Supplementary material for "Robust contrastive learning and nonlinear ICA in the presence of outliers"

Hiroaki Sasaki[1], **Takashi Takenouchi**[1,2], **Ricardo Monti**[3], **Aapo Hyvärinen**[4,5]
[1]Future University Hakodate, Hokkaido, Japan [2]RIKEN AIP, Tokyo, Japan [3]University College London, UK
[4]Université Paris-Saclay, Inria, CEA, France [5]University of Helsinki, Finland

Acknowledgement

The authors would like to thank Dr. Hiroshi Morioka for sharing his PCL codes with us. H.S. was supported by JSPS KAKENHI Grant Number 18K18107. T.T. was supported by JSPS KAKENHI Grant Number 20K03753. A.H. was supported by a Fellowship from CIFAR, and by the DATAIA convergence institute as part of the "Programme d'Investissement d'Avenir", (ANR-17-CONV-0003) operated by Inria.

A Proof of Theorem 1

Proof. Let us express the inverse of $\boldsymbol{f}$ in the data generative model (1) by $\boldsymbol{g}$ such that $\boldsymbol{s} = \boldsymbol{g}(\boldsymbol{x})$. The change of variables provides

$$\begin{aligned}
\log p(\boldsymbol{x}|\boldsymbol{u}) &= \log \{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) + \epsilon(\boldsymbol{u})\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})\} + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})| \\
&= \log \{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) + \epsilon(\boldsymbol{u})(\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) - p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}))\} + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})| \\
&= \log\left[p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})\left\{1 + \epsilon(\boldsymbol{u})\left(\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})} - 1\right)\right\}\right] + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})| \\
&= \log p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) + \epsilon(\boldsymbol{u})\left(\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})} - 1\right) + O(\epsilon(\boldsymbol{u})^2) + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|,
\end{aligned}$$

where we applied $\log(1+\epsilon(\boldsymbol{u})z) = \epsilon(\boldsymbol{u})z + O(\epsilon(\boldsymbol{u})^2)$ with a sufficiently small $\epsilon(\boldsymbol{u})$ on the last line. Then, the conditionally exponential family assumption (A2) gives

$$\log p(\boldsymbol{x}|\boldsymbol{u}) = \sum_{j=1}^{d_{\mathrm{x}}} \lambda_j(\boldsymbol{u}) q_j^\star(g_j(\boldsymbol{x})) + \epsilon(\boldsymbol{u})\left(\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})} - 1\right) + O(\epsilon(\boldsymbol{u})^2) + \log|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})| - \log Z(\boldsymbol{\lambda}(\boldsymbol{u})).$$

Contrasting two log-conditional densities of $\boldsymbol{x}$ given $\boldsymbol{u}$ and a fixed point $\boldsymbol{u}_0$ yields

$$\begin{aligned}
\log p(\boldsymbol{x}|\boldsymbol{u}) - \log p(\boldsymbol{x}|\boldsymbol{u}_0) = (\boldsymbol{\lambda}(\boldsymbol{u}) - \boldsymbol{\lambda}(\boldsymbol{u}_0))^\top \boldsymbol{q}^\star(\boldsymbol{g}(\boldsymbol{x})) + \left\{\epsilon(\boldsymbol{u})\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})} - \epsilon(\boldsymbol{u}_0)\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)}\right\} \\
+ O(\epsilon(\boldsymbol{u})^2) + O(\epsilon(\boldsymbol{u}_0)^2) - (\log Z(\boldsymbol{\lambda}(\boldsymbol{u})) - \log Z(\boldsymbol{\lambda}(\boldsymbol{u}_0))), \qquad (21)
\end{aligned}$$

where $\boldsymbol{\lambda}(\boldsymbol{u}) := (\lambda_1(\boldsymbol{u}), \ldots, \lambda_{d_{\mathrm{x}}}(\boldsymbol{u}))^\top$, $\boldsymbol{q}^\star(\boldsymbol{g}(\boldsymbol{x})) := (q_1^\star(g_1(\boldsymbol{x})), \ldots, q_{d_{\mathrm{x}}}^\star(g_{d_{\mathrm{x}}}(\boldsymbol{x})))^\top$, and note that the Jacobian $|\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|$ is cancelled out.

On the other hand, by the universal approximation assumption (A4), we obtain

$$\log p(\boldsymbol{x}|\boldsymbol{u}) - \log p(\boldsymbol{x}|\boldsymbol{u}_0) = (\boldsymbol{w}(\boldsymbol{u}) - \boldsymbol{w}(\boldsymbol{u}_0))^\top \boldsymbol{h}(\boldsymbol{x}) - \log e(\boldsymbol{u}) + \log e(\boldsymbol{u}_0). \qquad (22)$$

Then, equating (21) with (22) provides

$$\bar{\boldsymbol{\lambda}}(\boldsymbol{u})^\top \boldsymbol{q}^\star(\boldsymbol{g}(\boldsymbol{x})) + \epsilon(\boldsymbol{u})\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u})} - \epsilon(\boldsymbol{u}_0)\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)} + O(\epsilon(\boldsymbol{u})^2) + O(\epsilon(\boldsymbol{u}_0)^2) = \bar{\boldsymbol{w}}(\boldsymbol{u})^\top \boldsymbol{h}(\boldsymbol{x}) + \bar{\beta}(\boldsymbol{u}), \quad (23)$$

where $\bar{\boldsymbol{w}}(\boldsymbol{u}) := \boldsymbol{w}(\boldsymbol{u}) - \boldsymbol{w}(\boldsymbol{u}_0)$, $\bar{\boldsymbol{\lambda}}(\boldsymbol{u}) := \boldsymbol{\lambda}(\boldsymbol{u}) - \boldsymbol{\lambda}(\boldsymbol{u}_0)$, and $\bar{\beta}(\boldsymbol{u}) := \log e(\boldsymbol{u}) - \log e(\boldsymbol{u}_0) + \log Z(\boldsymbol{\lambda}(\boldsymbol{u})) - \log Z(\boldsymbol{\lambda}(\boldsymbol{u}_0))$.

Next, we multiply $\bar{\boldsymbol{\lambda}}(\boldsymbol{u})$ to the both sides of (23) and evaluate it at m points, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$. Finally, taking the summation for $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ yields

$$\underbrace{\left(\sum_{i=1}^m \bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)^\top\right)}_{\bar{\boldsymbol{\Lambda}}} \boldsymbol{q}^\star(\boldsymbol{s}) + \sum_{i=1}^m \left(\epsilon(\boldsymbol{u}_i)\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_i)}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_i)} - \epsilon(\boldsymbol{u}_0)\frac{\delta(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)}{p^\star(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0)}\right)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)$$
$$+ \sum_{i=1}^m \left\{O(\epsilon(\boldsymbol{u}_i)^2) + \epsilon(\boldsymbol{u}_0)^2)\right\}\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i) = \left(\sum_{i=1}^m \bar{\boldsymbol{w}}(\boldsymbol{u}_i)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i)^\top\right)\boldsymbol{h}(\boldsymbol{x}) + \sum_{i=1}^m \bar{\beta}(\boldsymbol{u}_i)\bar{\boldsymbol{\lambda}}(\boldsymbol{u}_i).$$

Applying the inverse of $\bar{\boldsymbol{\Lambda}}$ to both sides completes the proof. □

B Influence of outliers in general non-exponential case

As in Hyvärinen et al. [2019], we assume that the general conditional independence (4) holds, and that both the mixing function $\boldsymbol{f}$ in the data generative model (1) and a nonlinear feature $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_{d_\mathrm{x}}(\boldsymbol{x}))^\top$ are invertible. Then, by the change of variables from the generative model (1),

$$\boldsymbol{v}(\boldsymbol{y}) := \boldsymbol{s} = \boldsymbol{f}^{-1} \circ \boldsymbol{h}^{-1}(\boldsymbol{y}), \text{ and } \boldsymbol{y} := \boldsymbol{h}(\boldsymbol{x})$$

where $\circ$ denotes composition. For the case of no outliers, Hyvärinen et al. [2019] proved that the nonlinear feature $\boldsymbol{h}(\boldsymbol{x})$ asymptotically recover the source $\boldsymbol{s}$ up to elementwise invertible transformations by showing that the following equations hold: For all $i, j, k = 1, \ldots, d_\mathrm{x}$ with $j \neq k$,

$$\frac{\partial}{\partial y_j} v_i(\boldsymbol{y}) = 0, \ \frac{\partial}{\partial y_k} v_i(\boldsymbol{y}) = 0, \text{ and } \frac{\partial^2}{\partial y_j \partial y_k} v_i(\boldsymbol{y}) = 0. \quad (24)$$

Eqs.(24) indicate that each $v_i(= s_i)$ is a function of only one distinct element in $\boldsymbol{y}(= \boldsymbol{h}(\boldsymbol{x}))$, and thus $\boldsymbol{s}$ is identifiable by $\boldsymbol{h}(\boldsymbol{x})$ up to elementwise invertible transformations. By denoting the first- and second-order derivatives of $v_i(\boldsymbol{y})$ in (24) as

$$v_i^j := \frac{\partial}{\partial y_j} v_i(\boldsymbol{y}), \qquad v_i^{j,k} := \frac{\partial^2}{\partial y_j \partial y_k} v_i(\boldsymbol{y}),$$

Eqs.(24) for all i, j, k can be compactly expressed as the following matrix form:

$$\boldsymbol{M}(\boldsymbol{v}) = \boldsymbol{O}, \quad (25)$$

where $\boldsymbol{O}$ denotes the null matrix, and $\boldsymbol{M}(\boldsymbol{v})$ is a $(d_\mathrm{x}^2 - d_\mathrm{x})$ by $2d_\mathrm{x}$ matrix and defined by

$$\boldsymbol{M}(\boldsymbol{v}) := \begin{pmatrix} v_1^1 v_1^2 & v_2^1 v_2^2 & \cdots & v_{d_\mathrm{x}}^1 v_{d_\mathrm{x}}^2 & v_1^{12} & v_2^{12} & \cdots & v_{d_\mathrm{x}}^{12} \\ v_1^1 v_1^3 & v_2^1 v_2^3 & \cdots & v_{d_\mathrm{x}}^1 v_{d_\mathrm{x}}^3 & v_1^{13} & v_2^{13} & \cdots & v_{d_\mathrm{x}}^{13} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ v_1^{d_\mathrm{x}-1} v_1^{d_\mathrm{x}} & v_2^{d_\mathrm{x}-1} v_2^{d_\mathrm{x}} & \cdots & v_{d_\mathrm{x}}^{d_\mathrm{x}-1} v_{d_\mathrm{x}}^{d_\mathrm{x}} & v_1^{d_\mathrm{x}-1,d_\mathrm{x}} & v_2^{d_\mathrm{x}-1,d_\mathrm{x}} & \cdots & v_{d_\mathrm{x}}^{d_\mathrm{x}-1,d_\mathrm{x}} \end{pmatrix}.$$

However, in the presence of outliers, (25) does not hold in general. In order to investigate the influence from outliers, we define the following notations:

$$\widetilde{\boldsymbol{M}}(\boldsymbol{v}) := -\begin{pmatrix} v_1^1 v_2^2 & v_1^1 v_3^2 & \cdots & v_{d_\mathrm{x}-1}^1 v_{d_\mathrm{x}}^2 \\ v_1^1 v_2^3 & v_1^1 v_3^3 & \cdots & v_{d_\mathrm{x}-1}^1 v_{d_\mathrm{x}}^3 \\ \vdots & \vdots & \cdots & \vdots \\ v_1^{d_\mathrm{x}-1} v_2^{d_\mathrm{x}} & v_1^{d_\mathrm{x}-1} v_3^{d_\mathrm{x}} & \cdots & v_{d_\mathrm{x}-1}^{d_\mathrm{x}-1} v_{d_\mathrm{x}}^{d_\mathrm{x}} \end{pmatrix}$$

$$\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) := \left(\frac{\partial^2}{\partial v_1^2}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0), \ldots, \frac{\partial^2}{\partial v_{d_\mathrm{x}}^2}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0), \frac{\partial}{\partial v_1}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0), \ldots, \frac{\partial}{\partial v_{d_\mathrm{x}}}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0) \right)^\top$$

$$\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) := \left(\frac{\partial^2}{\partial v_1 \partial v_2}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0), \frac{\partial^2}{\partial v_1 \partial v_3}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0), \ldots, \frac{\partial^2}{\partial v_{d_\mathrm{x}-1} \partial v_{d_\mathrm{x}}}\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0) \right)^\top,$$

where $\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0) := \log p(\boldsymbol{v}|\boldsymbol{u}) - \log p(\boldsymbol{v}|\boldsymbol{u}_0)$, $\boldsymbol{u}_0$ denotes a fixed point of $\boldsymbol{u}$, $\widetilde{\boldsymbol{M}}(\boldsymbol{v})$ by a $(d_\mathrm{x}^2 - d_\mathrm{x})$ by $(d_\mathrm{x}^2 - d_\mathrm{x})$ matrix, while $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{u}_0)$ and $\widetilde{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{u}_0)$ is the $2d_\mathrm{x}$- and $(d_\mathrm{x}^2 - d_\mathrm{x})$-dimensional vectors, respectively. The following theorem is useful to understand how (25) is affected by outliers:

Theorem 3. *Assume that*

(B1) Data $\boldsymbol{x}$ is generated from (1) *where $\boldsymbol{f}$ is invertible.*

(B2) Conditional independence (4) *holds.*

(B3) For all $\boldsymbol{v}$ and $\boldsymbol{u}$, $\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$, $\frac{\delta^i(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$ and $\frac{\delta^{i,j}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$ are finite where $\delta^i(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial}{\partial v_i}\delta(\boldsymbol{v}|\boldsymbol{u})$ and $\delta^{i,j}(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial^2}{\partial v_i \partial v_j}\delta(\boldsymbol{v}|\boldsymbol{u})$.

(B4) The conditional density of $\boldsymbol{x}$ given $\boldsymbol{u}$ is universally approximated with an invertible feature extractor $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_{d_\mathrm{x}}(\boldsymbol{x}))$ as

$$\log \frac{p(\boldsymbol{x}|\boldsymbol{u})}{c(\boldsymbol{x})e(\boldsymbol{u})} = \sum_{i=1}^{d_\mathrm{x}} \psi(h_i(\boldsymbol{x}), \boldsymbol{u}), \tag{26}$$

where ψ, c and e are nonzero scalar functions.

Then, under the contaminated density model (5)*, the following holds at a fixed point $\boldsymbol{u}_0$:*

$$\boldsymbol{M}(\boldsymbol{v})\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) = \widetilde{\boldsymbol{M}}(\boldsymbol{v})\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0). \tag{27}$$

The proof is given in Section B.1. First of all, in the case of no outliers, Theorem 3 essentially recovers Theorem 1 in Hyvärinen et al. [2019]: When $\epsilon(\boldsymbol{u}) = 0$ for all $\boldsymbol{u}$, $\varphi(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{u}_0) = \sum_{i=1}^{d_\mathrm{x}} \{q^\star(v_i|\boldsymbol{u}_0) - q^\star(v_i|\boldsymbol{u}_0)\}$ and thus $\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) = \boldsymbol{0}$, which yields from (27)

$$\boldsymbol{M}(\boldsymbol{v})\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) = \boldsymbol{O}.$$

Then, applying an additional assumption called *Assumption of Variability* [Hyvärinen et al., 2019][1] recovers (25).

However, when $\epsilon(\boldsymbol{u}) \neq 0$, $\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ is a nonzero vector and thus the right-hand sides on (27) are nonzeros in general. Let us remind that the elements in $\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ are given by

$$\frac{\partial^2}{\partial v_i \partial v_j}\left\{\log p(\boldsymbol{v}|\boldsymbol{u}) - \log p(\boldsymbol{v}|\boldsymbol{u}_0)\right\}, \tag{28}$$

where $i, j = 1, \ldots, d_\mathrm{x}$ but $i \neq j$. Obviously, it is not easy to understand when these elements (28) are strongly deviated from zeros, yet the following proposition proved in Section B.2 gives some insight:

[1]The assumption mean that there exist $2d_\mathrm{x}+1$ points, $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{2d_\mathrm{x}}$ such that $\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_0), \ldots, \boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}_{2d_\mathrm{x}}, \boldsymbol{u}_0)$ are linear independent, implying that the conditional density $p^\star(\boldsymbol{v}|\boldsymbol{u})$ is sufficiently complex and diverse.

Proposition 1. *Assume than the conditional independence* (4) *holds and sources are generated from the contaminated density model* (5)*. Then, with sufficiently small* $\epsilon(\boldsymbol{u})$*,*

$$\frac{\partial^2}{\partial v_i \partial v_j} \log p(\boldsymbol{v}|\boldsymbol{u}) = \epsilon(\boldsymbol{u}) \left\{ \frac{\delta^{i,j}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - q^{\star i}(v_i|\boldsymbol{u}) \frac{\delta^j(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - q^{\star j}(v_j|\boldsymbol{u}) \frac{\delta^i(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} + q^{\star i}(v_i|\boldsymbol{u}) q^{\star j}(v_j|\boldsymbol{u}) \frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \right\} + O(\epsilon(\boldsymbol{u})^2), \quad (29)$$

where $i \neq j$ *and* $q^{\star i}(v_i|\boldsymbol{u}) := \frac{\partial}{\partial v_i} q^\star(v_i|\boldsymbol{u})$.

Eq.(29) allows us to more easily understand the implication of (28): $\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ would be strongly deviated from the zero vector when at least one of $\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$, $\frac{\delta^i(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$, $\frac{\delta^j(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$ and $\frac{\delta^{i,j}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}$ is large. The four ratio factors could be very large when smooth $\delta(\boldsymbol{v}|\boldsymbol{u})$ lies on the tails of $p^\star(\boldsymbol{v}|\boldsymbol{u})$.

B.1 Proof of Theorem 3

Proof. We first obtain the expression of $\log p(\boldsymbol{x}|\boldsymbol{u})$ by the change of variables the data generative model (1) as

$$\log p(\boldsymbol{x}|\boldsymbol{u}) = \log p(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) - \log Z(\boldsymbol{u}) + \log |\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|,$$

where $\boldsymbol{g} := \boldsymbol{f}^{-1}$, and $\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})$ and $\det$ denote the Jacobian and determinant, respectively. Then, the universal approximation assumption (B4) gives

$$\sum_{i=1}^{d_{\mathrm{x}}} \psi(h_i(\boldsymbol{x}), \boldsymbol{u}) = \log p(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) - \log Z(\boldsymbol{u}) + \log |\det \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})|.$$

To remove the Jacobian term and log-partition function, we compute the differences of the above equations at $\boldsymbol{u}$ and a fixed point $\boldsymbol{u}_0$ as

$$\sum_{i=1}^{d_{\mathrm{x}}} \bar{\psi}(h_i(\boldsymbol{x}), \boldsymbol{u}, \boldsymbol{u}_0) = \log p(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}) - \log p(\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{u}_0).$$

where

$$\bar{\psi}(h_i(\boldsymbol{x}), \boldsymbol{u}, \boldsymbol{u}_0) := \psi(h_i(\boldsymbol{x}), \boldsymbol{u}) - \psi(h_i(\boldsymbol{x}), \boldsymbol{u}_0).$$

By the further change of variables $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{h}^{-1}(\boldsymbol{y}))$,

$$\sum_{i=1}^{d_{\mathrm{x}}} \bar{\psi}(y_i, \boldsymbol{u}, \boldsymbol{u}_0) = \varphi(\boldsymbol{v}(\boldsymbol{y}), \boldsymbol{u}, \boldsymbol{u}_0), \quad (30)$$

where

$$\varphi(\boldsymbol{v}(\boldsymbol{y}), \boldsymbol{u}, \boldsymbol{u}_0) := \log p(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{u}) - \log p(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{u}_0).$$

Next, let us use or remind the following notations:

$$\varphi^l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) := \frac{\partial}{\partial v_l} \varphi(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0), \qquad \varphi^{l,m}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) := \frac{\partial^2}{\partial v_l \partial v_m} \varphi(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$$
$$v_l^k := \frac{\partial}{\partial y_k} v_l(\boldsymbol{y}), \qquad v_l^{j,k} := \frac{\partial^2}{\partial y_j \partial y_k} v_l(\boldsymbol{y}).$$

Taking the second-order partial derivative of the left-hand side on (30) with respect to y_j and y_k for $j \neq k$ yields

$$\frac{\partial^2}{\partial y_j \partial y_k} \sum_{i=1}^{d_{\mathrm{x}}} \bar{\psi}(y_i, \boldsymbol{u}, \boldsymbol{u}_0) = 0. \quad (31)$$

On the other hand, the second-order partial derivative of the right-hand side on (30) can be expressed as

$$\frac{\partial^2}{\partial y_j \partial y_k}\varphi(\boldsymbol{v}(\boldsymbol{y}), \boldsymbol{u}, \boldsymbol{u}_0) = \sum_{l=1}^{d_{\mathrm{x}}} \left[\varphi^{l,l}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^j v_l^k + \varphi^l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^{jk}\right] + \sum_{l=1}^{d_{\mathrm{x}}} \sum_{\substack{m=1 \\ m\neq l}}^{d_{\mathrm{x}}} \varphi^{l,m}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^j v_m^k. \tag{32}$$

Equating (31) with (32) under (30) gives the following equation:

$$\sum_{l=1}^{d_{\mathrm{x}}} \left[\varphi^{l,l}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^j v_l^k + \varphi^l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^{jk}\right] = -\sum_{l=1}^{d_{\mathrm{x}}} \sum_{\substack{m=1 \\ m\neq l}}^{d_{\mathrm{x}}} \varphi^{l,m}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) v_l^j v_m^k. \tag{33}$$

Regarding $j = 1, \ldots, d_{\mathrm{x}}$ and $k = 1, \ldots, d_{\mathrm{x}}$, we collect all of $v_l^j v_l^k$, v_l^{jk} and $-v_l^j v_m^k$ for $j \neq k$ and $l \neq m$ as $(d_{\mathrm{x}}^2 - d_{\mathrm{x}})$-dimensional vectors, $\boldsymbol{a}_l(\boldsymbol{v})$, $\boldsymbol{b}_l(\boldsymbol{v})$, and $\boldsymbol{c}_{l,m}(\boldsymbol{v})$, respectively. Then, (33) for all j and k (but $j \neq k$) can be expressed as

$$\sum_{l=1}^{d} \left[\varphi^{l,l}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) \boldsymbol{a}_l(\boldsymbol{v}) + \varphi^l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) \boldsymbol{b}_l(\boldsymbol{v})\right] = \sum_{l=1}^{d_{\mathrm{x}}} \sum_{\substack{m=1 \\ m\neq l}}^{d_{\mathrm{x}}} \varphi^{l,m}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) \boldsymbol{c}_{l,m}(\boldsymbol{v}).$$

Furthermore, the above equations for all l and m (but $l \neq m$) can be summarized as the following system of linear equations:

$$\boldsymbol{M}(\boldsymbol{v})\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0) = \widetilde{\boldsymbol{M}}(\boldsymbol{v})\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0),$$

where $\boldsymbol{M}(\boldsymbol{v}) := (\boldsymbol{a}_1(\boldsymbol{v}), \ldots, \boldsymbol{a}_n(\boldsymbol{v}), \boldsymbol{b}_1(\boldsymbol{v}), \ldots, \boldsymbol{b}_n(\boldsymbol{v}))$, $\widetilde{\boldsymbol{M}}(\boldsymbol{v}) := (\boldsymbol{c}_{1,2}(\boldsymbol{v}), \boldsymbol{c}_{1,3}(\boldsymbol{v}), \ldots, \boldsymbol{c}_{d_{\mathrm{x}}, d_{\mathrm{x}}-1}(\boldsymbol{v}))$, and $\boldsymbol{w}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ is the $2d_{\mathrm{x}}$-dimensional vector of all $\varphi^{l,l}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ and $\varphi^l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ for $l = 1, \ldots, d_{\mathrm{x}}$, and $\widetilde{\boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ is the $(d_{\mathrm{x}}^2 - d_{\mathrm{x}})$-dimensional vector of all $\varphi^{l,m}(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{u}_0)$ for $l, m = 1, \ldots, d_{\mathrm{x}}$ but $l \neq m$. Thus, the proof is completed. □

B.2 Proof of Proposition 1

Proof. We first define the following notations:

$$p^l(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial}{\partial v_l} p(\boldsymbol{v}|\boldsymbol{u}), \qquad p^{l,m}(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial^2}{\partial v_l \partial v_m} p(\boldsymbol{v}|\boldsymbol{u})$$
$$p^{\star l}(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial}{\partial v_l} p^\star(\boldsymbol{v}|\boldsymbol{u}), \qquad p^{\star l,m}(\boldsymbol{v}|\boldsymbol{u}) := \frac{\partial^2}{\partial v_l \partial v_m} p^\star(\boldsymbol{v}|\boldsymbol{u}).$$

Then, under the contaminated density model (5), we compute

$$\begin{aligned}\frac{\partial^2}{\partial v_l \partial v_m} \log p(\boldsymbol{v}|\boldsymbol{u}) &= \frac{\partial^2}{\partial v_l \partial v_m} \log p^\star(\boldsymbol{v}|\boldsymbol{u}) + \frac{\partial^2}{\partial v_l \partial v_m} \log \left\{1 + \epsilon(\boldsymbol{u}) \left(\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - 1\right)\right\} \\ &= \frac{\partial^2}{\partial v_l \partial v_m} \log \left\{1 + \epsilon(\boldsymbol{u}) \left(\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - 1\right)\right\},\end{aligned} \tag{34}$$

where we used the following relation under the conditional independence assumption (4):

$$\frac{\partial^2}{\partial v_l \partial v_m} \log p^\star(\boldsymbol{v}|\boldsymbol{u}) = \frac{\partial^2}{\partial v_l \partial v_m} \sum_{i=1}^{d_{\mathrm{x}}} q^\star(v_i|\boldsymbol{u}) = 0.$$

Next, we apply the Taylor expansion with sufficiently small $\epsilon(\boldsymbol{u})$ to the right-hand side on (34):

$$\begin{aligned}&\frac{\partial^2}{\partial v_l \partial v_m} \log \left\{1 + \epsilon(\boldsymbol{u}) \left(\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - 1\right)\right\} \\ &= \frac{\partial^2}{\partial v_l \partial v_m} \left\{\epsilon(\boldsymbol{u}) \left(\frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - 1\right) + O(\epsilon(\boldsymbol{u})^2)\right\} \\ &= \epsilon(\boldsymbol{u}) \left\{\frac{\delta^{l,m}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - \frac{\delta^l(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \frac{p^{\star m}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - \frac{\delta^m(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \frac{p^{\star l}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} + \frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \frac{p^{\star l}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \frac{p^{\star n}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})}\right\} + O(\epsilon(\boldsymbol{u})^2).\end{aligned} \tag{35}$$

Substituting (35) into (34) yields

$$\frac{\partial^2}{\partial v_l \partial v_m} \log p(\boldsymbol{v}|\boldsymbol{u}) = \epsilon(\boldsymbol{u}) \left\{ \frac{\delta^{l,m}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - q^{\star l}(v_l|\boldsymbol{u}) \frac{\delta^m(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} - q^{\star m}(v_m|\boldsymbol{u}) \frac{\delta^l(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} + q^{\star l}(v_l|\boldsymbol{u}) q^{\star m}(v_m|\boldsymbol{u}) \frac{\delta(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} \right\} + O(\epsilon(\boldsymbol{u})^2),$$

where we applied

$$\frac{p^{\star l}(\boldsymbol{v}|\boldsymbol{u})}{p^\star(\boldsymbol{v}|\boldsymbol{u})} = \frac{\partial}{\partial v_l} \log p^\star(\boldsymbol{v}|\boldsymbol{u}) = q^{\star l}(v_l|\boldsymbol{u}),$$

under the conditional independence assumption (4). Thus, the proof is completed.

□

C Proof of Theorem 2

C.1 Derivation of (13)

With $p(y, \boldsymbol{x}, \boldsymbol{u}) = p(\boldsymbol{x}, \boldsymbol{u}|y)p(y)$, we have

$$\begin{aligned}
d_\gamma(p(y|\boldsymbol{x}, \boldsymbol{u}), r(\boldsymbol{x}, \boldsymbol{u}); p(\boldsymbol{x}, \boldsymbol{u})) &:= -\frac{1}{\gamma} \log \iint \sum_{y=0}^{1} \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{y(\gamma+1)}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(y, \boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} \\
&= -\frac{1}{\gamma} \log \left[\iint \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x}, \boldsymbol{u}|y=1)p(y=1) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} \right. \\
&\qquad \left. + \iint \left\{ \frac{1}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x}, \boldsymbol{u}|y=0)p(y=0) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} \right] \\
&= -\frac{1}{\gamma} \log \left[\frac{1}{2} \underbrace{\iint \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u}}_{\text{(A)}} \right. \\
&\qquad \left. + \frac{1}{2} \iint \left\{ \frac{1}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x})p(\boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} \right], \qquad (36)
\end{aligned}$$

where $p(y=0) = p(y=1) = \frac{1}{2}$, $p(\boldsymbol{x}, \boldsymbol{u}|y=1) = p(\boldsymbol{x}, \boldsymbol{u})$, and $p(\boldsymbol{x}, \boldsymbol{u}|y=0) = p(\boldsymbol{x})p(\boldsymbol{u})$. Under the outlier model, the joint density can be expressed as

$$p(\boldsymbol{x}, \boldsymbol{u}) = p(\boldsymbol{x}|\boldsymbol{u})p(\boldsymbol{u}) = (1 - \epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}, \boldsymbol{u}) + \epsilon(\boldsymbol{u})\delta(\boldsymbol{x}, \boldsymbol{u}).$$

Then, the term (A) in (36) can be written as

$$\begin{aligned}
\text{(A)} &= \iint \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} \\
&= \iint \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} (1 - \epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u} + \iint \left\{ \frac{r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}}{1 + r(\boldsymbol{x}, \boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \epsilon(\boldsymbol{u})\delta(\boldsymbol{x}, \boldsymbol{u}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{u}
\end{aligned} \qquad (37)$$

Finally, substituting (37) into (36) yields

$$
\begin{aligned}
&d_\gamma(p(y|\boldsymbol{x},\boldsymbol{u}), r(\boldsymbol{x},\boldsymbol{u}); p(\boldsymbol{x},\boldsymbol{u})) \\
&= -\frac{1}{\gamma}\log\left[\frac{1}{2}\iint\left\{\frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}\right\}^{\frac{\gamma}{\gamma+1}}(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} + \frac{1}{2}\iint\left\{\frac{1}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}\right\}^{\frac{\gamma}{\gamma+1}} p(\boldsymbol{x})p(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\right. \\
&\qquad \left. + \frac{1}{2}\underbrace{\iint\left\{\frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}\right\}^{\frac{\gamma}{\gamma+1}}\epsilon(\boldsymbol{u})\delta(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}}_{=\nu}\right] \\
&= J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})] + O(\nu),
\end{aligned}
$$

where we applied the relation $\log(y+\nu) = \log(y) + O(\nu)$ with sufficiently small ν.

C.2 Proof of the minimizer (14)

C.2.1 Preliminaries

We use the following results in the main proof, which are derived from the Taylor expansion:

$$
\left(\frac{1}{(r+\eta\phi)^{\gamma+1}+1}\right)^{\frac{\gamma}{\gamma+1}} = \left(\frac{1}{r^{\gamma+1}+1}\right)^{\frac{\gamma}{\gamma+1}} - \eta\frac{\gamma r^\gamma\phi}{(r^{\gamma+1}+1)^{\frac{2\gamma+1}{\gamma+1}}} + \frac{\eta^2}{2}\frac{\{\gamma(\gamma+1)r^{2\gamma} - \gamma^2 r^{\gamma-1}\}\phi^2}{(r^{\gamma+1}+1)^{\frac{3\gamma+1}{\gamma+1}}} + O(\eta^3) \tag{38}
$$

$$
\left(\frac{(r+\eta\phi)^{\gamma+1}}{(r+\eta\phi)^{\gamma+1}+1}\right)^{\frac{\gamma}{\gamma+1}} = \left(\frac{r^{\gamma+1}}{r^{\gamma+1}+1}\right)^{\frac{\gamma}{\gamma+1}} + \eta\frac{\gamma r^{\gamma-1}\phi}{(r^{\gamma+1}+1)^{\frac{2\gamma+1}{\gamma+1}}} - \frac{\eta^2}{2}\frac{\{\gamma(\gamma+2)r^{2\gamma-1} - \gamma(\gamma-1)r^{\gamma-2}\}\phi^2}{(r^{\gamma+1}+1)^{\frac{3\gamma+2}{\gamma+1}}} + O(\eta^3). \tag{39}
$$

C.2.2 Main proof

Proof. Let us define $\tilde{J}[r] := \exp(-\gamma J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})])$, and then we derive a maximizer of $\tilde{J}[r]$ alternative to a minimizer of $J[r(\boldsymbol{x},\boldsymbol{u}); (1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}), p(\boldsymbol{x})p(\boldsymbol{u})])$. For $\eta > 0$ and a perturbation ϕ, with (38) and (39), we have

$$
\begin{aligned}
\tilde{J}[r+\eta\phi] = \tilde{J}[r] &+ \frac{\eta}{2}\int\int\left[\frac{\gamma r(\boldsymbol{x},\boldsymbol{u})^{\gamma-1}\phi(\boldsymbol{x},\boldsymbol{u})\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}) - r(\boldsymbol{x},\boldsymbol{u})p(\boldsymbol{x})p(\boldsymbol{u})\}}{(r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}+1)^{\frac{2\gamma+1}{\gamma+1}}}\right]\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \\
&+ \frac{\eta^2}{4}\iint\left[\frac{\left\{\gamma(\gamma+1)r(\boldsymbol{x},\boldsymbol{u})^{2\gamma} - \gamma^2 r(\boldsymbol{x},\boldsymbol{u})^{\gamma-1}\right\}p(\boldsymbol{x})p(\boldsymbol{u})\phi(\boldsymbol{x},\boldsymbol{u})^2}{(r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}+1)^{\frac{3\gamma+2}{\gamma+1}}}\right. \\
&\left. - \frac{\left\{\gamma(\gamma+2)r(\boldsymbol{x},\boldsymbol{u})^{2\gamma-1} - \gamma(\gamma-1)r(\boldsymbol{x},\boldsymbol{u})^{\gamma-2}\right\}(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u})\phi(\boldsymbol{x},\boldsymbol{u})^2}{(r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}+1)^{\frac{3\gamma+2}{\gamma+1}}}\right]\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} + O(\eta^3).
\end{aligned} \tag{40}
$$

The optimality condition is satisfied when the term of order η on the right-hand side of (40) equals to zero for arbitrary ϕ. Thus, an optimizer $r^\star(\boldsymbol{x},\boldsymbol{u})$ satisfies the following equation:

$$
(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}) - r(\boldsymbol{x},\boldsymbol{u})p(\boldsymbol{x})p(\boldsymbol{u}) = 0.
$$

Then, we obtain $r^\star(\boldsymbol{x},\boldsymbol{u})$ as

$$
r^\star(\boldsymbol{x},\boldsymbol{u}) = \frac{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u})}{p(\boldsymbol{x})p(\boldsymbol{u})} = \frac{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}|\boldsymbol{u})}{p(\boldsymbol{x})}.
$$

To investigate if $r^\star$ is a maximizer of $J[r]$, we compute the term of order η^2 on the right-hand side of (40) at $r = r^\star$ as

$$
-\iint\left[\frac{\gamma\left\{r^\star(\boldsymbol{x},\boldsymbol{u})^{2\gamma} + r^\star(\boldsymbol{x},\boldsymbol{u})^{\gamma-1}\right\}p(\boldsymbol{x})p(\boldsymbol{u})\phi(\boldsymbol{x},\boldsymbol{u})^2}{(r^\star(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}+1)^{\frac{3\gamma+1}{\gamma+1}}}\right]\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u},
$$

where we used the relation, $(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x},\boldsymbol{u}) = r^\star(\boldsymbol{x},\boldsymbol{u})p(\boldsymbol{x})p(\boldsymbol{u})$. This shows that the term of order η^2 is negative for any choices of ϕ. Thus, $r^\star$ is a maximizer, and the proof is completed. □

D Proof of Proposition 1

Proof. We first define the support of $\delta(\boldsymbol{x}|\boldsymbol{u})$ as

$$\mathcal{X}_{\boldsymbol{u}}^{\delta} := \{\boldsymbol{x} \mid \delta(\boldsymbol{x}|\boldsymbol{u}) > 0\}.$$

In the neighborhood of $r^\star(\boldsymbol{x},\boldsymbol{u}) = \frac{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}|\boldsymbol{u})}{p(\boldsymbol{x})}$, we can obtain

$$\begin{aligned}
\nu &= \iint_{\mathcal{X},\mathcal{U}} \left\{ \frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \epsilon(\boldsymbol{u})\delta(\boldsymbol{x},\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} \\
&= \int_{\mathcal{U}} \left[\int_{\mathcal{X}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}}{1+r(\boldsymbol{x},\boldsymbol{u})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \epsilon(\boldsymbol{u})\delta(\boldsymbol{x}|\boldsymbol{u})\mathrm{d}\boldsymbol{x} \right] p(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\
&\leq \int_{\mathcal{U}} \left[\int_{\mathcal{X}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{p^\star(\boldsymbol{x}|\boldsymbol{u})^{\gamma+1}}{\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}|\boldsymbol{u})\}^{\gamma+1} + p(\boldsymbol{x})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \delta(\boldsymbol{x}|\boldsymbol{u})\mathrm{d}\boldsymbol{x} \right] g_\gamma(\boldsymbol{u}) + O\left(\sup_{\boldsymbol{x},\boldsymbol{u}} |r(\boldsymbol{x},\boldsymbol{u}) - r^\star(\boldsymbol{x},\boldsymbol{u})| \right),
\end{aligned} \tag{41}$$

where $g_\gamma(\boldsymbol{u}) := (1-\epsilon(\boldsymbol{u}))^\gamma \epsilon(\boldsymbol{u})p(\boldsymbol{u})$. Since the generative model (1) is invertible, by the change of variables from $\boldsymbol{x}$ to $\boldsymbol{s}$, we obtain the following equation:

$$\begin{aligned}
&\int_{\mathcal{X}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{p^\star(\boldsymbol{x}|\boldsymbol{u})^{\gamma+1}}{\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{x}|\boldsymbol{u})\}^{\gamma+1} + p(\boldsymbol{x})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \delta(\boldsymbol{x}|\boldsymbol{u})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_{\mathcal{S}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{p^\star(\boldsymbol{s}|\boldsymbol{u})^{\gamma+1}}{\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{s}|\boldsymbol{u})\}^{\gamma+1} + p(\boldsymbol{s})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \delta(\boldsymbol{s}|\boldsymbol{u})\mathrm{d}\boldsymbol{s}.
\end{aligned} \tag{42}$$

Substituting (42) into (41) yields

$$\nu \leq \int_{\mathcal{U}} \left[\int_{\mathcal{S}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{p^\star(\boldsymbol{s}|\boldsymbol{u})^{\gamma+1}}{\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{s}|\boldsymbol{u})\}^{\gamma+1} + p(\boldsymbol{s})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \delta(\boldsymbol{s}|\boldsymbol{u})\mathrm{d}\boldsymbol{s} \right] g_\gamma(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + O\left(\sup_{\boldsymbol{x},\boldsymbol{u}} |r(\boldsymbol{x},\boldsymbol{u}) - r^\star(\boldsymbol{x},\boldsymbol{u})| \right).$$

Under the assumption that $\mathcal{S}_{\boldsymbol{u}}^{p^\star} \cap \mathcal{S}_{\boldsymbol{u}}^{\delta} = \emptyset$, we can easily conform that for $\gamma > 0$,

$$\int_{\mathcal{S}_{\boldsymbol{u}}^{\delta}} \left\{ \frac{p^\star(\boldsymbol{s}|\boldsymbol{u})^{\gamma+1}}{\{(1-\epsilon(\boldsymbol{u}))p^\star(\boldsymbol{s}|\boldsymbol{u})\}^{\gamma+1} + p(\boldsymbol{s})^{\gamma+1}} \right\}^{\frac{\gamma}{\gamma+1}} \delta(\boldsymbol{s}|\boldsymbol{u})\mathrm{d}\boldsymbol{s} = 0.$$

The above equation gives

$$\nu \leq O\left(\sup_{\boldsymbol{x},\boldsymbol{u}} |r(\boldsymbol{x},\boldsymbol{u}) - r^\star(\boldsymbol{x},\boldsymbol{u})| \right).$$

Thus, the proof is completed. □

E Proof of Proposition 2

E.1 Main proof

Proof. We first define the following notations:

$$
\begin{aligned}
L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) &= \frac{1}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}} \\
S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) &= \{L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})(1-L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}))\}^{\gamma/(1+\gamma)} \\
\boldsymbol{\alpha}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) &= L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{1/(1+\gamma)} S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) \frac{\partial \log r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{\partial \boldsymbol{\theta}} \\
\boldsymbol{\beta}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) &= (1-L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}))^{1/(1+\gamma)}\, S_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) \frac{\partial \log r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{\partial \boldsymbol{\theta}} \\
\boldsymbol{C}_{\boldsymbol{\theta}} &= \iint \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{\top} \bar{p}(\boldsymbol{x},\boldsymbol{u}) - \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{\top} \bar{p}(\boldsymbol{x}) \bar{p}(\boldsymbol{u}) \right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}.
\end{aligned}
$$

Then, our proof relies on the following lemma proved in Section E.2:

Lemma 1. *The influence function of $\hat{\boldsymbol{\theta}}$ under the γ-cross entropy* (11) *satisfies the following equations:*

- *Under the contamination model (A),*

$$
\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} \mathrm{IF}(\bar{\boldsymbol{x}}) = \int \left\{ \boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u}) - \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u}) \right\} p^{\star}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{43}
$$

  where IF *denotes the influence function,*

- *Under the contamination model (B),*

$$
\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} \mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) = -\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\bar{\boldsymbol{u}}) p^{\star}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u}) p^{\star}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} - \iint \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u}) p^{\star}(\boldsymbol{x}) p^{\star}(\boldsymbol{u}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}. \tag{44}
$$

First of all, under Assumption (19), we observe that

$$
\lim_{\|\boldsymbol{x}\| \to \infty} \boldsymbol{\alpha}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) = \lim_{\|\boldsymbol{x}\| \to \infty} \boldsymbol{\beta}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}) = \mathbf{0} \tag{45}
$$

because $L_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{1/(1+\gamma)} \leq 1$. From Lemma 1, regarding the contamination model (A), (45) ensures that the right-hand side of (43) approaches to $\mathbf{0}$ as $\|\bar{\boldsymbol{x}}\| \to \infty$. Thus,

$$
\lim_{\|\bar{\boldsymbol{x}}\| \to \infty} \|\mathrm{IF}(\bar{\boldsymbol{x}})\| = \mathbf{0}.
$$

Furthermore, under Assumption (18),

$$
\sup_{\boldsymbol{x},\boldsymbol{u}} \|\boldsymbol{\alpha}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})\| < \infty, \qquad \sup_{\boldsymbol{x},\boldsymbol{u}} \|\boldsymbol{\beta}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})\| < \infty. \tag{46}
$$

This indicates that

$$
\sup_{\bar{\boldsymbol{x}}} \|\mathrm{IF}(\bar{\boldsymbol{x}})\| < \infty,
$$

and therefore $\hat{\boldsymbol{\theta}}$ is B-robust.

For the the contamination model (B), (46) holds, which indicates that the right-hand side of (44) is uniformly bounded even when $r_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})$, $r_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u})$ or $r_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\bar{\boldsymbol{u}})$ diverge through the influence of the outliers $(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})$. Thus, we have

$$
\sup_{\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}} \|\mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}})\| < \infty. \tag{47}
$$

□

E.2 Proof of Lemma 1

Proof. $\hat{\boldsymbol{\theta}}$ associated with the densities $p^\star(\boldsymbol{x},\boldsymbol{u})$ and $p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})$ is a solution of the following estimating equation:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\boldsymbol{\theta}} d_\gamma(p^\star(y|\boldsymbol{x},\boldsymbol{u}), r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}); p^\star(\boldsymbol{x},\boldsymbol{u}))\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
&\propto \left[\iint \left\{ \left(\frac{r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}\right)^{-1/(1+\gamma)} \frac{\frac{\partial}{\partial\boldsymbol{\theta}} r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{(1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)})^2} p^\star(\boldsymbol{x},\boldsymbol{u}) \right.\right. \\
&\qquad \left.\left. - \left(\frac{1}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}\right)^{-1/(1+\gamma)} \frac{\frac{\partial}{\partial\boldsymbol{\theta}} r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{(1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)})^2} p^\star(\boldsymbol{x})p^\star(\boldsymbol{u}) \right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
&= \iint \left\{\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u}) p^\star(\boldsymbol{x},\boldsymbol{u}) - \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u}) p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}
\end{aligned} \tag{48}$$

On the other hand, $\hat{\boldsymbol{\theta}}_\epsilon$ is a solution of the the estimating equation with the contaminated densities:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\boldsymbol{\theta}} d_\gamma(\bar{p}(y|\boldsymbol{x},\boldsymbol{u}), r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u}); \bar{p}(\boldsymbol{x},\boldsymbol{u}))\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\epsilon} \\
&\propto \left[\iint \left\{ \left(\frac{r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}\right)^{-1/(1+\gamma)} \frac{\frac{\partial}{\partial\boldsymbol{\theta}} r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{(1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)})^2} \bar{p}(\boldsymbol{x},\boldsymbol{u}) \right.\right. \\
&\qquad \left.\left. - \left(\frac{1}{1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}\right)^{-1/(1+\gamma)} \frac{\frac{\partial}{\partial\boldsymbol{\theta}} r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)}}{(1+r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{u})^{(\gamma+1)})^2} \bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u}) \right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\epsilon} \\
&= \iint \left\{\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u}) \bar{p}(\boldsymbol{x},\boldsymbol{u}) - \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u}) \bar{p}(\boldsymbol{x})\bar{p}(\boldsymbol{u})\right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}.
\end{aligned} \tag{49}$$

Next, we obtain (43) and (44) under two contamination models separately by combining (49) with (48).

E.2.1 Contamination model (A)

We remind that $\bar{p}(\boldsymbol{x},\boldsymbol{u})$, $\bar{p}(\boldsymbol{x})$, and $\bar{p}(\boldsymbol{u})$ in the contamination model (A) are defined as

$$\begin{aligned}
\bar{p}(\boldsymbol{x},\boldsymbol{u}) &= (1-\epsilon)p^\star(\boldsymbol{x},\boldsymbol{u}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x})p^\star(\boldsymbol{u}) \\
\bar{p}(\boldsymbol{x}) &= (1-\epsilon)p^\star(\boldsymbol{x}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x}) \\
\bar{p}(\boldsymbol{u}) &= p^\star(\boldsymbol{u}),
\end{aligned}$$

with $\bar{\delta}_{\bar{\boldsymbol{z}}}(\boldsymbol{z})$ is the Dirac delta function having a point mass at $\bar{\boldsymbol{z}}$. Then, the right-hand side of (49) is given by

$$(1-\epsilon)\underbrace{\iint \left\{\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u}) p^\star(\boldsymbol{x},\boldsymbol{u}) - \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u}) p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\right\} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}}_{(*)} + \epsilon\int \left\{\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\bar{\boldsymbol{x}},\boldsymbol{u}) - \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\bar{\boldsymbol{x}},\boldsymbol{u})\right\} p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \tag{50}$$

Taylor expansion of (50) around $\hat{\boldsymbol{\theta}}$ is given by

$$\mathbf{0} = \frac{\partial d_\gamma}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{\partial d_\gamma}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\hat{\boldsymbol{\theta}}_\epsilon - \hat{\boldsymbol{\theta}}) + O(\|\hat{\boldsymbol{\theta}}_\epsilon - \hat{\boldsymbol{\theta}}\|^2). \tag{51}$$

Since it follows from (48) that the term $(*)$ in (50) vanishes as $\hat{\boldsymbol{\theta}}_\epsilon = \hat{\boldsymbol{\theta}}$, we obtain (43) by taking the limit of $\epsilon \to 0$ because $\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} = \frac{\partial d_\gamma}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ in (51).

E.2.2 Contamination model (B)

In the contamination model (B), the auxiliary variables $\boldsymbol{u}$ are also contaminated as follows:

$$\begin{aligned}\bar{p}(\boldsymbol{x},\boldsymbol{u}) &= (1-\epsilon)p^\star(\boldsymbol{x},\boldsymbol{u}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x})\bar{\delta}_{\bar{\boldsymbol{u}}}(\boldsymbol{u})\\ \bar{p}(\boldsymbol{x}) &= (1-\epsilon)p^\star(\boldsymbol{x}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{x}}}(\boldsymbol{x})\\ \bar{p}(\boldsymbol{u}) &= (1-\epsilon)p^\star(\boldsymbol{u}) + \epsilon\bar{\delta}_{\bar{\boldsymbol{u}}}(\boldsymbol{u}).\end{aligned}$$

Then, the right-hand side of (49) is given by

$$\begin{aligned}&\iint \boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x},\boldsymbol{u}) - \iint \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u} + \epsilon\left[\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) - \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\bar{\boldsymbol{u}})p^\star(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right.\\ &\left. - \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\bar{\boldsymbol{x}},\boldsymbol{u})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \iint \boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x},\boldsymbol{u}) + 2\iint \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_\epsilon}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\right] + O(\epsilon^2), \qquad (52)\end{aligned}$$

By following the derivation in Section E.2.1, we have

$$\begin{aligned}\boldsymbol{C}_{\hat{\boldsymbol{\theta}}}\mathrm{IF}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) &= -\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\bar{\boldsymbol{u}})p^\star(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\\ &\quad + \iint \boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x},\boldsymbol{u}) - 2\iint \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\\ &= -\boldsymbol{\alpha}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\bar{\boldsymbol{u}}) + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\bar{\boldsymbol{u}})p^\star(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\bar{\boldsymbol{x}},\boldsymbol{u})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \iint \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x},\boldsymbol{u})p^\star(\boldsymbol{x})p^\star(\boldsymbol{u})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u},\end{aligned}$$

where (48) was applied in the last line. Thus, the proof is completed. □

F Robust property in multiclass classification

Here, we perform a similar robust analysis as Theorem 2 in multiclass classification. We first specifically assume that the following $\tilde{\nu}$ is sufficiently small:

$$\tilde{\nu} := \int \frac{\sum_u r(u,\boldsymbol{x})^\gamma \delta(\boldsymbol{x}|u)\epsilon(u)p(u)}{(\sum_{u'} r(u',\boldsymbol{x})^{\gamma+1})^{\frac{\gamma}{\gamma+1}}}\mathrm{d}\boldsymbol{x}. \qquad (53)$$

Then, the outlier model (5) decomposes the γ-cross entropy (20) into

$$\begin{aligned}&d_\gamma(p(u|\boldsymbol{x}), r(u,\boldsymbol{x}); p(\boldsymbol{x}))\\ &= -\frac{1}{\gamma}\log\int \frac{\sum_u r(u,\boldsymbol{x})^\gamma p(u|\boldsymbol{x})}{(\sum_{u'} r(u',\boldsymbol{x})^{\gamma+1})^{\frac{\gamma}{\gamma+1}}}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= -\frac{1}{\gamma}\log\left[\int \frac{\sum_u r(u,\boldsymbol{x})^\gamma (1-\epsilon(u))p^\star(\boldsymbol{x}|u)p(u)}{(\sum_{u'} r(u',\boldsymbol{x})^{\gamma+1})^{\frac{\gamma}{\gamma+1}}}\mathrm{d}\boldsymbol{x} + \int \frac{\sum_u r(u,\boldsymbol{x})^\gamma \delta(\boldsymbol{x}|u)\epsilon(u)p(u)}{(\sum_{u'} r(u',\boldsymbol{x})^{\gamma+1})^{\frac{\gamma}{\gamma+1}}}\mathrm{d}\boldsymbol{x}\right]\\ &= d_\gamma(r(u,\boldsymbol{x}), p^\star(\boldsymbol{x}|u); (1-\epsilon(u))p(u)) + O(\tilde{\nu}), \qquad (54)\end{aligned}$$

where we employed $\log(y+z) = \log(y) + O(z)$ with sufficiently small z and $p(u|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|u)p(u)}{p(\boldsymbol{x})} = \frac{(1-\epsilon(u))p^\star(\boldsymbol{x}|u)p(u)+\epsilon(u)\delta(\boldsymbol{x}|u)p(u)}{p(\boldsymbol{x})}$, and

$$d_\gamma(r(u,\boldsymbol{x}), p^\star(\boldsymbol{x}|u); (1-\epsilon(u))p(u)) := -\frac{1}{\gamma}\log\left[\sum_u \left\{\int \frac{r(u,\boldsymbol{x})^\gamma p^\star(\boldsymbol{x}|u)}{(\sum_{u'} r(u',\boldsymbol{x})^{\gamma+1})^{\frac{\gamma}{\gamma+1}}}\mathrm{d}\boldsymbol{x}\right\}(1-\epsilon(u))p(u)\right].$$

Eq.(54) indicates that minimization of $d_\gamma(p(u|\boldsymbol{x}), r(u,\boldsymbol{x}); p(\boldsymbol{x}))$ approximately equals to minimization of $d_\gamma(r(u,\boldsymbol{x}), p^\star(\boldsymbol{x}|u); (1-\epsilon(u))p(u))$ when $\tilde{\nu}$ is sufficiently small. In addition, $d_\gamma(r(u,\boldsymbol{x}), p^\star(\boldsymbol{x}|u); (1-\epsilon(u))p(u))$ is minimized at

$$r(u,\boldsymbol{x}) = p^\star(\boldsymbol{x}|u) \qquad (55)$$

because it is the γ-cross entropy to $p^\star(\boldsymbol{x}|u)$ under the measure $(1-\epsilon(u))p(u)$. The minimizer (55) is desirable in terms of the universal approximation assumptions (A4) and (B4).

Next, we discuss when $\tilde{\nu}$ is sufficiently small. Following Section D, in the neighborhood of $p^\star(\boldsymbol{x}|u)$,

$$\begin{aligned}\tilde{\nu} &\leq \sum_u \left\{ \int \frac{p^\star(\boldsymbol{x}|u)^\gamma \delta(\boldsymbol{x}|u)}{(\sum_{u'} p^\star(\boldsymbol{x}|u')^{\gamma+1})^{\frac{\gamma}{\gamma+1}}} \mathrm{d}\boldsymbol{x} \right\} \epsilon(u)p(u) + O(\sup_{u,\boldsymbol{x}} |r(u,\boldsymbol{x}) - p^\star(\boldsymbol{x}|u)|) \\ &= \sum_u \left\{ \int \frac{p^\star(\boldsymbol{s}|u)^\gamma \delta(\boldsymbol{s}|u)}{(\sum_{u'} p^\star(\boldsymbol{s}|u')^{\gamma+1})^{\frac{\gamma}{\gamma+1}}} \mathrm{d}\boldsymbol{s} \right\} \epsilon(u)p(u) + O(\sup_{u,\boldsymbol{x}} |r(u,\boldsymbol{x}) - p^\star(\boldsymbol{x}|u)|), \end{aligned} \tag{56}$$

where we performed the change of variables from $\boldsymbol{x}$ to $\boldsymbol{s}$ under the data generate model (1). When $\sum_{u'} p^\star(\boldsymbol{s}|u')^{\gamma+1} \neq 0$ and the supports of $p^\star(\boldsymbol{s}|u)$ and $\delta(\boldsymbol{s}|u)$ are mutually disjoint as in Section D, we can make the same implication as Proposition 1: $\tilde{\nu}$ can be sufficiently small in the neighborhood of $p^\star(\boldsymbol{x}|u)$ when $p^\star(\boldsymbol{s}|u)$ and $\delta(\boldsymbol{s}|u)$ are clearly separated. This clear separation possibly happens on a situation where $\delta(\boldsymbol{s}|u)$ lies on the tails of $p^\star(\boldsymbol{s}|u)$ as seen in common contamination by outliers. Thus, the γ-cross entropy for multiclass classification would be also robust against outliers.

As discussed in Section 4.3, the non-robustness of TCL can be understood in terms of the γ-cross entropy. TCL employs the multiclass logistic regression whose cross-entropy can be obtained as a limit of $\gamma = 0$ in the γ-cross entropy (20). The multiclass logistic regression cannot also fulfill the robustness condition: It follows from the definition (53) that $\tilde{\nu}$ cannot be sufficiently small when $\gamma = 0$. Thus, this implies that TCL is sensitive to outliers.

G Experimental details

G.1 Robust time contrastive learning

Source vectors with time segment length 512 was first generated from (5): Following (2), given a time segment label u, the target density $p^\star(\boldsymbol{s}|u)$ was conditionally independent Laplace densities with means 0 and different scales across time segments, which were randomly determined from the uniform density on $[0, \frac{1}{\sqrt{2}}]$. Regarding the outlier density $\delta(\boldsymbol{s}|u)$, two types of densities were used: An independent Laplace density with mean 0 and scale 3.0, and a mixture of mean-modulated two Gaussians. More precisely, the outlier density of the mixture is given by $\delta(\boldsymbol{s}|u) = 0.5N(1+3y(u), 0.5)+0.5N(-1-3y(u), 0.5)$ where $N(a, b)$ is a Gaussian density with mean a and standard deviation b, and $y(u)$ is fixed at the scale parameter of $p^\star(\boldsymbol{s}|u)$ above in the u-th time segment. We set $\epsilon(u) = \epsilon$ for all time segments u. The total numbers of segments and of data samples were $K = 256$ and $T = 512 \times 256$, respectively. The dimensionality of data was $d_{\mathrm{x}} = 10$ in Table 1, while $d_{\mathrm{x}} = 5$ in Table 2. Then, data $\boldsymbol{x}$ was generated according to (1) where $\boldsymbol{f}(\boldsymbol{s})$ was modelled by a three-layer neural network (Table 1) or two-layer (Table 2) neural network with the leaky ReLU activation function and random weights. The numbers of all hidden and output units were the same as the dimensionality of data (i.e., d_{x}). As preprocessing, we performed whitening based on the γ-divergence [Chen et al., 2013].

ICA features $\boldsymbol{h}(\boldsymbol{x})$ both in RTCL and TCL were modelled by a three layer neural network where the number of hidden units was $4d_{\mathrm{x}}$, but the final layer was d_{x}. Regarding the activation functions, the final layer employed the absolute value function, while the other hidden layers were the max-out function [Goodfellow et al., 2013] with two groups. ℓ_2 regularization was employed with the regularization parameter 10^{-4}. When $\epsilon < 0.1$, we optimized the network parameters both in RTCL and TCL with 0.001 learning rate using the Adam optimizer for $1,000$ epochs with mini-batch size 256, while we updated the parameters for $3,000$ epochs for $\epsilon = 0.1$. We empirically observed that more iterations for parameter update are often needed to escape from bad local optima when the contamination ratio is larger. No postprocessing was applied. The performance was measured by the absolute value of the Pearson correlation coefficient between learned ICA features $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{s}$ without outliers.

When comparing with iVAE, the same neural network has been employed both in RTCL and iVAE, but the activation functions were all ReLU except for the final layer. We optimized the network parameters with 0.001 learning rate using the Adam optimizer for $1,000$ epochs, and the mini-batch size was 256.

G.2 Robust permutation contrastive learning

First, the temporally dependent T sources were generated from

$$\log p^{\star}(\boldsymbol{s}(t)|\boldsymbol{s}(t-1)) = -\sum_{i=1}^{d_{\mathrm{x}}} |s_i(t) - \rho s_i(t-1)| + C,$$

where C denotes a constant and the auto-regressive coefficient ρ was fixed at 0.7. The total number of sources was $T = 65,536$. Then, we randomly replaced the sources with outliers based on a constant contamination ratio ϵ, which were generated from the independent Gaussian density with the mean 0 and scale 3. Data $\boldsymbol{x}$ was generated as the nonlinear mixing of the sources with outliers by a three-layer neural networks with the leaky ReLU function and random weights according to (1). The same whitening preprocessing above was performed based on the γ-divergence [Chen et al., 2013].

Based on the universal approximation assumption in Hyvärinen and Morioka [2017, Theorem 1 and Eq.(12)] or Appendix B, we restrict the model r as

$$r(\boldsymbol{x}(t), \boldsymbol{u}(t)) = \exp\left(\sum_{i=1}^{d_{\mathrm{x}}} \psi_i(h_i(\boldsymbol{x}(t)), h_i(\boldsymbol{u}(t))\right),$$

with a neural network $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_{d_{\mathrm{x}}}(\boldsymbol{x}))^{\top}$. Following Hyvärinen and Morioka [2017], $\psi_i(h_i(\boldsymbol{x}), h_i(\boldsymbol{u}))$ was further modelled by

$$|a_{i,1}h_i(\boldsymbol{x}) + a_{i,2}h_i(\boldsymbol{u}) + b_i| - (\bar{a}_i h_i(\boldsymbol{x}) + \bar{b}_i)^2 + c,$$

where $a_{i,1}, a_{i,2}, b_i, \bar{a}_i, \bar{b}_i, c$ are parameters to be estimated from data. ICA features $\boldsymbol{h}(\boldsymbol{x})$ both in RPCL and PCL were modelled by a three layer feedforward neural network where the number of hidden units is $4d_{\mathrm{x}}$. No activation function was applied in the last layer, while the max-out function was employed in the other layers. ℓ_2 regularization was employed with the regularization parameter 10^{-4}. Then, we optimized the parameters in RPCL and PCL using the Adam optimizer with 0.001 learning rate for $1,000$ epochs with mini-batch size 128. The performance was measured by the absolute Pearson correlation coefficient between learned ICA features $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{s}$ without outliers.

H Visualization of outliers on fMRI data

Fig.1 visualizes the time series data from the Parahippocampal brain region, and demonstrates the presence of outliers in many of the segments.

References

P. Chen, H. Hung, O. Komori, S.-Y. Huang, and S. Eguchi. Robust independent component analysis via minimum γ-divergence estimation. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):614–624, 2013.

I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 1319–1327. PMLR, 2013.

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 460–469. PMLR, 2017.

A. Hyvärinen, H. Sasaki, and R. E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 859–868, 2019.

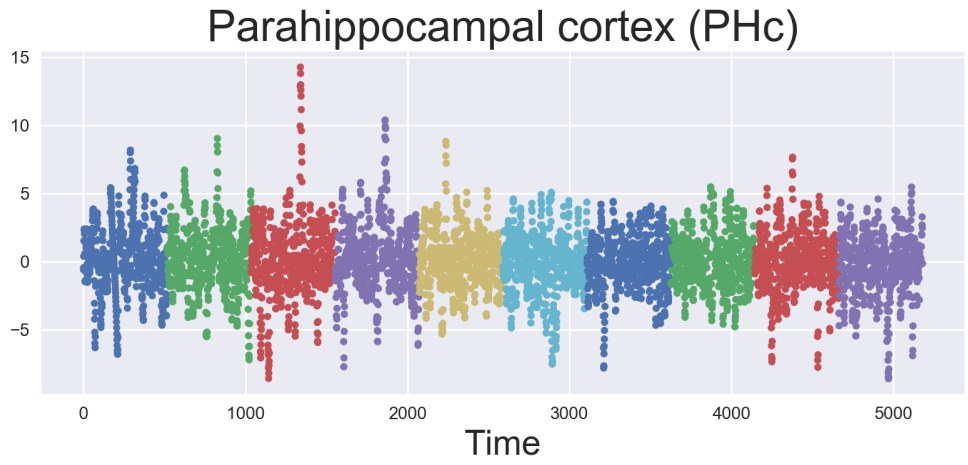

Figure 1: Subset of time series data corresponding to the Parahippocampal (PHc) region taken from the Hippocampal fMRI dataset. Different colors denote distinct segments, which in this case correspond to fMRI measurements from the same subject on distinct days.