t-DIVERGENCE:A NEW DIVERGENCE MEASURE WITH APPLICATION TO ROBUST STATISTICS & CLUSTERING

Debolina Paul
Affiliate, Machine Learning Research Group,
Indian Statistical Institute,
Kolkata, India

Saptarshi Chakraborty
Department of Statistics,
University of California,
Berkeley, USA

Swagatam Das
Electronics & Communication Sciences Unit,
Indian Statistical Institute,
Kolkata, India

ABSTRACT

This paper introduces the t-divergence, a novel divergence measure associated with the inverse tangent function. We investigate its intriguing consistent and outlier-robust features, particularly its quasi-metric properties and role in establishing weak convergence. Additionally, we showcase the efficacy of this divergence measure family in feature-weighted clustering for high-dimensional data.

1 PROPOSED t-DIVERGENCE

Practitioners often select divergence measures for their resilience against outliers. Employing less sensitive loss functions such as ℓ_1, Huber, or Geman-McClure typically guarantees this robustness. Nonetheless, these functions often lack smoothness, presenting challenges for derivative-based optimization methods. Our paper introduces the t-divergence, showcasing its efficacy in robust statistical estimation through comprehensive theoretical and experimental analysis. Formally, let Ω be the sample space and let $\mathcal{F}$ be a σ-algebra defined on it. Suppose $\mu : \mathcal{F} \to [0, \infty)$ be a measure defined on $(\Omega, \mathcal{F})$. Let D_μ be the set of all measures on $(\Omega, \mathcal{F})$, dominated by μ and $\int \frac{dP}{d\mu} \tan^{-1} \frac{dP}{d\mu} d\mu < \infty$, i.e. $D_\mu = \left\{ P : P \ll \mu \text{ and } \int \frac{dP}{d\mu} \tan^{-1} \frac{dP}{d\mu} d\mu < \infty \right\}$. Let, $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. The t-divergence, $D : D_\mu \times D_\mu \to [0, \infty)$, between measures $P, Q \in D_\mu$, is defined as

$$D(P,Q) = \int (p(x) - q(x)) \tan^{-1}(p(x) - q(x)) d\mu(x)$$

Some intriguing properties of the proposed t-divergence are as follows:

1. $D(P,Q) \geq 0$. Moreover, $D(P,Q) = 0$ *iff* $p = q$, a.e. $[\mu]$.
2. $D(P,Q) < \infty$, for all $P, Q \in D_\mu$.
3. Let D_μ^p denote the set of all probability measures, dominated by μ. Unlike many other divergence measures, we observe that $0 \leq D(P,Q) < \infty$ for any $P, Q \in D_\mu^p$. This is because $D_\mu^p \subseteq D_\mu$.
4. The t-divergence is symmetric, i.e. $D(P,Q) = D(Q,P)$.
5. $D(P,Q) \leq \pi TV(P,Q)$, where $TV(P,Q)$ is the total variation distance between P and Q.
6. Suppose $\{P_n\}_{n \geq 1}$ be a sequence of probability measures in D_μ^p. Also let P be another probability measure, dominated by μ. Then $\lim_{n\to\infty} D(P_n, P) = 0$ implies $P_n \to P$ in distribution.
7. $D(\cdot, \cdot)$ is a near-metric (Burgin, 2017), with $\rho = 2$.

2 APPLICATIONS

Application to Robust Statistical Inference Suppose $X_1, \dots, X_n$ are i.i.d. according to the distribution G. Let $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a family of distributions, indexed by θ. We assume that both $G, F_\theta \ll \mu$, for some dominating measure μ, for all $\theta \in \Theta$. Let $g = \frac{dG}{d\mu}$ and $f_\theta = \frac{dF_\theta}{d\mu}$ be respectively. We define the minimum t-functional as: $T(G) = \arg\min_{\theta\in\Theta} D(G, F_\theta)$.In the context of minimum divergence based estimation, one tries to minimise $D(G, F_\theta)$, w.r.t. θ, in order to obtain a point estimate of θ. Since in practice, the distribution G is unknown, one uses proxies (such as kernel density estimates or empirical cumulative distribution function) for G, based on the observed data $X_1, \dots, X_n$. Let this estimate be $\hat{G}_n$, which has a density $\hat{g}_n$ w.r.t μ. The estimate for θ, based on the data is given by $\hat{\theta} = T(G_n) = \arg\min_{\theta\in\Theta} D(\hat{G}_n, F_\theta)$. We call this estimator as the minimum t-estimator. It can be shown that the t-estimator exists and is unique under mild regularity conditions (refer to Theorem 2). Additionally, the minimum t-estimator is robust under certain regularity conditions. This is done by deriving its influence function and showing that its bounded. The influence function, $IF(\cdot)$ of a functional $T(\cdot)$ is defined though the following equation: $IF(y;T,G) = \lim_{\epsilon\downarrow 0}\frac{1}{\epsilon}(T((1-\epsilon)G+\epsilon\delta_y)-T(G))$, where δ_y denotes the degenerate distribution, putting all its mass at y. Informally, we state the following theorem.

Theorem 1. *Under mild assumptions, the first order influence function $IF(y;T,G)$ for the minimum t-functional is given by* $\big(\sum_{x\in\mathcal{X}}\big[\frac{2f^2_{\theta_0}(x)u_{\theta_0}(x)u'_{\theta_0}(x)}{(1+(f_{\theta_0}(x)-g(x))^2)^2}+\rho(f_{\theta_0}(x)-g(x))f_{\theta_0}(x)(u^2_{\theta_0}(x)+u'_{\theta_0}(x))\big]\big)^{-1}\big(\frac{2f_{\theta_0}(y)u_{\theta_0}(y)}{(1+(f_{\theta_0}(y)-g(y))^2)^2}-\sum_{x\in\mathcal{X}}\frac{g(x)f_{\theta_0}(x)u_{\theta_0}(x)}{(1+(f_{\theta_0}(x)-g(x))^2)^2}\big)$*,where* $\rho(x)=\tan^{-1}(x)+\frac{x}{1+x^2}$ *and* $u_\theta(x)=\frac{\partial}{\partial\theta}\log f_\theta(x)$.

The first-order influence function for the minimum t-estimator remains bounded as shown in Theorem 1 when $f_\theta(y)u_\theta(y)$ is bounded across all $\theta\in\Theta$ and for all $y\in\mathcal{X}$. This condition is satisfied by exponential families and numerous commonly used distributions.

To demonstrate the robustness of the t-estimator, we conduct experiments with 100 datapoints which consist of $(1-\epsilon)\%$ from $Binomial(50,\theta)$ and $\epsilon\%$ from $Uniform(40,\dots,50)$, with $\epsilon\in(0,45)$ and true $\theta=0.5$. We estimate θ under the binomial model using various methods, including maximum likelihood (MLE), median, minimum squared Hellinger estimate, minimum total variation estimate, and minimum t-estimate. We repeat this experiment 100 times and plot the average estimate for θin Figure 1. The results highlight the vulnerability of MLE and the median to even small outlier fractions. Conversely, the minimum t-estimate demonstrates outlier robustness, performing comparably to the minimum Hellinger and minimum total variation estimates.

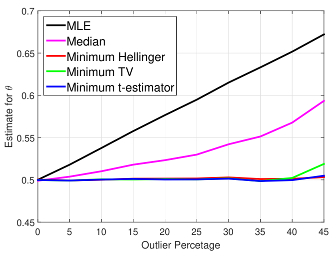

Figure 1: Comparison of different estimates of θ at $\theta=0.5$ in terms of average point estimate under different levels of contamination.

Application to Clustering We use the t-divergence induced loss as opposed to the squared error or Minkowski loss in the Weighted k-means algorithm to justtify its performance in practice. Our experimental results show that using the robust t-divergence induced loss improves the performance of Weighted k-means, even in a high-dimensional setting, where the number of features (p) far exceeds the number of observations (n). Given data points $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in\mathbb{R}^p$, the objective function is thus given by,

$$f(\boldsymbol{\Theta},\boldsymbol{w})=\sum_{i=1}^n\min_{1\le j\le k}\sum_{l=1}^p w_l^\beta(x_{il}-\theta_{jl})\tan^{-1}(x_{il}-\theta_{jl}),\quad \text{subject to } \sum_{l=1}^p w_l=1 \tag{1}$$

A block coordinate descent algorithm is used to minimize the objective function equation 1 (detailed derivation given in the supplementary document), similar in spirit to Lloyd's k-means (Lloyd, 1982).

Table 1: Average ARI values for different peer algorithms on real data benchmarks ($+$ ($\approx$) denotes statistically significant (equivalent) results w.r.t. the best performing algorithm of that row; The last row indicates the average rank in terms of ARI).

Datasets	*k*-means	*Wk*-means	Minkowski	Sparse	*Wk*-means (Huber)	*Wk*-means (*t*)
Wine	$0.364^+(5)$	$0.561^+(4)$	$0.104^+(6)$	$0.806^\approx(2)$	$0.761^\approx(3)$	**0.830**(1)
WBDC	$0.490^+(3)$	$0.013^+(6)$	$0.106^+(5)$	$0.491^+(2)$	$0.486^+(4)$	**0.730**(1)
Lymphoma	$0.394^+(6)$	$0.768^+(4)$	$0.618^+(5)$	$0.848^+(2)$	$0.790^+(3)$	**0.947**(1)
Leukemia	$0.683^+(3)$	$0.213^+(6)$	$0.401^+(5)$	$0.727^+(2)$	$0.581^+(4)$	**0.944**(1)
Appendicitis	$0.229^+(4.5)$	$0.213^+(6)$	$0.229^+(4.5)$	$0.446^\approx(2)$	$0.251^+(3)$	**0.452**(1)
Brain	$0.436^+(4)$	$0.432^+(5)$	$0.392^+(6)$	$0.446^+(3)$	$0.451^+(2)$	**0.534**(1)
Colon	$0.016^+(5)$	$0.001^+(6)$	$0.014^+(4)$	$0.088^+(3)$	$0.102^+(2)$	**0.447**(1)
Average Rank	4.36	5.43	5.07	2.29	3	**1**

The study validates the efficacy of a weighted k-means algorithm using t-divergence induced loss on real-life datasets from Arizona State University and UCI Machine Learning Repository. The algorithm's performance was compared to classical k-means, Wk-means (Huang et al., 2005), MW-k-means (De Amorim & Mirkin, 2012), Sparse k-means (Witten & Tibshirani, 2010), and Wk-means with Huber loss using Adjusted Rand Index (ARI) as a performance indicator. Experiments, conducted involved running each algorithm 20 times on the same randomly chosen centroids until convergence. The weighted k-means with t-divergence induced loss showed enhanced performance over peer algorithms, as shown in Table 1, with statistical significance confirmed by Wilcoxon's signed-rank test at a 5% level. This robust approach notably improved the W-k-means algorithm's performance on benchmark datasets.

3 URM Statement

This is to confirm that all the authors of this paper satisfy the URM criteria of the ICLR 2024 tiny papers track.

References

Mark Burgin. *Semitopological Vector Spaces: Hypernorms, Hyperseminorms, and Operators*. CRC Press, 2017.

Saptarshi Chakraborty and Swagatam Das. On the strong consistency of feature-weighted k-means clustering in a nearmetric space. *Stat*, 8(1):e227, 2019.

Saptarshi Chakraborty, Debolina Paul, and Swagatam Das. On consistent entropy-regularized k-means clustering with feature weight learning: Algorithm and statistical analyses. *IEEE Transactions on Cybernetics*, 2022.

Renato Cordeiro De Amorim and Boris Mirkin. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075, 2012.

Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

Debolina Paul and Swagatam Das. A bayesian non-parametric approach for automatic clustering with feature weighting. *Stat*, 9(1):e306, 2020.

Debolina Paul, Saptarshi Chakraborty, and Swagatam Das. On the uniform concentration bounds and large sample properties of clustering with bregman divergences. *Stat*, 10(1):e360, 2021.

Debolina Paul, Saptarshi Chakraborty, Swagatam Das, and Jason Xu. Implicit annealing in kernel spaces: A strongly consistent clustering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5862–5871, 2022.

David Pollard. Strong consistency of k-means clustering. *The annals of statistics*, pp. 135–140, 1981.

Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

A Proof of Properties

Proof of Property 1. *Follows trivially as* $\forall x, y \in \mathbb{R}$, $(x-y)\tan^{-1}(x-y) \geq 0$ *and equality holds iff* $x = y$.

Proof of Property 2. *Let* $p = \frac{dP}{d\mu}$ *and* $q = \frac{dQ}{d\mu}$. *Since we know* $(x-y)\tan^{-1}(x-y) \leq 2\big[(x-z)\tan^{-1}(x-z) + (z-y)\tan^{-1}(z-y)\big]$, $\forall x, y, z \in \mathbb{R}$, *we observe that for all* $x \in \Omega$, $0 \leq (p(x)-q(x))\tan^{-1}(p(x)-q(x)) \leq 2[p(x)\tan^{-1}p(x) + q(x)\tan^{-1}q(x)]$. *Integrating w.r.t.* μ, *we get,*

$$D(P,Q) \leq 2\int (p\tan^{-1}p + q\tan^{-1}q)d\mu < \infty.$$

Proof of Property 3. *If* $P \in D_\mu^p$. *Let* $p = \frac{dP}{d\mu}$. *Then,* $\int p\tan^{-1}p d\mu \leq \frac{\pi}{2}\int p d\mu = \frac{\pi}{2} < \infty$.

Proof of Property 4. *Follows trivially since* x *and* $\tan^{-1}x$ *are both odd functions.*

Proof of Property 5. *We know that for all* $z \in \mathbb{R}$, $|\tan^{-1}(z)| \leq \frac{\pi}{2}$. *Thus, for all* $z \in \mathbb{R}$, $z \leq z\tan^{-1}(z) \leq |\frac{\pi}{2}z|$. *Thus, we have,* $z\tan^{-1}(z) \leq \frac{\pi}{2}|z|$, *for all* $z \in \mathbb{R}$. *Now,* $D(P,Q) = \int(p-q)\tan^{-1}(p-q)d\mu \leq \int \frac{\pi}{2}|p-q|d\mu = \pi\frac{1}{2}\int|p-q|d\mu = \pi TV(P,Q)$.

Proof of Property 6. *We will first show that $D(P_n, P) \to 0$ implies $TV(P_n, P) \to 0$. We fix $\epsilon > 0$. Thus, there exists $N_\epsilon \in \mathbb{N}$, such that $n \geq N_\epsilon$ implies $D(P_n, P) < \epsilon$. For any $\delta > 0$,*

$$\begin{aligned}
\epsilon &> \int (p_n - p)\tan^{-1}(p_n - p)d\mu \\
&= \int_{|p_n-p|>\delta} (p_n - p)\tan^{-1}(p_n - p)d\mu + \int_{|p_n-p|<\delta} (p_n - p)\tan^{-1}(p_n - p)d\mu \\
&\geq \tan^{-1}\delta \int_{|p_n-p|>\delta} |p_n - p|d\mu + \int_{|p_n-p|<\delta} (p_n - p)\tan^{-1}(p_n - p)d\mu \\
&= \tan^{-1}\delta \int |p_n - p|d\mu + \int\limits_{|p_n-p|\leq\delta} [(p_n - p)\tan^{-1}(p_n - p) - \tan^{-1}\delta|p_n - p|]d\mu \\
&\geq \tan^{-1}\delta \int |p_n - p|d\mu - \int\limits_{|p_n-p|\leq\delta} \tan^{-1}\delta|p_n - p|d\mu \\
&\geq \tan^{-1}\delta \int |p_n - p|d\mu - \delta\tan^{-1}\delta.
\end{aligned}$$

Thus, $TV(P_n, P) < \frac{\epsilon}{\tan^{-1}\delta} + \delta$, for all $\delta > 0$. Thus, $TV(P_n, P) \leq \inf_{\delta>0}\left(\frac{\epsilon}{\tan^{-1}\delta} + \delta\right)$, which can be made smaller that η, for any prefixed $\eta > 0$, if ϵ is chosen small enough. Thus, $TV(P_n, P) \to 0$ as $n \to \infty$. Now, let $g : \Omega \to \mathbb{R}$ be any bounded continuous function,

$$|\int gp_n d\mu - \int gp d\mu| \leq \int |g||p_n - p|d\mu \leq \sup_{x\in\Omega}|g(x)|TV(P_n, P) \to 0.$$

Thus, $P_n \xrightarrow{\mathcal{L}} P$, i.e. P_n converges to P in distribution.

Proof of Property 7. *The non-negativity, identity of indiscernibles and symmetry properties of $D(P, Q)$ have been showed before. What remains to show is that $D(P, Q) \leq 2(D(P, Q)+D(R, Q))$ for all $P, Q, R \in D_\mu$. Let $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ and $r = \frac{dR}{d\mu}$. We note that $D(P, Q) = \int(p - q)\tan^{-1}(p-q)d\mu \leq \int 2\left[(p-r)\tan^{-1}(p-r)+(r-q)\tan^{-1}(r-q)\right]d\mu = 2(D(P, Q)+D(R, Q))$.*

B EXISTENCE OF THE t-ESTIMATE

Theorem 2. *Let the parametric family $\mathcal{F}$ be identifiable and let Θ be a compact subset of $\mathbb{R}^p$. We also assume that $f_\theta(\cdot)$ is continuous a.e. $[\mu]$. Then the following holds:*

1. *For all $G \ll \mu$, $T(G)$ exists.*
2. *If $T(G)$ is unique, then $T(\cdot)$ is continuous at G, under the total variation topology, i.e. $T(G_n) \to T(G)$, whenever $\int |g_n - g|d\mu \to 0$. Here $g_n = \frac{dG_n}{d\mu}$.*
3. *$T(F_\theta) = \theta$ for all $\theta \in \Theta$.*

Proof. **Proof of part (1)**: Let $t_n \to t$ be a sequence of parameter values in Θ. Let $h(t) = D(G\|F_t)$. We observe that:

$$\begin{aligned}
|D(G\|F_{t_n}) - D(G\|F_t)| &= \left|\int \left[(g - f_{t_n})\tan^{-1}(g - f_{t_n}) - (g - f_t)\tan^{-1}(g - f_t)\right]d\mu\right| \\
&\leq \int \left|(g - f_{t_n})\tan^{-1}(g - f_{t_n}) - (g - f_t)\tan^{-1}(g - f_t)\right|d\mu. \qquad (2)
\end{aligned}$$

We note that,

$$\begin{aligned}
\left|(g - f_{t_n})\tan^{-1}(g - f_{t_n}) - (g - f_t)\tan^{-1}(g - f_t)\right| &\leq |(g - f_{t_n})\tan^{-1}(g - f_{t_n})| + |(g - f_t)\tan^{-1}(g - f_t)| \\
&\leq \frac{\pi}{2}\left[|(g - f_{t_n})| + |(g - f_t)|\right] \\
&\leq \frac{\pi}{2}\left[2g + f_{t_n} + f_t\right]. \qquad (3)
\end{aligned}$$

We note that the LHS of equation equation 3 is μ-integrable and thus by simple application of Dominated Convergence Theorem (DCT) the RHS of equation 2 converges to 0 as $n \to \infty$. Hence $h(t)$ is continuous on Θ, which is compact. Hence $h(\cdot)$ attains its minimum on Θ.

Proof of part (2): Let $\{G_n\}_{n\geq 1}$ converges to G in total variation sense, i.e. $\int |g_n(x) - g(x)| d\mu(x) \to 0$, as $n \to \infty$. We define $h_n(t) = D(G_n \| F_t)$. We also assume that $\theta_n = T(G_n)$ and $\theta = T(G)$ are also defined uniquely. We observe the following,

$$\begin{aligned} |h_n(t) - h(t)| &= \left| \int \left[(g_n - f_t) \tan^{-1}(g_n - f_t) - (g - f_t) \tan^{-1}(g - f_t) \right] d\mu \right| \\ &= \int \left| \tan^{-1}(\xi_x) + \frac{\xi_x}{1 + \xi_x^2} \right| |g_n - g| d\mu \\ &\leq \left(\frac{\pi}{2} + 1 \right) \int |g_n - g| d\mu \end{aligned} \tag{4}$$

Equation equation 4 follows from applying first order Taylor's expansion on the function $x \tan^{-1}(x)$. Here ξ_x lies between $g_n(x) - f_t(x)$ and $(g(x) - f_t(x))$. From the above calculations we conclude that $\lim_{n\to\infty} \sup_{t\in\Theta} |h_n(t) - h(t)| = 0$. From the definition of θ_n and θ, we observe that:

$$\begin{aligned} & |h(\theta_n) - h(\theta)| \\ =& h(\theta_n) - h(\theta) \\ =& (h(\theta_n) - h_n(\theta_n)) + (h_n(\theta_n) - h_n(\theta)) + (h_n(\theta) - h(\theta)) \\ \leq& (h(\theta_n) - h_n(\theta_n)) + (h_n(\theta) - h(\theta)) & (5) \\ \leq& 2 \sup_{t\in\Theta} |h_n(t) - h(t)| & (6) \\ & \longrightarrow 0 \text{ as } n \longrightarrow \infty \end{aligned}$$

Equation equation 5 follows from the fact that, since θ_n is the minimiser if $h_n(\cdot)$, $h_n(\theta_n) \leq h_n(\theta)$. Thus, we get, $\lim_{n\to\infty} h(\theta_n) = h(\theta)$. We will now show that $\theta_n \to \theta$. We assume the contrary. Suppose $\theta_n \not\to \theta$. We note that $\{\theta_n\}_{n\geq 1}$ is a sequence in the compact set Θ. Thus it has a converging sub-sequence, say, $\{\theta_{n_l}\}_{l\geq 1}$, such that $\theta_{n_l} \to \theta_1$, where, $\theta_1 \neq \theta$. By the continuity of $h(\cdot)$, $h(\theta_{n_l}) \to h(\theta_1)$. This implies that $h(\theta) = h(\theta_1)$, since there cannot be two limit for the converging sequence $h(\theta_{n_l})$. Thus, $h(\theta) = h(\theta_1)$ gives us a contradiction, since, $T(G)$ is assumed to be unique. Thus, $\theta_n \to \theta$.

Proof of part (3): Since the parametric family $\mathcal{F}$ is identifiable, $D(F_\theta \| F_t) = 0$, only at the value $t = \theta$. Thus $T(F_\theta) = \theta$, uniquely. □

C Proof of Theorem 1

We first derive the estimating equation for a minimum t-estimator. Let $\theta_0 = T(G)$ and θ_0 is an interior point of Θ, then θ_0 satisfies the following equation

$$\left[\frac{\partial}{\partial\theta} \int (f_\theta(x) - g(x)) \tan^{-1}(f_\theta(x) - g(x)) d\mu(x) \right] \Bigg|_{\theta=\theta_0} = 0.$$

Assuming the differentiability under the integral sign, we get,

$$\left[\int \left[\tan^{-1}(f_\theta(x) - g(x)) + \frac{f_\theta(x) - g(x)}{1 + (f_\theta(x) - g(x))^2} \right] \frac{\partial f_\theta(x)}{\partial\theta} d\mu(x) \right] \Bigg|_{\theta=\theta_0} = 0.$$

Thus, θ_0 satisfies the following equation.

$$\int \rho(f_\theta(x) - g(x)) f_\theta(x) u_\theta(x) d\mu(x) = 0. \tag{7}$$

Here $\rho(x) = \tan^{-1}(x) + \frac{x}{1+x^2}$ and $u_\theta(x) = \frac{\partial}{\partial\theta} \log f_\theta(x)$. Equation equation 7 gives the estimating equation for minimum t-estimator.

We will make the following technical assumptions.

A1 The support of F_θ is independent of θ.

A2 There exists $\eta > 0$ such that $T(G_\epsilon)$ is an interior point of Θ, for all $\epsilon < 0$ and for all $y \in \mathbb{R}$. Here $G_\epsilon = (1-\epsilon)G + \epsilon\delta_y$.

A3 $T(\cdot)$ is Gateaux differentiable at G.

A4 The derivative on the Left hand side of equation 7 is permitted under the integral sign.

Let $G_\epsilon = (1-\epsilon)G + \epsilon\Delta_y$. Here Δ_y denotes the degenerate distribution at y. We take μ to be the counting measure. Let $\theta_0 = T(G)$ and $\theta_\epsilon = T(G_\epsilon)$ be defined uniquely for all $\epsilon \geq 0$. The value of the influence function at y is given by $IF(y) = \frac{\partial\theta_\epsilon}{\partial\epsilon}\Big|_{\epsilon=0}$. Observe that $g_\epsilon = (1-\epsilon)g + \epsilon\delta_y$ is the density of G_ϵ w.r.t. μ. Here $\delta_y(x) = 1$ if $x = y$ and is 0, otherwise. Before we proceed, we observe that $\rho'(x) = \frac{2}{(1+x^2)^2}$. From the estimating equation equation 7, we observe that

$$\begin{aligned} &\sum_{x\in\mathcal{X}} \rho(f_{\theta_\epsilon}(x) - g_\epsilon(x)) f_{\theta_\epsilon}(x) u_{\theta_\epsilon}(x) = 0 \\ \implies &\sum_{x\in\mathcal{X}} \rho(f_{\theta_\epsilon}(x) - (1-\epsilon)g(x) - \epsilon\delta_y(x)) f_{\theta_\epsilon}(x) u_{\theta_\epsilon}(x) = 0 \end{aligned}$$

Differentiating both sides w.r.t. ϵ and assuming that the derivative can be passed inside the summation, we get,

$$\begin{aligned} \sum_{x\in\mathcal{X}} \Bigg[&\frac{2}{(1+(f_{\theta_\epsilon}(x) - g_\epsilon(x))^2)^2} (f'_{\theta_\epsilon}(x)\theta_\epsilon{}' + g(x) - \delta_y(x)) f_{\theta_\epsilon}(x) u_{\theta_\epsilon}(x) \\ &+ \rho(f_{\theta_\epsilon}(x) - g_\epsilon(x)) f'_{\theta_\epsilon}(x)\theta_\epsilon{}' u_{\theta_\epsilon}(x) + \rho(f_{\theta_\epsilon}(x) - g_\epsilon(x)) f_{\theta_\epsilon}(x) u'_{\theta_\epsilon}(x)\theta_\epsilon{}' \Bigg] = 0. \end{aligned}$$

Substituting $\epsilon = 0$ in the above equation, we get,

$$\begin{aligned} \sum_{x\in\mathcal{X}} \Bigg[&\frac{2}{(1+(f_{\theta_0}(x) - g(x))^2)^2} (f'_{\theta_0}(x) IF(y) + g(x) - \delta_y(x)) f_{\theta_0}(x) u_{\theta_0}(x) \\ &+ \rho(f_{\theta_0}(x) - g(x)) f'_{\theta_0}(x) IF(y) u_{\theta_0}(x) + \rho(f_{\theta_0}(x) - g(x)) f_{\theta_0}(x) u'_{\theta_0}(x) IF(y) \Bigg] = 0. \end{aligned}$$

Thus,

$$\begin{aligned} IF(y) = &\sum_{x\in\mathcal{X}} \left[\frac{2f'_{\theta_0}(x) f_{\theta_0}(x) u'_{\theta_0}(x)}{(1+(f_{\theta_0}(x) - g(x))^2)^2} + \rho(f_{\theta_0}(x) - g(x)) f'_{\theta_0}(x) u_{\theta_0}(x) + \rho(f_{\theta_0}(x) - g(x)) f_{\theta_0}(x) u'_{\theta_0}(x) \right] \\ = &\sum_{x\in\mathcal{X}} \frac{2(\delta_y(x) - g(x)) f_{\theta_0}(x) u_{\theta_0}(x)}{(1+(f_{\theta_0}(x) - g(x))^2)^2} \\ = &\frac{2f_{\theta_0}(y) u_{\theta_0}(y)}{(1+(f_{\theta_0}(y) - g(y))^2)^2} - \sum_{x\in\mathcal{X}} \frac{g(x) f_{\theta_0}(x) u_{\theta_0}(x)}{(1+(f_{\theta_0}(x) - g(x))^2)^2}. \end{aligned}$$

simplifying the above equation, we get,

$$IF(y) = \frac{\frac{2f_{\theta_0}(y) u_{\theta_0}(y)}{(1+(f_{\theta_0}(y) - g(y))^2)^2} - \sum_{x\in\mathcal{X}} \frac{g(x) f_{\theta_0}(x) u_{\theta_0}(x)}{(1+(f_{\theta_0}(x) - g(x))^2)^2}}{\sum_{x\in\mathcal{X}} \frac{2f^2_{\theta_0}(x) u_{\theta_0}(x) u'_{\theta_0}(x)}{(1+(f_{\theta_0}(x) - g(x))^2)^2} + \rho(f_{\theta_0}(x) - g(x)) f_{\theta_0}(x)(u^2_{\theta_0}(x) + u'_{\theta_0}(x))}.$$

D APPLICATION TO CLUSTERING

D.1 CLUSTERING ALGORITHM

The proposed clustering algorithm with t-divergence has been proposed at Algorithm 1.

Algorithm 1 Weighted k-means with t-divergence induced loss

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\beta > 1$. **Output**: Cluster assignment matrix $\boldsymbol{U}$, feature weight vector $\boldsymbol{w}$.

repeat

Step 1: Update $\boldsymbol{U}$ by $u_{ij}^{(t+1)} \leftarrow \begin{cases} 1 & \text{if } j = \arg\min\limits_{1 \leq j' \leq k} \sum_{l=1}^{p} w_l^{(t)\beta}(x_{il} - \theta_{j'l}^{(t)}) \tan^{-1}(x_{il} - \theta_{j'l}^{(t)}), \\ 0 & \text{Otherwise.} \end{cases}$

Step 2: Update $\boldsymbol{\Theta}$ by taking $\theta_{jl}^{(t+1)} \leftarrow \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} u_{ij}^{(t+1)}(x_{il} - \theta) \tan^{-1}(x_{il} - \theta)$ by the Newton-Raphson method.

Setp 3: Update $\boldsymbol{w}$ by taking $w_l^{(t+1)} \leftarrow \frac{1/D_l^{(\beta-1)}}{\sum_{m=1}^{p} 1/D_m^{(\beta-1)}}$, where $D_l = \sum_{i=1}^{n} u_{ij}^{(t+1)}(x_{il} - \theta_{jl}^{(t+1)}) \tan^{-1}(x_{il} - \theta_{jl}^{(t+1)})$.

until objective equation 1 converges.

D.2 STRONG CONSISTENCY

The proposed t-divergence based clustering enjoy elegant theoretical properties such as strong consistency under general assumptions. This property can be guaranteed through standard tools available in the literature (Pollard, 1981; Paul & Das, 2020; Paul et al., 2021; 2022; Chakraborty et al., 2022). We note that since the t-divergence induced loss is a near-metric property, the strong consistency of the (global) minimizers of equation 1 can be assessed through the works of Chakraborty & Das (2019). We assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independently and identically distributed according to the distribution $\mathbb{P}$. We assume that $\mathbb{P}$ has finite first moment, i.e. $\mathbb{E}(\|\boldsymbol{X}_1\|_1) < \infty$. It is easy to see that the t-divergence induced loss satisfies all the assumptions A1-A6 of (Chakraborty & Das, 2019). A7 can be assessed by observing that $\int \sum_{l=1}^{p}(x_l - \theta_l) \tan^{-1}(x_l - \theta_l) d\mathbb{P} \leq \frac{\pi}{2} \int \sum_{l=1}^{p} |x_l - \theta_l| d\mathbb{P} \leq \frac{\pi}{2} \int \sum_{l=1}^{p}(|x_l| + |\theta_l|) d\mathbb{P} = \frac{\pi}{2}(\mathbb{E}(\|\boldsymbol{X}_1\|_1) + \|\boldsymbol{\theta}\|_1) < \infty$. Thus, we have the following theorem guaranteeing the strong consistency of the W-k-means algorithm under the t-loss.