\right]\\
&\quad=\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[\gamma\log\bar{c}+\gamma\log p(X_1)-\gamma(1+\gamma)\log\bar{c}-\gamma(1+\gamma)\log q(X_1)+\gamma^2\psi(k)\right]\\
&\qquad\qquad-\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\gamma\psi(k)-\gamma\log\bar{c}-\gamma\log q(Y_1)\right]\\
&\quad=\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[\log p^\gamma(X_1)-(1+\gamma)\log q^\gamma(X_1)\right]+\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\log q^\gamma(Y_1)\right]\\
&\qquad-\frac{\gamma}{(1+\gamma)}\log\bar{c}+\frac{\gamma}{(1+\gamma)}\psi(k)+\frac{\gamma}{(1+\gamma)}\log\bar{c}-\frac{\gamma}{(1+\gamma)}\psi(k)\\
&\quad=\frac{1}{\gamma(1+\gamma)}\mathbb{E}_{X_1\sim p}\left[\log p^\gamma(X_1)\right]-\frac{1}{\gamma}\mathbb{E}_{X_1\sim p}\left[\log q^\gamma(X_1)\right]+\frac{1}{1+\gamma}\mathbb{E}_{Y_1\sim q}\left[\log q^\gamma(Y_1)\right].
\end{aligned}$$

Therefore, Eq. (14) is asymptotically unbiased. The claim is proved. □

If $-k<\kappa:=\gamma<0$, the asymptotic unbiasedness also holds.

Theorem 9 (Asymptotic unbiasedness)**.** *Let $-k<\kappa:=\gamma<0$. Suppose that Assumptions 2-4 are satisfied. Let $\exists\delta_0$ s.t. $\forall\delta\in(0,\delta_0)$, $\int_{\mathcal{M}}H(x,p,\delta,1)q(x)\mathrm{d}x<\infty$, and that p is bounded from above. Let $\mathrm{supp}(p)\supseteq\mathrm{supp}(q)$. Then, the estimator in Eq. (7) is asymptotically unbiased.*

Proof. This theorem can be shown in the same way as Theorem 2. □

By combining the results of Theorem 8 and 9, Theorem 2 can be shown.

D.2 Proofs of Theorem 3

Proof. Recalling the default formulation of γ-divergence estimator in Eq. (7), we can see

$$
\begin{aligned}
&\widehat{D}_\gamma(X^n\|Y^m)\\
&= \frac{1}{\gamma(1+\gamma)}\log\left(\frac{1}{n}\sum_{i=1}^n \hat{p}_k^\gamma(X_i)\right) - \frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^n \hat{q}_k^\gamma(X_i)\right) + \frac{1}{1+\gamma}\log\left(\frac{1}{m}\sum_{j=1}^m \hat{q}_k^\gamma(Y_j)\right)\\
&= \frac{1}{\gamma(1+\gamma)}\log\left(\frac{1}{n}\sum_{i=1}^n p^\gamma(X_i)\right) - \frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^n q^\gamma(X_i)\right) + \frac{1}{1+\gamma}\log\left(\frac{1}{m}\sum_{j=1}^m q_k^\gamma(Y_j)\right)\\
&\quad - \frac{1}{\gamma(1+\gamma)}\log\left(\frac{1}{n}\sum_{i=1}^n p^\gamma(X_i)\right) + \frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^n q^\gamma(X_i)\right) - \frac{1}{1+\gamma}\log\left(\frac{1}{m}\sum_{j=1}^m q_k^\gamma(Y_j)\right)\\
&\quad + \frac{1}{\gamma(1+\gamma)}\log\left(\frac{1}{n}\sum_{i=1}^n \hat{p}_k^\gamma(X_i)\right) - \frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^n \hat{q}_k^\gamma(X_i)\right) + \frac{1}{1+\gamma}\log\left(\frac{1}{m}\sum_{j=1}^m \hat{q}_k^\gamma(Y_j)\right)\\
&= \frac{1}{\gamma(1+\gamma)}\log\left(\frac{1}{n}\sum_{i=1}^n p^\gamma(X_i)\right) - \frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^n q^\gamma(X_i)\right) + \frac{1}{1+\gamma}\log\left(\frac{1}{m}\sum_{j=1}^m q_k^\gamma(Y_j)\right)\\
&\quad + \frac{1}{\gamma(1+\gamma)}\log\frac{\frac{1}{n}\sum_{i=1}^n \hat{p}_k^\gamma(X_i)}{\frac{1}{n}\sum_{i=1}^n p^\gamma(X_i)} - \frac{1}{\gamma}\log\frac{\frac{1}{n}\sum_{i=1}^n \hat{q}_k^\gamma(X_i)}{\frac{1}{n}\sum_{i=1}^n q^\gamma(X_i)} + \frac{1}{1+\gamma}\log\frac{\frac{1}{m}\sum_{j=1}^m \hat{q}_k^\gamma(Y_j)}{\frac{1}{m}\sum_{j=1}^m q_k^\gamma(Y_j)}.
\end{aligned}
$$

The first, second and third terms converge to the expectation of $p^\gamma(x)$, $q^\gamma(x)$ and $q^\gamma(y)$, and therefore these terms converge to $D_\gamma(p\|q)$ almost surely because the sum of almost surely convergence terms also converges almost surely [34].

(i) According to Lemma 7, $\hat{p}_k^\gamma(x) \to_p p^\gamma(x)$ for almost all of x. In addition, according to the fact that the sum of random variables that converge in probability converges almost surely [34], we obtain

$$
\frac{1}{n}\sum_{i=1}^n \hat{p}_k^\gamma(X_i) \overset{\text{a.s.}}{\to} \mathbb{E}_{p(x)}[p^\gamma(x)].
$$

Therefore,

$$
\frac{1}{\gamma(1+\gamma)}\log\frac{\frac{1}{n}\sum_{i=1}^n \hat{p}_k^\gamma(X_i)}{\frac{1}{n}\sum_{i=1}^n p^\gamma(X_i)} \overset{\text{a.s.}}{\to} \frac{1}{\gamma(1+\gamma)}\log\frac{\mathbb{E}_{p(x)}[p^\gamma(x)]}{\mathbb{E}_{p(x)}[p^\gamma(x)]} = 0.
$$

(ii) According to Lemma 8, $\hat{q}_k^\gamma(x) \to_p q^\gamma(x)$ for almost all of x. In the same way of (i), we obtain

$$
\frac{1}{\gamma(1+\gamma)}\log\frac{\frac{1}{n}\sum_{i=1}^n \hat{q}_k^\gamma(X_i)}{\frac{1}{n}\sum_{i=1}^n q^\gamma(X_i)} \overset{\text{a.s.}}{\to} \frac{1}{\gamma(1+\gamma)}\log\frac{\mathbb{E}_{p(x)}[q^\gamma(x)]}{\mathbb{E}_{p(x)}[q^\gamma(x)]} = 0.
$$

(iii) According to Lemma 9, we obtain

$$
\frac{1}{1+\gamma}\log\frac{\frac{1}{m}\sum_{j=1}^m \hat{q}_k^\gamma(Y_i)}{\frac{1}{m}\sum_{j=1}^m q^\gamma(Y_i)} \overset{\text{a.s.}}{\to} \frac{1}{1+\gamma}\log\frac{\mathbb{E}_{q(y)}[q^\gamma(y)]}{\mathbb{E}_{q(y)}[q^\gamma(y)]} = 0
$$

From (i) to (iii), we obtain

$$
\widehat{D}_\gamma(X^n\|Y^m) \overset{\text{a.s.}}{\to} D_\gamma(p\|q),
$$

and the claim is proved. □

E Detail of Data Discrepancy Measure

In this section, we introduce data discrepancy measures.

E.1 Distance between Summary Statistics

An ABC often uses the distance between the summary statistics: $S(X^n)$ and $S(Y^m)$ as the discrepancy measure. If we use the Euclidian distance, the discrepancy measure can be expressed as

$$D_S(X^n, Y^m) = \|S(X^n) - S(Y^m)\|.$$

However, it is difficult to choose the summary statistic S for each task properly. One way to bypass this difficulty is the Bayesian indirect inference method [24, 25].

Bayesian Indirect method The aim of the Bayesian indirect method is to construct the summary statistics from an auxiliary model: $\{p_A(x|\phi) : \phi \in \Phi\}$ (see Drovandi et al. [25] for general review). Drovandi and Pettitt [24] proposed to use the maximum likelihood estimation (MLE) of the auxiliary model as summary statistics. Formally,

$$S(Y^m) = \hat{\phi}(Y^m) = \operatorname*{argmax}_{\phi \in \Phi} \prod_{j=1}^{m} p_A(Y_j|\phi).$$

We set $p_A(x|\phi)$ as d-dimensional Gaussian with parameter ϕ in our experiments. In this setting, the summary statistics are merely the sample mean and covariance of Y^m. Furthermore, we adopted the auxiliary likelihood (AL) proposed by Gleim and Pigorsch [31] as a data discrepancy:

$$D_{\mathrm{AL}}(X^n, Y^m) = \frac{1}{m} \log p_A(Y^m|\hat{\phi}(Y^m)) - \frac{1}{m} \log p_A(Y^m|\hat{\phi}(X^n)).$$

Outlier-Robust Function as Summary Statistics Ruli et al. [68] proposed the robust M-estimator Ψ as the summary statistics to deal with the outliers in the observed data. For example, we can use the Huber function as

$$\Psi(x - \mu) = \begin{cases} -c \ \ (x - \mu < -c), \\ x - \mu \ \ (|x - \mu| \leq 0), \\ c \ \ (x - \mu > c), \end{cases}$$

where μ is a mean of x. We adopted this function as the summary statistics and applied for the AL in the above. Formally,

$$D_{\mathrm{ALH}}(X^n, Y^m) = \frac{1}{m} \log p_A(S_\Phi(Y^m)|\hat{\phi}(S_\Phi(Y^m))) - \frac{1}{m} \log p_A(S_\Phi(Y^m)|\hat{\phi}(S_\Phi(X^n))).$$

Further, we set $c_1 = 1.345$ for mean and $c_2 = 2.07$ for covariance (see Huber et al. [39]).

E.2 Maximum Mean Discrepancy (MMD) based Approach

MMD method Smola et al. [73] and Berlinet and Thomas-Agnan [6] defined the kernel embedding for a probability distribution $g(x)$ as

$$\mu_g = \int k(\cdot, x) g(x) \mathrm{d}x,$$

where k is a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Therefore, μ_g is an element in the reproducing kernel Hilbert space (RKHS): $\mathcal{H}$.

The maximum mean discrepancy (MMD) [33] between the probability distributions g_0 and g_1 is the distance between the kernel embeddings μ_{g_0} and μ_{g_1} in RKHS $\mathcal{H}$, defined as

$$\mathrm{MMD}^2(g_0, g_1) = \|\mu_{g_0} - \mu_{g_1}\|_{\mathcal{H}}^2.$$

Park et al. [60] applied an unbiased estimator of $\mathrm{MMD}^2(p_{\theta^*}, q_\theta)$ as the data discrepancy in ABC. The squared estimator of MMD is defined as

$$D_{\mathrm{MM}}^2(X^n, Y^m) = \frac{\sum_{1 \leq i \neq j \leq n} k(X_i, X_j)}{n(n-1)} + \frac{\sum_{1 \leq i \neq j \leq m} k(Y_i, Y_j)}{m(m-1)} - \frac{2 \sum_{i=1}^{n} \sum_{j=1}^{m} k(X_i, Y_j)}{nm}. \tag{15}$$

In the same way of Park et al. [60] and Jiang et al. [42], we chose a Gaussian kernel with the bandwidth being the median of $\{\|X_i - X_j : 1 \leq i \neq j \leq n\|\}$ in our experiments. Then, the time cost of D_{MM} is $\mathcal{O}((n+m)^2)$ which is caused to compute the $(n+m) \times (n+m)$ pairwise distance matrix.

Median-of-mean to Kernel (MONK) method Lerasle et al. [49] proposed the outlier-robust MMD estimator computed by using the median-of-mean (MON) estimator. MON estimators are expected to enjoy the outlier-robustness thanks to the median step.

For any mapping function $h : \mathcal{X} \mapsto \mathbb{R}$ and any non-empty subset $S \subseteq \{1, 2, \ldots, n\}$, denote by $\mathbb{P}_S = |S|^{-1} \sum_{i \in S} \delta_{X_i}$ the empirical measure associated to the subset x_S and $\mathbb{P}_S h = |S|^{-1} \sum_{i \in S} h(X_i)$. For simplification, we express $\mu_S = \mu_{\mathbb{P}_S}$. Let n is divisible by $Q \in \mathbb{Z}^+$ and let $(S_q)_{q \in Q}$ denote a partition of $\{1, 2, \ldots, n\}$ into subsets with the same cardinality $|S_q| = N/Q$. We also mention that q is different from the distribution of Y^m with parameter θ defined as q_θ. Then, the MON is defined as

$$\mathrm{MON}_Q[h] = \mathrm{med}_q\{\mathbb{P}_{S_q}, h\} = \mathrm{med}_q\{k(h, \mu_{S_q})\},$$

where $h \in \mathcal{H}$ in the second equality is a consequence of the mean-reproducing property of $\mu_{\mathbb{P}}$. When we choose $Q = 1$, the MON estimator is equal to the classical mean as $\mathrm{MON}_1 = n^{-1} \sum_{i=1}^n h(X_i)$.

Lerasle et al. [49] defined the minimax MON-based estimator associated with Kernel k (MONK) as

$$\hat{\mu}_{\mathbb{P},Q} = \hat{\mu}_{\mathbb{P},Q}(X^n) \in \operatorname*{argmin}_{f \in \mathcal{H}} \sup_{g \in \mathcal{H}} \tilde{J}(f, g),$$

where for all $f, g \in \mathcal{H}$

$$\tilde{J}(f, g) = \mathrm{MON}_Q\left[x \mapsto \|f - k(\cdot, x)\|_{\mathcal{H}}^2 - \|g - k(\cdot, x)\|_{\mathcal{H}}^2\right].$$

When we choose $Q = 1$, we obtain the classical empirical mean based estimator as $\mu_{\mathbb{P},1} = n^{-1} \sum_{i=1}^n k(\cdot, X_i)$.

The MON-based MMD estimator on $X^n \sim g_0$ and $Y^m \sim g_1$ is defined as

$$\widehat{\mathrm{MMD}}_Q(g_0, g_1) = \sup_f \mathrm{med}_{q \in Q}\{k(f, \mu_{S_{q,g_0}} - \mu_{S_{q,g_1}})\},$$

where $\mu_{S_{q,g_0}} = \mu_{\mathbb{P}_{S_q, X_i}}$ and $\mu_{S_{q,g_1}} = \mu_{\mathbb{P}_{S_q, Y_i}}$. Again, when we choose $Q = 1$, this is equal to the classical V-statistic-based MMD estimator [33] in the previous paragraph. The (unbiased) U-statistic based MONK estimator also could be obtained in the same way as Eq. (15) (see Lerasle et al. [49]).

The MONK estimator has the time cost $\mathcal{O}(n^3)$ and therefore $\mathcal{O}((n+m)^3)$ when we use it as the data discrepancy in ABC. It is too expensive to apply for a large sample size. Leonenko et al. [48] also proposed the faster algorithm to compute the MONK estimator, called MONK BCD-Fast, which has $\mathcal{O}((n+m)^3/Q^2)$ time cost. We adopted this algorithm in our experiments and set $Q = 11$. Furthermore, we adopted the RBF kernel with bandwidth $\sigma = 1$, which is also used in Lerasle et al. [49].

E.3 Wasserstein Distance

Jiang et al. [42] mentioned that the estimator of the q-Wasserstein distance could be used as a data discrepancy for ABC. Let ψ be a distance on $\mathcal{X} \subseteq \mathbb{R}^d$. The q-Wasserstein distance between g_0 and g_1 is defined as

$$\mathcal{W}_q(g_0, g_1) = \left[\inf_{\tau \in \Gamma(g_0, g_1)} \int_{\mathcal{X} \times \mathcal{X}} \psi(x, y)^q \mathrm{d}\tau(x, y)\right]^{1/q},$$

where $\Gamma(g_0, g_1)$ is the set of all joint distribution $\tau(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that τ has marginals g_0 and g_1. We also mention that q is different from the distribution of Y^m with parameter θ defined as q_θ. When we set $q = 2$ and ψ be the Euclidean distance, the data discrepancy based on the q-Wasserstein distance is given by

$$D_{\mathcal{W}2}(X^n, Y^m) = \min_\tau \left[\sum_{i=1}^n \sum_{j=1}^m \tau_{ij} \|X_i - Y_j\|^2\right]^{1/2} \text{ s.t. } \tau \mathbf{1}_m = \mathbf{1}_n, \tau^\top \mathbf{1}_n = \mathbf{1}_m, 0 \leq \tau_{ij} \leq 1,$$

where $\tau = \{\tau_{ij}; 1 \leq i \leq n, 1 \leq j \leq m\}$ is a $n \times m$ matrix and $\mathbf{1}_n, \mathbf{1}_m$ are vectors filled with n pieces or m pieces of 1, respectively.

When we want to solve the optimization problem of $D_{\mathcal{W}2}$ exactly on multivariate distributions ($d > 1$), we have the time cost $\mathcal{O}((n+m)^3 \log(n+m))$ [14]. It is a high cost significantly and therefore Cuturi [21] and Cuturi and Doucet [22] proposed approximate optimization algorithms which reduce the time cost to $\mathcal{O}((n+m)^2$. We used this algorithm in our experiments. For univariate distributions, i.e., $d = 1$, if $n = m$ and $\psi(x, y) = |x - y|$, the q-Wasserstein distance has an explicit form as

$$\left(\frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|^q\right)^{1/q},$$

and in this special case, the time cost is $\mathcal{O}(n \log n)$ [42].

E.4 Classification Accuracy Method

The classification accuracy discrepancy (CAD) has been proposed by Gutmann et al. [35]. The idea of this method is on the basis of the belief that it is easier to distinguish the observed data X^n and the synthetic data Y^m when θ is different significantly to the true parameter θ^* than to do so when θ resembles θ^*.

The CAD sets the labels of $\{X_i\}_{i=1}^n$ as class 0 and $\{Y_j\}_{j=1}^m$ as class 1 at first. In short, it yields an augmented data set as

$$\mathcal{D} = \{(X_1, 0), (X_2, 0), \ldots, (X_n, 0), (Y_1, 1), (Y_2, 1), \ldots, (Y_m, 1)\},$$

and then trains a prediction classifier $h : x \mapsto \{0, 1\}$.

Gutmann et al. [35] defined classifiability between X^n and Y^m as the K-fold cross-validation classification accuracy and proposed to use it for ABC as a data discrepancy. The data discrepancy based on the CAD is defined as

$$D_{\mathrm{CAD}}(X^n, Y^m) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{D}_k|}\left[\sum_{i:(X_i,0)\in\mathcal{D}_k}(1 - \hat{h}_k(X_i)) + \sum_{j:(Y_j,1)\in\mathcal{D}_k}\hat{h}_k(Y_j)\right],$$

where $\mathcal{D}_k$ is the k-fold subset of $\mathcal{D}$, $|\mathcal{D}_k|$ is the size of $\mathcal{D}_k$ and $\hat{h}_k$ is the trained predictor on the data set $\mathcal{D} \setminus \mathcal{D}_k$.

The discrepancy via linear Discriminant Analysis (LDA) is computationally cheaper than other classifiers, which is $\mathcal{O}(n+m)$; however, Gutmann et al. [35] explicitly noted that LDA does not work for some models, e.g., the moving average models (see Figure 2 in Gutmann et al. [35]). Therefore, in our experiments, we set $K = 5$ and h to be the logistic regression with L_1 regularization and the gradient boosting classifier.

E.5 KL-divergence estimation via k-NN

KL-divergence between the density functions p and q is defined as

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{M}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{16}$$

where $\mathcal{M}$ is a support of p. It indicates zero if and only if $p = q$ for almost everywhere. Pérez-Cruz [61] proposed to estimate the density firstly by using k-NN density estimation and plug these estimators into Eq. (16). Given i.i.d. samples, X^n and Y^m, we can estimate $D_{\mathrm{KL}}(p\|q)$ by using the k-NN density estimator expressed in Eqs. (4) and (5) as follows:

$$\widehat{D}_{\mathrm{KL}}(p\|q) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{p}_k(X_i)}{\hat{q}_k(X_i)} = \frac{d}{n}\sum_{i=1}^{n}\log\frac{\rho_k(i)}{\nu_k(i)} + \log\frac{m}{n-1}. \tag{17}$$

This estimator enjoys asymptotical properties such as asymptotical unbiasedness, L_2-consistency and almost sure convergence ([61, 79]). If we use 1-NN density estimation, the above estimator (17) can be expressed as

$$\widehat{D}_{\mathrm{KL}}(p\|q) = \frac{d}{n}\sum_{i=1}^{n}\log\frac{\min_j \|X_i - Y_j\|_2}{\min_{j\neq i}^{n}\|X_i - X_j\|_2} + \log\frac{m}{n-1}, \tag{18}$$

where $\|\cdot\|_2$ means l_2-norm.

Jiang et al. [42] proposed to use this estimator (18) as the data discrepancy in the ABC framework. As ABC involves $2n$ operations of nearest neighbor search, Jiang et al. [42] also proposed to use KD trees [5, 52]. The time cost thus is $\mathcal{O}((n \vee m) \log(n \vee m))$ on average, where we denote $\max\{a, b\}$ as $a \vee b$.

According to Theorem 1 in [42], the asymptotic ABC posterior is a restriction of the prior π on the region $\{\theta \in \Theta : D(g_{\theta^*} \| g_\theta) < \epsilon\}$.

Theorem 10 (Theorem 1 in [42])**.** *Let the data discrepancy measure $D(X^n, Y^m)$ in Algorithm 1 converges to some real number $D(p_{\theta^*}, q_\theta)$ almost surely as the data size $n \to \infty$, $m/n \to \alpha > 0$. Then, the ABC posterior distribution $\pi(\theta | X^n; D, \epsilon)$ defined by (1) converges to $\pi(\theta | D(p_{\theta^*}, q_\theta) < \epsilon)$ for any θ. That is,*

$$\lim_{n \to \infty} \pi(\theta | X^n; D, \epsilon) = \pi(\theta | D(p_{\theta^*}, q_\theta) < \epsilon) \propto \pi(\theta) \mathbb{1}\{D(p_{\theta^*}, q_\theta) < \epsilon\}.$$

Jiang et al. [42] also showed the behavior of the ABC posterior based on KL-divergence estimator.

Corollary 2 (Corollary 1 in [42])**.** *Let the data size $n \to \infty$, $m/n \to \alpha > 0$. Let us define $\pi(\theta | D_{\mathrm{KL}}(p_{\theta^*}, q_\theta) < \epsilon)$ as the posterior under $D_{\mathrm{KL}}(p_{\theta^*}, q_\theta) < \epsilon)$. If Algorithm 1 uses $\widehat{D}_{\mathrm{KL}}$ defined by Eq. (18) as the data discrepancy measure, then the ABC posterior distribution $\pi(\theta | X^n; \widehat{D}_{\mathrm{KL}}, \epsilon)$ defined by Eq. (1) converges to $\pi(\theta | D_{\mathrm{KL}}(p_{\theta^*}, q_\theta) < \epsilon)$ for any θ. That is,*

$$\lim_{n \to \infty} \pi(\theta | X^n; D_{\mathrm{KL}}, \epsilon) = \pi(\theta | D_{\mathrm{KL}}(p_{\theta^*}, q_\theta) < \epsilon) \propto \pi(\theta) \mathbb{1}\{D_{\mathrm{KL}}(p_{\theta^*}, q_\theta) < \epsilon\}.$$

It is known that the maximum likelihood estimator minimizes the KL-divergence between the empirical distribution of p_{θ^*} and q_θ. ABC with D_{KL} shares the same idea to find θ with small KL-divergence.

F Details of Experimental Settings

In this section, we summarize the details of the model settings we used in experiments.

F.1 Gaussian Mixture Model (GM)

The univariate Gaussian mixture model is the most fundamental benchmark model in ABC literature [71, 81, 42]. We adopted a bivariate Gaussian mixture model with the true parameters $p^* = 0.3$, $\mu_0^* = (0.7, 0.7)$ and $\mu_1^* = (-0.7, -0.7)$, where p^* means the mixture ratio and μ_0^*, μ_1^* are sub-population means of Gaussian distribution. Therefore, the set of the true parameter is $\theta^* = (p^*, \mu_0^*, \mu_1^*)$ The generative process of data is as follows:

$$\begin{aligned} &Z \sim \mathrm{Bernoulli}(p), \\ &[X | Z = 0] \sim \mathcal{N}(\mu_0, [0.5, -0.3; -0.3, 0.5]), \\ &[X | Z = 1] \sim \mathcal{N}(\mu_1, [0.25, 0; 0, 0.25]). \end{aligned}$$

We set the $n = 500$ observed data and the prior on the unknown parameter $\theta = (p, \mu_0, \mu_1)$ as $p \sim \mathrm{Uniform}[0, 1]$ and $\mu_0, \mu_1 \sim \mathrm{Uniform}[-1, 1]^2$.

F.2 *M/G/*1-queueing Model (MG1)

Queuing models are usually easy to simulate from; however, it is difficult to conduct inference because these have no intractable likelihoods. The $M/G/1$-queuing model well has been studied in ABC context [13, 27, 42]. The M, G and 1 means *Memoryless* which follows some arrival process, *General holding time distribution* and *single server*, respectively. In this model, the service times follows $\mathrm{Uniform}[\theta_1, \theta_2]$ and the inter arrival times are exponentially distributed with rate θ_3. Each datum is a 5-dimensional vector consisting of the first five inter departure times $x = (x_1, x_2, x_3, x_4, x_5)$ after the queue starts from empty [42].

We adopted this model with the true parameters $\theta^* = (1, 5, 0.2)$. We set the $n = 500$ observed data and the prior on the unknown parameter $\theta = (\theta_1, \theta_2, \theta_3)$ as $\theta_1 \sim \mathrm{Uniform}[0, 10]$, $\theta_2 - \theta_1 \sim \mathrm{Uniform}[0, 10]$ and $\theta_3 \sim \mathrm{Uniform}[0, 0.5]$.

F.3 Bivariate Beta Model (BB)

The bivariate beta model was proposed as a model with 8 parameters $\theta = (\theta_1, \ldots, \theta_8)$ by Arnold and Ng [2]. The generative process of data is as follows:

$$
\begin{aligned}
U_i &\sim \mathrm{Gamma}(\theta_i, 1)\ (i = 1, \ldots, 8), \\
V_1 &= \frac{U_1 + U_5 + U_7}{U_3 + U_6 + U_8}, \\
V_2 &= \frac{U_2 + U_5 + U_8}{U_4 + U_6 + U_7}, \\
Z_1 &= \frac{V_1}{1 + V_1}, \\
Z_2 &= \frac{V_2}{1 + V_2}.
\end{aligned}
$$

Then, $Z = (Z_1, Z_2)$ follows a bivariate beta distribution. Crackel and Flegal [19] reconsidered as a 5-parameter sub-model by restricting $\theta_3, \theta_4, \theta_5 = 0$. Jiang et al. [42] used the 5-parameter models for ABC experiments and therefore we also adopted this with the true parameter $\theta^* = (3, 2.5, 2, 1.5, 1)$ as a benchmark model.

We set the $n = 500$ observed data and the prior on the unknown parameter $\theta = (\theta_1, \theta_2, \theta_6, \theta_7, \theta_8)$ as $\theta_1, \theta_2, \theta_6, \theta_7, \theta_8 \sim \mathrm{Uniform}[0, 5]^5$.

F.4 Moving-average Model of Order 2 (MA2)

Marin et al. [53] used the moving-average model of order 2 as a benchmark model. We adopted this model with 10-length time series and unobserved noise error term Z_j, which follows Student's t-distribution with 5 degrees of freedom. Therefore, the generative process of data is

$$
Y_j = Y_j + \theta_1 Y_{j-1} + \theta_2 Y_{j-2}\ \ (j = 1, 2, \ldots, 10).
$$

We also assumed this model has the true parameter $\theta^* = (0.6, 0.2)$. We then set the $n = 200$ observed data and the prior on the unknown parameter $\theta = (\theta_1, \theta_2)$ as $\theta_1, \theta_2 \sim \mathrm{Uniform}[-2, 2] \times \mathrm{Uniform}[-1, 1]$.

F.5 Multivariate g-and-k Distribution (GK)

The univariate g-and-k distribution is defined by its inverse distribution function as

$$
F^{-1}(x) = A + B\left[1 + c\frac{1 - \exp(-g z_x)}{1 + \exp(-g z_x)}\right](1 + z_x^2)^k z_x,
$$

where z_x is the x-th quantile of the standard normal distribution, and the parameters A, B, g, k are related to location, scale, skewness and kurtosis, respectively. The hyper-parameter c is conventionally chose as $c = 0.8$ [27]. As the inversion transform method can conveniently sample from this distribution by drawing $Z \sim N(0, 1)$ i.i.d. and then transforming them to be g-and-k distributed random variables. Rayner and Macgillivray [66] mentioned that the univariate g-and-k distribution had no analytical form of the density function, and the numerical evaluation of the likelihood function is costly. Therefore, ABC is often used on it [62, 27, 1]. Furthermore, Drovandi and Pettitt [24] and [50] has also considered the multivariate g-and-k distribution.

In our experiments, we set a 5-dimensional g-and-k distribution. The generative steps are as follows:

$$
\begin{aligned}
&\text{Draw: } Z = (Z_1, \ldots, Z_5) \sim \mathcal{N}(0, \Sigma), \\
&\text{Transform: } Z,
\end{aligned}
$$

where Σ is sparse matrix which has $\Sigma_{ii} = 1$ and $\Sigma_{ii} = \rho$ if $|i - j| = 1$ or 0 otherwise. We used the transformation for Z that changes marginally as the univariate g-and-k distribution does. We also adopted this model with the true parameters $\theta^* = (A^*, B^*, g^*, k^*, \rho^*)$, where $A^* = 3$, $B^* = 1$, $g^* = 2$, $k^* = 0.5$ and $\rho^* = -0.3$. We set the $n = 500$ observed data and the prior on the unknown parameter $\theta = (A, B, g, k, \rho)$ as $A, B, g, k \sim \mathrm{Uniform}[0, 4]$ and ρ is sampled from $\mathrm{Uniform}[0, 1]$ and is transformed by $2\sqrt{3}(\rho - 0.5)/3$.

G Additional Results for Experiments in Section 5

We summarize the additional mean-squared-error (MSE) results for the experiments in Section 5. Furthermore, we report the simulation error results based on the energy distance.

G.1 MSEs for All Parameters

The following table shows the experimental results of MSEs for all parameters in the experiments of Section 5. From these results, our method almost outperforms the other baseline methods, especially when the observed data have heavy contamination.

Table 2: Experimental results of 8 baseline methods for 5 benchmark models on MSE and standard error of all parameters. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores of γ-divergence estimator are picked up from the all of experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	GM	MG1	BB	MA2	GK
AL (Indirect)	0%	0.350 (0.419)	0.940 (0.851)	0.946 (0.412)	0.006 (0.004)	0.155 (0.144)
	10%	0.805 (0.669)	0.556 (0.448)	1.538 (0.251)	1.094 (0.033)	0.870 (0.275)
	20%	0.734 (0.882)	2.888 (1.222)	1.557 (0.229)	1.125 (0.022)	1.374 (0.439)
AL with Huber (Robust Indirect)	0%	0.097 (0.261)	0.734 (1.369)	1.092 (0.456)	0.029 (0.030)	0.199 (0.116)
	10%	0.920 (0.033)	0.370 (0.369)	1.948 (0.140)	1.017 (0.154)	1.066 (0.180)
	20%	1.000 (0.025)	0.836 (0.567)	2.441 (0.700)	2.275 (0.998)	0.872 (0.300)
Classification (L_1 + Logistic)	0%	1.324 (0.088)	4.018 (0.664)	1.076 (0.430)	0.459 (0.410)	1.076 (0.384)
	10%	0.270 (0.242)	6.422 (0.554)	0.680 (0.213)	0.757 (0.138)	1.240 (0.290)
	20%	0.212 (0.250)	8.394 (0.051)	0.709 (0.276)	0.810 (0.112)	1.477 (0.145)
Classification (Boosting)	0%	1.564 (0.075)	0.022 (0.033)	**0.204** (**0.123**)	**0.004** (**0.002**)	**0.074** (**0.076**)
	10%	1.495 (0.218)	0.005 (0.006)	**0.315** (**0.334**)	**0.005** (**0.005**)	0.187 (0.121)
	20%	0.639 (0.686)	0.017 (0.017)	0.346 (0.136)	0.008 (0.007)	0.179 (0.090)
MMD	0%	0.054 (0.105)	0.617 (0.413)	0.326 (0.179)	**0.004** (**0.003**)	0.240 (0.141)
	10%	0.760 (0.500)	0.333 (0.229)	0.366 (0.253)	0.079 (0.024)	**0.165** (**0.094**)
	20%	1.342 (0.339)	1.237 (0.764)	0.823 (0.175)	0.382 (0.054)	0.559 (0.281)
MONK-BCD Fast	0%	0.647 (0.203)	0.113 (0.115)	0.424 (0.205)	0.049 (0.040)	0.362 (0.348)
	10%	0.719 (0.164)	0.114 (0.145)	0.524 (0.243)	0.054 (0.060)	0.326 (0.110)
	20%	0.714 (0.211)	0.160 (0.204)	0.753 (0.403)	0.102 (0.077)	0.282 (0.139)
q-Wasserstein	0%	0.009 (0.011)	0.419 (0.235)	0.317 (0.210)	0.009 (0.006)	0.189 (0.153)
	10%	1.349 (0.311)	0.188 (0.110)	1.880 (0.165)	0.255 (0.051)	0.305 (0.129)
	20%	1.371 (0.296)	3.384 (1.116)	1.967 (0.257)	0.432 (0.104)	0.585 (0.252)
KL-divergence	0%	0.005 (0.003)	0.089 (0.058)	0.406 (0.129)	**0.004** (**0.004**)	0.240 (0.152)
	10%	0.007 (0.004)	0.102 (0.064)	0.346 (0.123)	0.012 (0.006)	0.377 (0.159)
	20%	**0.004** (**0.003**)	0.113 (0.069)	**0.270** (**0.132**)	0.051 (0.027)	0.578 (0.290)
γ-divergence (proposed)	0%	**0.002** (**0.006**)	**0.003** (**0.025**)	0.405 (0.194)	0.005 (0.008)	0.260 (0.140)
	10%	**0.004** (**0.002**)	**0.001** (**0.025**)	0.418 (0.150)	**0.005** (**0.080**)	0.228 (0.140)
	20%	**0.004** (**0.002**)	**0.003** (**0.017**)	0.314 (0.296)	**0.004** (**0.010**)	**0.170** (**0.146**)

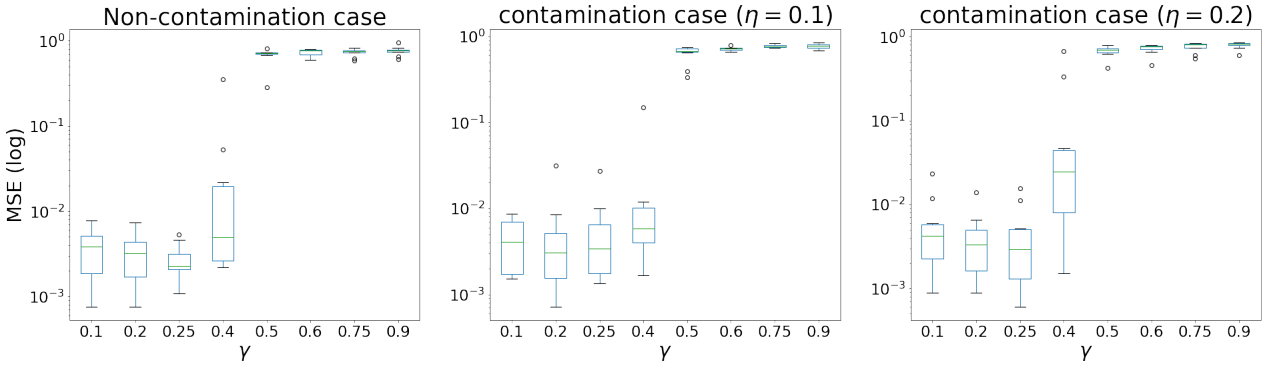

Figure 6: All of the experimental results of our method for the GM model based on MSE.

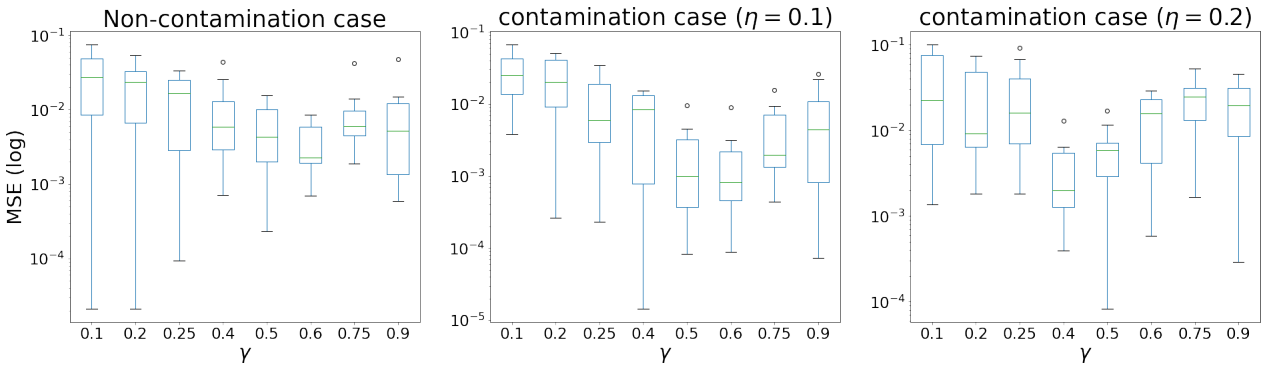

Figure 7: All of the experimental results of our method for the MG1 model based on MSE.

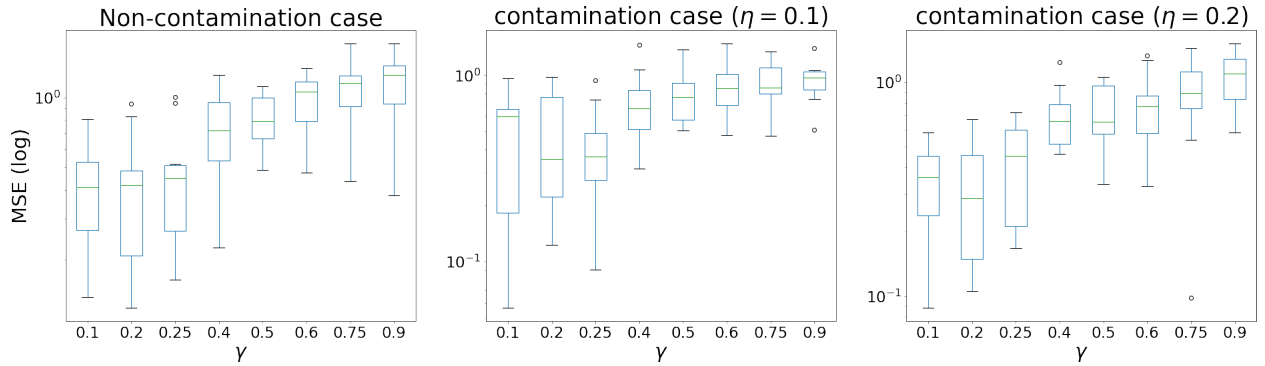

Figure 8: All of the experimental results of our method for the BB model based on MSE.

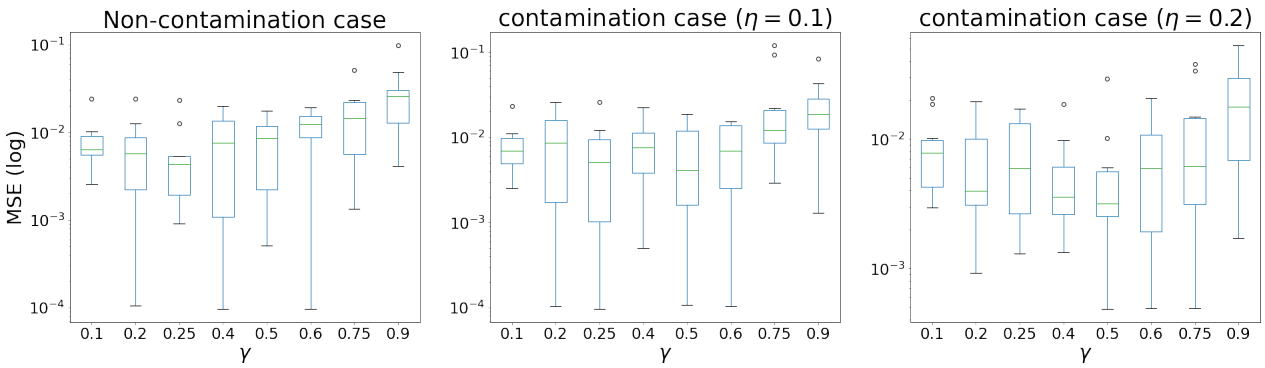

Figure 9: All of the experimental results of our method for the MA2 model based on MSE.

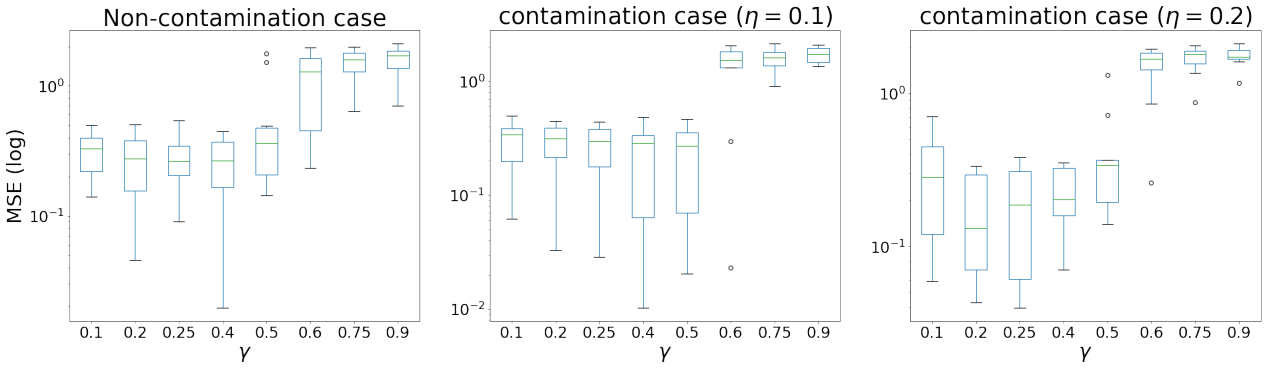

Figure 10: All of the experimental results of our method for the GK model based on MSE.

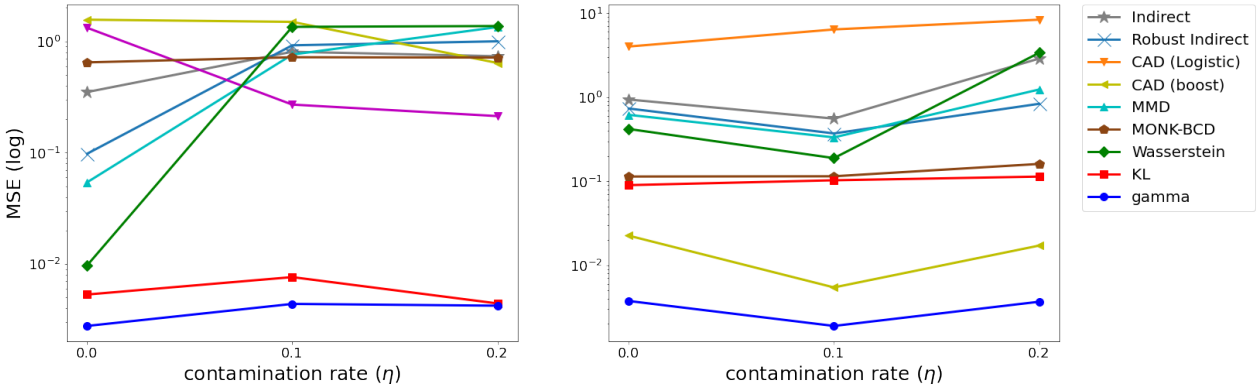

Figure 11: Experimental results for the GM and the MG1 model based on MSE.

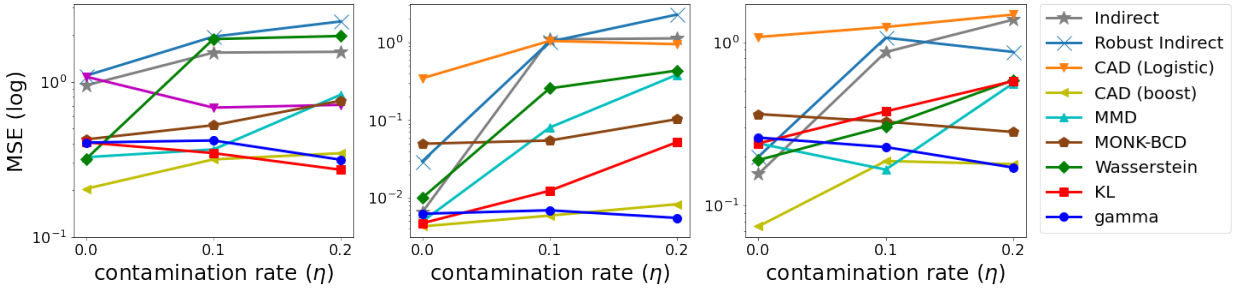

Figure 12: Experimental results for the BB, the MA2, and the GK model based on MSE.

G.2 MSEs for Individual Parameters and Simulation Error

Here, we report the MSE results for each parameter and simulation error in all experiments in Section 5.

G.2.1 Gaussian Mixture Model (GM)

The following table shows the experimental results of MSEs for each parameter in Gaussian mixture experiments. From these results, our method achieves almost a better performance than that of the other baseline methods, especially when the observed data have heavy contamination.

Table 3: Experimental results of 8 baseline methods for Gaussian mixture model on MSE and standard error of each parameter. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores for γ-divergence estimator are picked up from the all of experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	p	$\mu_{0\{0\}}$	$\mu_{0\{1\}}$	$\mu_{1\{0\}}$	$\mu_{1\{1\}}$
AL (Indirect)	0%	0.024 (0.028)	0.868 (1.047)	0.851 (1.027)	0.003 (0.004)	**0.001** (**0.001**)
	10%	0.060 (0.030)	0.898 (1.093)	0.867 (1.046)	0.912 (1.107)	1.290 (1.047)
	20%	0.044 (0.021)	0.879 (1.067)	0.924 (1.129)	0.907 (1.100)	0.915 (1.122)
AL with Huber (Robust Indirect)	0%	0.008 (0.008)	0.252 (0.697)	0.223 (0.608)	0.002 (0.002)	**0.001** (**0.002**)
	10%	0.112 (0.141)	0.007 (0.006)	0.006 (0.004)	2.225 (0.159)	2.252 (0.055)
	20%	0.169 (0.151)	0.025 (0.003)	0.020 (0.006)	2.399 (0.058)	2.387 (0.078)
Classification (L_1 + Logistic)	0%	0.060 (0.023)	2.249 (0.152)	2.224 (0.139)	1.015 (0.333)	1.070 (0.351)
	10%	0.047 (0.024)	0.187 (0.545)	**0.005** (**0.006**)	0.951 (0.945)	0.158 (0.468)
	20%	0.079 (0.046)	0.219 (0.609)	0.014 (0.007)	0.545 (0.829)	0.206 (0.613)
Classification (Boosting)	0%	0.179 (0.021)	2.010 (0.169)	2.026 (0.128)	1.802 (0.273)	1.804 (0.216)
	10%	0.162 (0.034)	2.031 (0.107)	1.952 (0.103)	1.668 (0.581)	1.663 (0.570)
	20%	0.067 (0.039)	0.955 (0.963)	0.959 (0.962)	0.603 (0.924)	0.610 (0.934)
MMD	0%	0.005 (0.004)	0.247 (0.529)	0.013 (0.008)	**0.001** (**0.001**)	0.002 (0.002)
	10%	0.082 (0.061)	1.657 (0.831)	1.402 (0.920)	0.320 (0.637)	0.340 (0.675)
	20%	0.114 (0.041)	2.141 (0.116)	2.130 (0.138)	1.241 (0.835)	1.084 (0.897)
MONK-Fast	0%	0.009 (0.005)	1.592 (0.577)	1.620 (0.494)	0.007 (0.011)	0.005 (0.008)
	10%	0.013 (0.011)	1.699 (0.227)	1.689 (0.383)	**0.002** (**0.002**)	0.192 (0.562)
	20%	0.032 (0.042)	1.792 (0.268)	1.547 (0.556)	0.193 (0.561)	0.007 (0.006)
q-Wasserstein	0%	**0.001** (**0.001**)	0.023 (0.032)	0.018 (0.029)	**0.001** (**0.001**)	0.003 (0.004)
	10%	0.044 (0.026)	0.978 (0.777)	0.880 (0.859)	2.411 (0.058)	2.430 (0.049)
	20%	0.018 (0.018)	1.004 (0.804)	0.767 (0.654)	2.550 (0.050)	2.518 (0.068)
KL-divergence	0%	0.003 (0.002)	0.010 (0.018)	0.003 (0.003)	0.002 (0.003)	0.007 (0.005)
	10%	0.007 (0.018)	0.010 (0.013)	0.011 (0.010)	0.004 (0.004)	0.004 (0.005)
	20%	**0.004** (**0.003**)	0.004 (0.004)	**0.006** (**0.013**)	0.003 (0.006)	**0.002** (**0.004**)
γ-divergence (proposed)	0%	0.003 (0.002)	**0.006** (**0.007**)	**0.001** (**0.001**)	**0.001** (**0.001**)	**0.001** (**0.001**)
	10%	**0.002** (**0.004**)	**0.006** (**0.007**)	0.007 (0.008)	**0.002** (**0.002**)	**0.003** (**0.004**)
	20%	**0.001** (**0.001**)	0.005 (0.006)	0.009 (0.018)	**0.002** (**0.002**)	**0.002** (**0.003**)

G.2.2 *M/G/*1-queueing Model (MG1)

The following table shows the experimental results of MSEs for each parameter in *M/G/*1-queueing Model experiments. From these results, our method achieves almost a better performance than that of the other baseline methods, especially when the observed data have heavy contamination.

Table 4: Experimental results of 8 baseline methods for $M/G/1$-queueing model on MSE and standard error of each parameter. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores for γ-divergence estimator are picked up from the all of experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	θ_1	θ_2	θ_3
AL (Indirect)	0%	0.083 (0.069)	2.737 (2.547)	**0.0001** (**0.0002**)
	10%	1.008 (0.749)	0.660 (0.845)	0.0009 (0.0008)
	20%	4.804 (3.593)	3.859 (2.911)	0.003 (0.001)
AL with Huber (Robust Indirect)	0%	0.202 (0.242)	2.001 (4.142)	**0.0001** (**0.0002**)
	10%	0.998 (1.802)	0.113 (0.118)	0.0007 (0.0003)
	20%	1.339 (1.221)	1.167 (1.119)	0.001 (0.0007)
Classification (L_1 + Logistic)	0%	0.078 (0.082)	11.961 (1.990)	0.016 (0.015)
	10%	0.078 (0.077)	19.180 (1.692)	0.009 (0.010)
	20%	0.308 (0.429)	24.861 (0.414)	0.013 (0.015)
Classification (Boosting)	0%	0.015 (0.022)	0.051 (0.081)	0.0002 (0.0002)
	10%	0.013 (0.018)	0.002 (0.006)	0.0008 (0.0006)
	20%	0.022 (0.038)	0.027 (0.049)	0.0009 (0.0004)
MMD	0%	0.528 (0.567)	1.323 (0.873)	**0.0001** ($>$ **1e** $-$ **6**)
	10%	0.630 (0.667)	0.368 (0.346)	0.0004 (0.0003)
	20%	0.655 (0.732)	3.053 (2.477)	0.002 (0.0004)
MONK-BCD Fast	0%	0.019 (0.024)	0.318 (0.335)	0.003 (0.004)
	10%	0.043 (0.031)	0.298 (0.447)	0.001 (0.002)
	20%	0.182 (0.284)	0.295 (0.579)	0.004 (0.005)
q-Wasserstein	0%	0.174 (0.221)	1.082 (0.772)	**0.0001** ($>$ **1e** $-$ **6**)
	10%	0.175 (0.201)	0.389 (0.321)	**0.00009** ($>$ **1e** $-$ **6**)
	20%	0.393 (0.547)	9.758 (3.177)	0.0008 ($> 1e-6$)
KL-divergence	0%	0.124 (0.186)	0.145 (0.139)	**0.0001** (**0.0002**)
	10%	0.160 (0.132)	0.147 (0.142)	0.0002 (0.0003)
	20%	0.249 (0.185)	0.090 (0.060)	0.001 (0.0008)
γ-divergence	0%	**0.009** (**0.007**)	$>$ **1e** $-$ **5** (**0.0001**)	0.001 (0.002)
	10%	**0.005** (**0.007**)	$>$ **1e** $-$ **5** (**0.0002**)	0.0003 (0.0002)
	20%	**0.008** (**0.010**)	**0.002** (**0.003**)	**0.0002** (**0.0003**)

G.2.3 Bivariate Beta Model (BB)

The following table shows the experimental results of MSEs for each parameter in bivariate-beta model experiments. From these results, our method fails to reduce the effects of outliers. Furthermore, the KL-divergence method works well, even if the observed data are heavily contaminated. We will investigate the reason why this phenomenon occurs as future work. We believe this may be due to the way the contamination of the data occurs.

Table 5: Experimental results of 8 baseline methods for the Bivariate-Beta model on MSE and standard error of each parameter. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores for γ-divergence estimator are picked up from the all of experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	θ_1	θ_2	θ_6	θ_7	θ_8
AL (Indirect)	0%	1.065 (0.538)	1.304 (0.927)	1.365 (1.228)	0.823 (0.617)	0.175 (0.092)
	10%	0.852 (0.645)	1.713 (1.110)	3.066 (0.245)	1.621 (0.086)	0.438 (0.156)
	20%	0.768 (0.419)	2.044 (1.026)	2.908 (0.159)	1.618 (0.185)	0.446 (0.142)
AL with Huber (Robust Indirect)	0%	0.788 (0.466)	1.763 (0.770)	2.038 (1.576)	0.800 (0.799)	0.071 (0.082)
	10%	1.917 (0.546)	3.883 (0.503)	2.279 (0.462)	0.992 (0.266)	0.668 (0.092)
	20%	2.125 (2.138)	1.504 (1.028)	2.028 (1.549)	2.892 (3.149)	3.656 (2.863)
Classification (L_1 + Logistic)	0%	1.135 (0.464)	1.757 (1.118)	1.918 (1.291)	0.412 (0.397)	0.158 (0.235)
	10%	0.833 (0.668)	0.848 (0.692)	0.589 (0.669)	0.687 (0.358)	0.443 (0.143)
	20%	0.715 (0.451)	1.994 (1.213)	**0.141** (**0.149**)	0.312 (0.266)	0.381 (0.148)
Classification (Boosting)	0%	**0.309** (**0.396**)	**0.482** (**0.460**)	0.113 (0.138)	**0.036** (**0.034**)	0.080 (0.049)
	10%	0.622 (1.000)	**0.328** (**0.530**)	0.268 (0.287)	0.315 (0.251)	**0.044** (**0.071**)
	20%	0.571 (0.461)	**0.307** (**0.337**)	0.210 (0.145)	0.546 (0.470)	0.095 (0.129)
MMD	0%	0.756 (0.593)	0.668 (0.370)	**0.085** (**0.094**)	0.059 (0.073)	0.061 (0.064)
	10%	0.653 (0.984)	0.458 (0.527)	0.245 (0.267)	0.391 (0.247)	0.081 (0.100)
	20%	0.774 (0.581)	0.980 (0.784)	1.320 (0.517)	0.796 (0.431)	0.246 (0.163)
MONK-BCD Fast	0%	0.729 (0.365)	0.564 (0.611)	0.538 (0.896)	0.220 (0.128)	0.071 (0.109)
	10%	0.792 (0.638)	0.931 (0.916)	0.678 (0.898)	**0.138** (**0.146**)	0.079 (0.084)
	20%	0.851 (0.709)	1.270 (0.950)	1.189 (1.080)	0.359 (0.802)	0.096 (0.073)
q-Wasserstein	0%	0.373 (0.414)	0.635 (0.622)	0.379 (0.331)	0.128 (0.116)	0.070 (0.106)
	10%	1.663 (0.553)	3.042 (0.736)	2.774 (0.320)	1.364 (0.206)	0.559 (0.139)
	20%	1.871 (0.322)	3.255 (1.137)	2.688 (0.322)	1.392 (0.127)	0.629 (0.079)
KL-divergence	0%	0.794 (0.503)	0.871 (0.408)	0.214 (0.207)	0.065 (0.064)	0.086 (0.090)
	10%	**0.323** (**0.341**)	0.911 (0.734)	**0.238** (**0.323**)	0.205 (0.206)	0.055 (0.090)
	20%	0.568 (0.344)	0.439 (0.383)	0.222 (0.257)	**0.049** (**0.050**)	**0.074** (**0.085**)
γ-divergence	0%	0.639 (0.599)	1.114 (0.632)	0.169 (0.255)	0.051 (0.050)	**0.052** (**0.101**)
	10%	0.897 (0.500)	0.551 (0.581)	0.377 (0.514)	0.162 (0.205)	0.102 (0.133)
	20%	**0.350** (**0.356**)	0.689 (0.552)	0.359 (0.314)	0.096 (0.114)	**0.074** (**0.082**)

G.2.4 Moving-average Model of Order 2 (MA2)

The following table shows the experimental results of MSEs for each parameter in the Moving-average Model of Order 2 experiments. From these results, our method achieves almost a better performance than that of the other baseline methods, especially when the observed data have heavy contamination.

Table 6: Experimental results of 8 baseline methods for the Moving-average model of order 2 on MSE and standard error of each parameter. We performed ABC over 10 trials in 10 different datasets. Lower values are better. The scores for γ-divergence estimator are picked up the best score from all of the experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	θ_1	θ_2
Indirect	0%	0.008 (0.008)	0.004 (0.002)
	10%	1.679 (0.060)	0.508 (0.029)
	20%	1.737 (0.047)	0.514 (0.018)
Robust Indirect	0%	0.035 (0.032)	0.023 (0.030)
	10%	1.563 (0.100)	0.470 (0.270)
	20%	4.251 (1.966)	0.299 (0.189)
Classification (L1 + Logistic)	0%	0.775 (0.772)	0.143 (0.119)
	10%	1.023 (0.127)	0.491 (0.271)
	20%	1.395 (0.134)	0.226 (0.151)
Classification (Boosting)	0%	0.004 (0.002)	0.004 (0.003)
	10%	**0.004** (**0.005**)	0.007 (0.008)
	20%	**0.006** (**0.006**)	0.009 (0.015)
MMD	0%	0.006 (0.006)	**0.002** (**0.002**)
	10%	0.121 (0.025)	0.038 (0.036)
	20%	0.547 (0.089)	0.218 (0.063)
MONK-BCD Fast	0%	0.063 (0.064)	0.035 (0.047)
	10%	0.086 (0.110)	0.022 (0.032)
	20%	0.170 (0.151)	0.034 (0.028)
q-Wasserstein	0%	0.017 (0.013)	**0.002** (**0.004**)
	10%	0.153 (0.050)	0.357 (0.088)
	20%	0.423 (0.134)	0.442 (0.102)
KL-divergence	0%	0.004 (0.005)	0.004 (0.004)
	10%	0.007 (0.008)	0.016 (0.007)
	20%	0.045 (0.025)	0.058 (0.034)
γ-divergence (proposed)	0%	**0.003** (**0.005**)	0.008 (0.009)
	10%	0.008 (0.006)	**0.002** (**0.002**)
	20%	**0.006** (**0.005**)	**0.003** (**0.003**)

G.2.5 Multivariate g-and-k Distribution (GK)

The following table shows the experimental results of MSEs for each parameter in Multivariate g-and-k Distribution model experiments. From these results, our method achieves almost a better performance than that of the other baseline methods, especially when the observed data have heavy contamination.

Table 7: Experimental results of 8 baseline methods for the Multivariate g-and-k distribution model on MSE and standard error of each parameter. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores for γ-divergence estimator are picked up the best score from all of the experimental results in Figure 6-10. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	A	B	g	k	ρ
AL (Indirect)	0%	0.080 (0.116)	0.119 (0.116)	0.505 (0.697)	0.063 (0.030)	0.009 (0.013)
	10%	0.294 (0.537)	3.135 (1.220)	0.796 (0.584)	0.088 (0.026)	0.039 (0.007)
	20%	1.209 (1.668)	4.985 (1.142)	0.600 (0.569)	0.039 (0.036)	0.039 (0.005)
AL with Huber (Robust Indirect)	0%	0.052 (0.053)	0.151 (0.172)	0.763 (0.534)	**0.020** (**0.016**)	0.008 (0.010)
	10%	0.150 (0.099)	4.531 (0.899)	0.606 (0.556)	**0.003** (**0.006**)	0.039 (0.003)
	20%	0.248 (0.121)	3.439 (1.686)	0.546 (0.399)	0.110 (0.074)	0.017 (0.004)
Classification (L_1 + Logistic)	0%	0.109 (0.045)	0.340 (0.083)	1.732 (0.739)	2.910 (1.808)	0.290 (0.232)
	10%	0.397 (0.131)	3.217 (1.354)	2.362 (0.215)	0.016 (0.012)	0.209 (0.144)
	20%	0.201 (0.113)	5.401 (0.802)	1.632 (0.591)	0.019 (0.023)	0.130 (0.089)
Classification (Boosting)	0%	0.009 (0.009)	**0.016** (**0.014**)	**0.317** (**0.382**)	**0.020** (**0.022**)	0.008 (0.004)
	10%	0.024 (0.028)	0.285 (0.279)	0.588 (0.511)	0.020 (0.017)	0.017 (0.008)
	20%	0.035 (0.036)	0.377 (0.312)	**0.447** (**0.448**)	0.014 (0.018)	0.020 (0.005)
MMD	0%	0.021 (0.019)	0.130 (0.128)	0.958 (0.711)	0.063 (0.122)	0.026 (0.053)
	10%	0.054 (0.028)	0.190 (0.196)	**0.526** (**0.441**)	0.040 (0.041)	0.018 (0.005)
	20%	0.299 (0.166)	1.729 (1.117)	0.714 (0.483)	0.021 (0.024)	0.033 (0.006)
MONK-BCD Fast	0%	0.009 (0.011)	0.071 (0.146)	0.593 (0.457)	1.063 (1.919)	0.076 (0.143)
	10%	**0.009** (**0.008**)	0.114 (0.160)	1.175 (0.402)	0.316 (0.253)	0.018 (0.025)
	20%	0.016 (0.013)	0.195 (0.494)	0.842 (0.526)	0.222 (0.237)	0.133 (0.162)
q-Wasserstein	0%	0.028 (0.037)	0.025 (0.022)	0.859 (0.769)	0.028 (0.030)	**0.006** (**0.010**)
	10%	0.190 (0.156)	0.502 (0.414)	0.722 (0.675)	0.087 (0.035)	0.023 (0.010)
	20%	0.530 (0.133)	1.474 (0.769)	0.790 (0.827)	0.109 (0.038)	0.022 (0.007)
KL-divergence	0%	**0.007** (**0.006**)	0.042 (0.040)	1.103 (0.752)	0.040 (0.032)	**0.006** (**0.006**)
	10%	0.015 (0.022)	0.149 (0.348)	1.663 (0.545)	0.038 (0.028)	0.018 (0.010)
	20%	0.066 (0.069)	0.993 (1.128)	1.766 (0.732)	0.030 (0.024)	0.033 (0.004)
γ-divergence	0%	0.046 (0.016)	0.065 (0.038)	1.105 (0.591)	0.080 (0.135)	**0.006** (**0.005**)
	10%	0.033 (0.014)	**0.041** (**0.039**)	1.028 (0.757)	0.030 (0.029)	**0.007** (**0.006**)
	20%	**0.008** (**0.008**)	**0.020** (**0.016**)	0.809 (0.575)	**0.007** (**0.008**)	**0.009** (**0.004**)

G.2.6 All of Simulation Error

The following table shows the experimental results of simulation errors (energy distance) in the experiments of Section 5. From these results, our method also outperforms the other baseline methods, especially when the observed data have heavy contamination.

Table 8: Experimental results of 8 baseline methods for 5 benchmark models on simulation error (energy distance) and its standard error. We performed ABC over 10 trials on 10 different datasets. Lower values are better. The scores of γ-divergence estimator are picked up from the all of experimental results in Figure 13-17. Bold-faces indicate the best score per contamination rate.

Discrepancy measure	Outlier	GM	MG1	BB	MA2	GK
AL (Indirect)	0%	0.199 (0.169)	0.270 (0.114)	0.070 (0.030)	0.056 (0.017)	0.260 (0.093)
	10%	0.349 (0.244)	0.408 (0.182)	0.254 (0.034)	0.475 (0.014)	0.460 (0.211)
	20%	0.263 (0.082)	0.906 (0.225)	0.251 (0.026)	0.501 (0.022)	0.724 (0.351)
AL with Huber (Robust Indirect)	0%	0.200 (0.221)	0.223 (0.090)	0.063 (0.019)	0.064 (0.015)	0.300 (0.144)
	10%	0.966 (0.040)	0.345 (0.101)	0.300 (0.018)	0.466 (0.033)	0.470 (0.038)
	20%	1.005 (0.043)	0.509 (0.168)	0.232 (0.061)	0.403 (0.028)	0.689 (0.124)
Classification (L_1 + Logistic)	0%	0.157 (0.027)	0.453 (0.019)	0.066 (0.021)	0.180 (0.065)	0.422 (0.086)
	10%	0.434 (0.148)	0.605 (0.049)	0.132 (0.039)	0.354 (0.028)	0.678 (0.124)
	20%	0.443 (0.125)	0.779 (0.079)	0.142 (0.028)	0.413 (0.040)	0.873 (0.058)
Classification (Boosting)	0%	0.112 (0.006)	0.169 (0.104)	0.042 (0.030)	**0.048** (**0.018**)	**0.138** (**0.048**)
	10%	0.150 (0.062)	0.359 (0.130)	0.049 (0.023)	**0.052** (**0.015**)	0.193 (0.068)
	20%	0.273 (0.100)	0.293 (0.106)	0.052 (0.024)	0.062 (0.024)	0.181 (0.069)
MMD	0%	0.059 (0.042)	0.233 (0.075)	0.055 (0.026)	0.055 (0.024)	0.231 (0.076)
	10%	0.249 (0.120)	0.275 (0.098)	0.048 (0.023)	0.121 (0.026)	0.317 (0.093)
	20%	0.229 (0.119)	0.593 (0.084)	0.070 (0.025)	0.262 (0.031)	0.419 (0.116)
MONK-BCD Fast	0%	0.283 (0.069)	0.312 (0.130)	0.049 (0.030)	0.066 (0.017)	0.393 (0.402)
	10%	0.279 (0.086)	0.266 (0.189)	0.069 (0.041)	0.073 (0.032)	0.295 (0.156)
	20%	0.262 (0.091)	0.362 (0.251)	0.075 (0.040)	0.094 (0.031)	0.245 (0.174)
q-Wasserstein	0%	**0.051** (**0.021**)	0.200 (0.060)	**0.037** (**0.018**)	0.066 (0.027)	0.215 (0.082)
	10%	0.671 (0.289)	0.175 (0.041)	0.298 (0.020)	0.238 (0.035)	0.344 (0.088)
	20%	0.755 (0.213)	0.599 (0.098)	0.307 (0.018)	0.320 (0.051)	0.587 (0.067)
KL-divergence	0%	0.066 (0.024)	0.125 (0.050)	0.055 (0.024)	0.064 (0.016)	0.198 (0.074)
	10%	0.098 (0.079)	0.178 (0.094)	**0.041** (**0.015**)	0.073 (0.027)	0.155 (0.081)
	20%	0.085 (0.044)	0.322 (0.123)	**0.038** (**0.025**)	0.117 (0.022)	0.271 (0.100)
γ-divergence (proposed)	0%	0.060 (0.028)	**0.096** (**0.042**)	0.066 (0.026)	0.049 (0.018)	0.195 (0.030)
	10%	**0.076** (**0.044**)	**0.099** (**0.041**)	0.048 (0.022)	0.055 (0.018)	**0.138** (**0.047**)
	20%	**0.060** (**0.018**)	**0.121** (**0.085**)	0.043 (0.019)	**0.060** (**0.017**)	**0.140** (**0.439**)

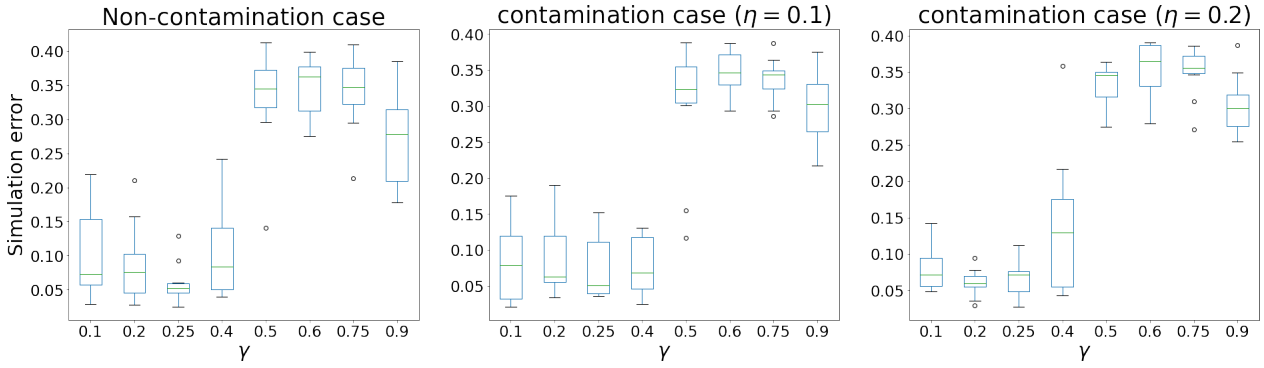

Figure 13: All of the experimental results of our method for the GM model based on simulation error.

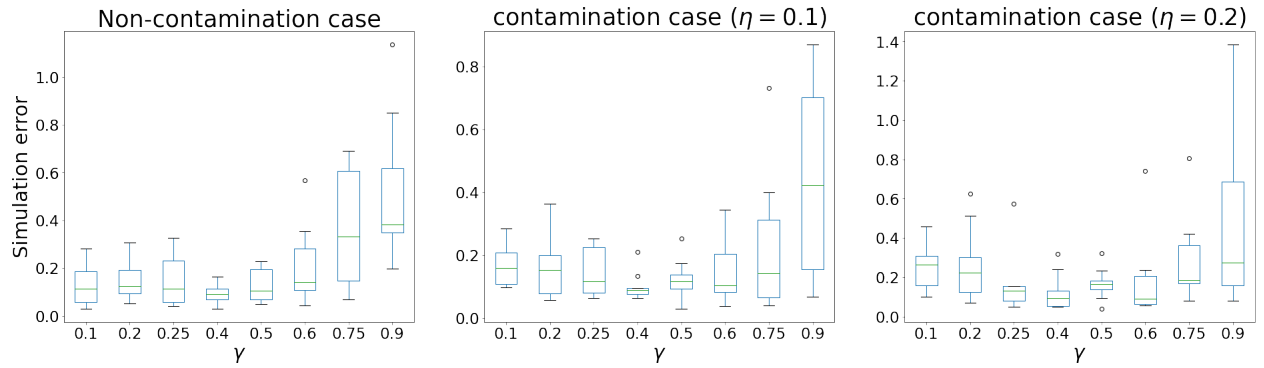

Figure 14: All of the experimental results of our method for the MG1 model based on simulation error.

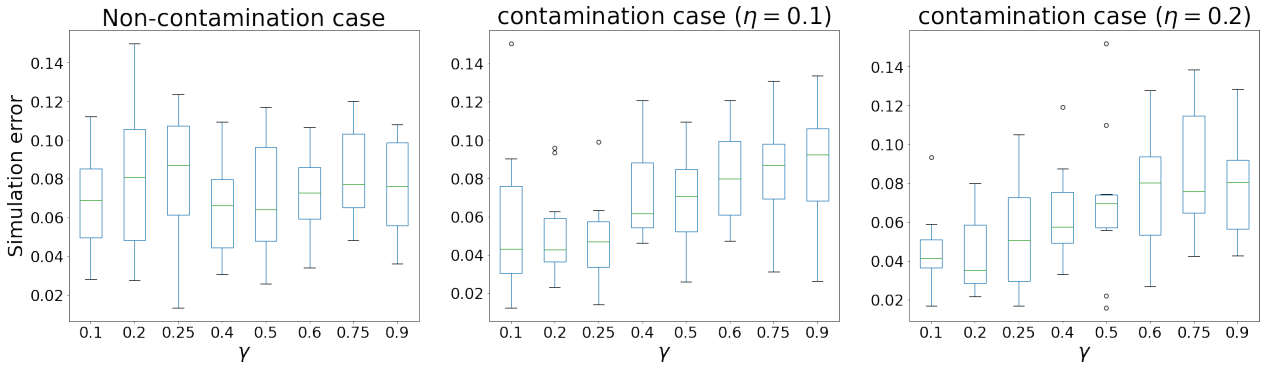

Figure 15: All of the experimental results of our method for the BB model based on simulation error.

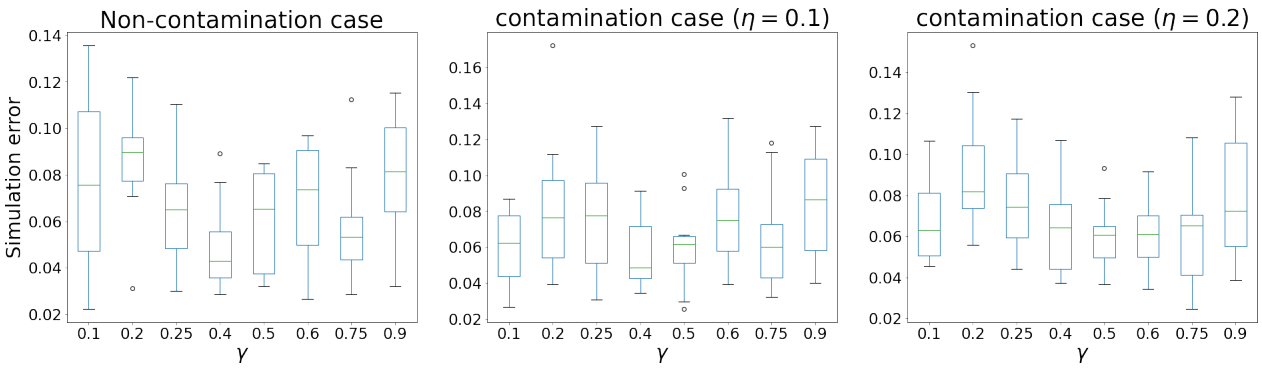

Figure 16: All of the experimental results of our method for the MA2 model based on simulation error.

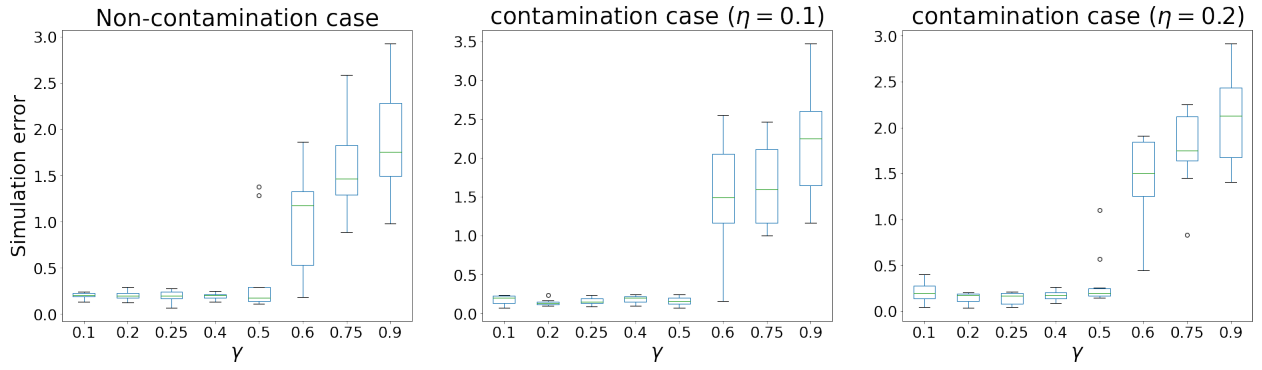

Figure 17: All of the experimental results of our method for the GK model based on simulation error.

G.3 ABC posterior via our method and the second-best method

In this section, we report the ABC posterior distributions of our method for all experiments in Section 5 when $\eta = 0.2$, and compare them with those of the second-best method.

G.3.1 Gaussian Mixture Model (GM)

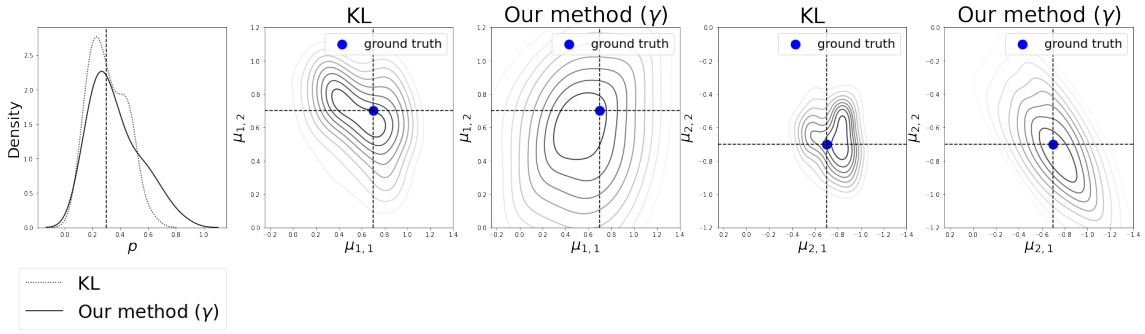

Figure 18: ABC posterior via our method and KL method.

G.3.2 *M*/*G*/1-queueing Model (MG1)

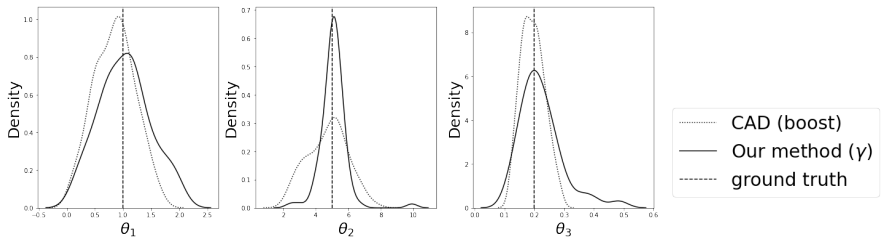

Figure 19: ABC posterior via our method and classification method with boosting.

G.3.3 Bivariate Beta Model (BB)

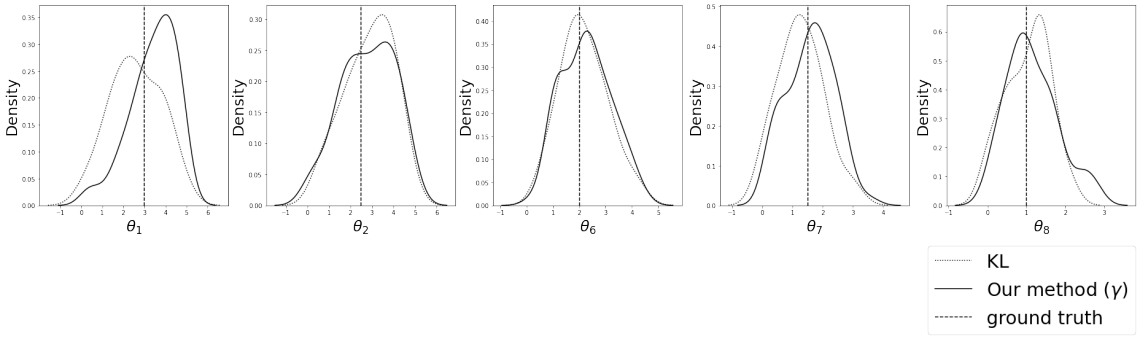

Figure 20: ABC posterior via our method and KL method.

G.3.4 Moving-average Model of Order 2 (MA2)

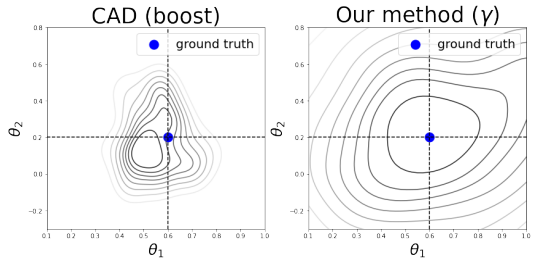

Figure 21: ABC posterior via our method and classification method with boosting.

G.3.5 Multivariate g-and-k Distribution (GK)

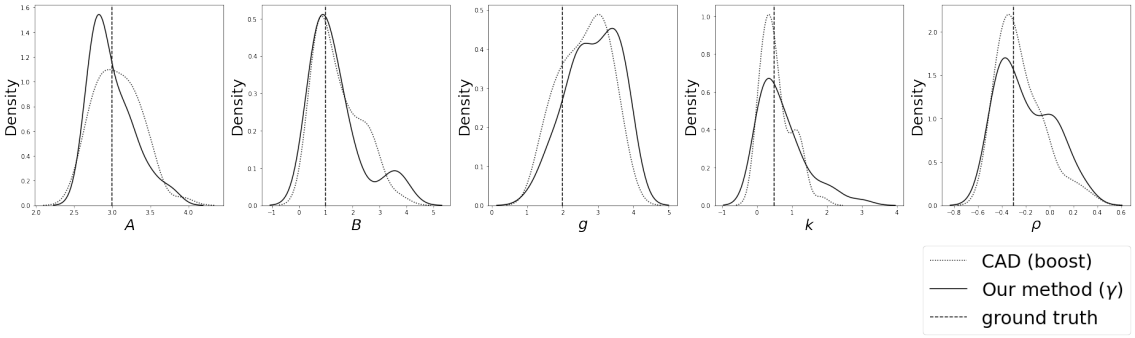

Figure 22: ABC posterior via our method and classification method with boosting.