Convex Clustering Redefined: Robust Learning with the Median of Means Estimator

Sourav De[1*], Koustav Chowdhury[1*], Bibhabasu Mandal[1*], Sagar Ghosh[2], Swagatam Das[3], Debolina Paul[4], Saptarshi Chakraborty[5]

A Supplementary Material

A.1 ADAM update rule for optimization

For a chosen $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ the update rule to be followed is as follows:

$$\hat{\mathbf{m}}_i^{(t)} = \frac{\mathbf{m}_i^{(t)}}{1-\boldsymbol{\beta}_1^t},$$

$$\hat{\mathbf{v}}_i^{(t)} = \frac{\mathbf{v}_i^{(t)}}{1-\boldsymbol{\beta}_2^t},$$

$$\mathbf{u}_i^{(t+1)} = \mathbf{u}_i^{(t)} - \frac{\alpha \hat{\mathbf{m}}_i^{(t)}}{\sqrt{\hat{\mathbf{v}}_i^{(t)}+\epsilon}}.$$

A.2 Proof of Theorem 1

Theorem 1. *Suppose $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{nd}$ is a vector of independent bounded random variables, with mean 0, covariance matrix $\sigma^2 \boldsymbol{I}$ and $|\epsilon_i| \leq M$, for all $i = 1, \ldots, nd$. Suppose that $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{I}}_{B_{l_t}}$ are obtained from minimizing equation (10), then if $\gamma' \geq \frac{M}{ndb\sqrt{n}}$ the following holds with probability at least 1 - δ:*

$$\frac{1}{2ndb}\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}} \leq M^2 \left(\frac{\sqrt{db/n}+d}{\sqrt{ndb}} + c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)} + c\frac{\log\left(\frac{1}{\delta}\right)}{ndb} \right) + \gamma'\frac{|\mathcal{E}|}{4} + \gamma' \left[\sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2 + \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2^2 \right]$$

Proof. Let $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}_{\boldsymbol{\beta}}^\top$ be the singular value decomposition (SVD) of $\boldsymbol{D}$, where $\boldsymbol{V}_{\boldsymbol{\beta}} \in \mathbb{R}^{nd\times(n-1)d}$. We construct $\boldsymbol{V}_{\boldsymbol{\alpha}} \in \mathbb{R}^{nd\times d}$ such that $\boldsymbol{V} = [\boldsymbol{V}_{\boldsymbol{\alpha}}, \boldsymbol{V}_{\boldsymbol{\beta}}]$ is an $nd \times nd$ orthonormal matrix.

Next, define $\boldsymbol{\beta} = \boldsymbol{V}_{\boldsymbol{\beta}}^\top \boldsymbol{u}$, $\boldsymbol{\alpha} = \boldsymbol{V}_{\boldsymbol{\alpha}}^\top \boldsymbol{u}$ and $\gamma' = \frac{\gamma}{2nd}$. Also $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Lambda}$ and $\boldsymbol{Z}^-$ is the left inverse of $\boldsymbol{Z}$. Thus, the optimization problem becomes:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{I}_{B_{l_t}}} \frac{1}{2ndb}\left(\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right)^\top \boldsymbol{I}_{B_{l_t}} \left(\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right) + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2 \tag{1}$$

Now, let $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{I}}_{B_{l_t}}$ be the minimiser of the above cost function. Then, by definition,

$$\begin{aligned}
&\frac{1}{2ndb}\left\|\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}} - \boldsymbol{V}_{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}} + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\hat{\boldsymbol{\beta}}\right\|_2^2 \\
&\leq \frac{1}{2ndb}\left\|\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right\|^2_{\boldsymbol{I}_{B_{l_t}}} + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2 \\
&\leq \frac{1}{2ndb}\left\|\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}} + \frac{1}{2ndb}\boldsymbol{\epsilon}^\top(\boldsymbol{I}_{B_{l_t}} - \hat{\boldsymbol{I}}_{B_{l_t}})\boldsymbol{\epsilon} + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2 \\
&\leq \frac{1}{2ndb}\left\|\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{\beta}\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}} + \frac{1}{n}M^2 + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2
\end{aligned}$$

On further simplification, we get the following,

$$\frac{1}{2ndb}\left\|\boldsymbol{V}_{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) + \boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}} + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\hat{\boldsymbol{\beta}}\right\|_2^2 \leq \frac{1}{ndb}G\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{I}}_{B_{l_t}}\right) + \frac{1}{n}M^2 + \gamma' \sum_{(i,j)\in\mathcal{E}} \left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2$$

where $G(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{I}}_{B_{l_t}}) = \boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}(\boldsymbol{V}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \boldsymbol{V}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))$. Since $\hat{\boldsymbol{\alpha}}$ is the minimiser, we can choose $\hat{\boldsymbol{\alpha}}$ such that $\boldsymbol{x} - \boldsymbol{V}_{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}} - \boldsymbol{V}_{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}} = 0$. Therefore, $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top \boldsymbol{\epsilon}$. Now, we can bound,

[*]These authors contributed equally.

$$
\begin{aligned}
&\frac{1}{ndb}\left|G(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{I}}_{B_{l_t}})\right| \\
&= \frac{1}{ndb}\left|\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right)+\boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right)\right| \\
&= \frac{1}{ndb}\left|\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}+\boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right)\right| \\
&\leq \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}+\frac{1}{ndb}\left|\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right| \\
&= \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}+\frac{1}{ndb}\left|\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{Z}^{-}\boldsymbol{Z}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right| \\
&= \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon} \\
&\quad+\frac{1}{ndb}\left|\sum_{(i,j)\in\mathcal{E}}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{Z}^{-}_{\mathcal{C}(i,j)}\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right| \\
&\leq \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon} \\
&\quad+\frac{1}{ndb}\sum_{(i,j)\in\mathcal{E}}\left|\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{Z}^{-}_{\mathcal{C}(i,j)}\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right| \\
&\leq \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon} \\
&\quad+\frac{1}{ndb}\sum_{(i,j)\in\mathcal{E}}\left(\left\|(\boldsymbol{Z}^{-}_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_2\right. \\
&\qquad\qquad\left.\cdot\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2\right) \\
&\leq \frac{1}{ndb}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon} \\
&\quad+\frac{1}{ndb}\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^{-}_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_2 \\
&\qquad\qquad\cdot\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2
\end{aligned}
$$

Next, we derive high-probability bounds for the terms $\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}$ and $\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^{-}_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_2$. Now, using the Hanson Wright Inequality,

$$
\begin{aligned}
&\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon} \\
&\leq \mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\right] \\
&\quad+c\left(M\sqrt{r}\,\mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\left\|\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\right\|_{sp}\right]\right) \\
&\quad+c\left(rM^2\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\left\|I_{B_{l_t}}\right\|\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\operatorname{tr}\left(\boldsymbol{\epsilon}^\top \hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\right)\right] \\
&\quad+c\left(M\sqrt{r}\,\mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\left\|\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\right\|_{sp}\|\boldsymbol{\epsilon}\|_2\right]\right) \\
&\quad+c\left(rM^2\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}1\right) \\
&= \mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\operatorname{tr}\left(\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)\right] \\
&\quad+c\left(M\sqrt{r}\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\left\|\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\right\|_{sp}\mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2\right]+rM^2\right) \\
&\leq \mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\sqrt{\operatorname{tr}\left(\hat{\boldsymbol{I}}^2_{B_{l_t}}\right)}\cdot\right. \\
&\qquad\left.\sqrt{\operatorname{tr}\left(\left(\boldsymbol{V}_\alpha\boldsymbol{V}_\alpha^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)^\top\left(\boldsymbol{V}_\alpha\boldsymbol{V}_\alpha^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)\right)}\right] \\
&\qquad+c\left(M\sqrt{r}\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\left\|\hat{\boldsymbol{I}}_{B_{l_t}}\right\|_{sp}\left\|\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\right\|_{sp}\mathbb{E}\left[M\sqrt{nd}\right]\right) \\
&\qquad+crM^2 \\
&= \mathbb{E}\left[\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\sqrt{db}\sqrt{\operatorname{tr}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)}\right] \\
&\qquad+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= \sqrt{db}\,\mathbb{E}\left[\sqrt{\operatorname{tr}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)}\right]+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= \sqrt{db}\,\mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)}\right]+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&\leq Md\sqrt{nb}\,\mathbb{E}\left[\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)}\right]+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&\leq Md\sqrt{nb}\sqrt{\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)\right]}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= Md\sqrt{nb}\sqrt{\operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right]\right)}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= Md\sqrt{nb}\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right]\right)}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= Md\sqrt{nb}\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top(\sigma^2 I)\right)}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= Md\sigma\sqrt{nb}\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\right)}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= M^2d\sqrt{nb}\sqrt{\operatorname{tr}\left(\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{V}_{\boldsymbol{\alpha}}\right)}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= M^2d\sqrt{ndb}+c\left(M^2\sqrt{ndr}+rM^2\right) \\
&= M^2\left(d\sqrt{ndb}+c\sqrt{ndr}+cr\right)
\end{aligned}
$$

Thus, from the above analysis, we get

$$\mathbb{P}\left(\frac{1}{ndb}\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\boldsymbol{\epsilon}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\right.$$
$$\left.\geq M^2\left(\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{r}+\frac{cr}{ndb}\right)\right)\leq e^{-r}$$

Taking $r=\log\left(\frac{1}{\delta}\right)$, we get,

$$\mathbb{P}\left(\frac{1}{ndb}\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\boldsymbol{\epsilon}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\right.$$
$$\left.\geq M^2\left(\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\right)\leq\delta$$

Thus, with probability atleast $1-\delta$,

$$\frac{1}{ndb}\boldsymbol{\epsilon}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}\leq\frac{1}{ndb}\sup_{\boldsymbol{I}_{B_{l_t}}\in\mathcal{I}}\boldsymbol{\epsilon}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{V}_{\boldsymbol{\alpha}}{}^\top\boldsymbol{\epsilon}$$
$$\leq M^2\left(\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)$$

Let $y_j=\boldsymbol{e}_j^\top(\boldsymbol{Z}^-_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}$. Now, y_j is a univariate, bounded random variable with $|y_j|\leq\frac{M}{\sqrt{n}}$. Thus,

$$\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^-_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_\infty=\max_j|y_j|\leq\frac{M}{\sqrt{n}}$$
$$\Rightarrow\frac{1}{ndb}\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^-_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_2$$
$$\leq\frac{1}{ndb}\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^-_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_\infty\leq\frac{M}{ndb\sqrt{n}}$$

Note $\gamma'\geq\frac{M}{ndb\sqrt{n}}$ implies that,

$$\gamma'\geq\frac{1}{ndb}\max_{(i,j)\in\mathcal{E}}\left\|(\boldsymbol{Z}^-_{\mathcal{C}(i,j)})^\top\boldsymbol{V}_{\boldsymbol{\beta}}{}^\top\hat{\boldsymbol{I}}_{B_{l_t}}\boldsymbol{\epsilon}\right\|_2.$$

So we get,

$$\frac{1}{ndb}\left|G(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{I}}_{B_{l_t}})\right|\leq M^2\left(\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}\right.$$
$$\left.+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)+\gamma'\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2$$

holds with probability atleast $1-\delta$. Now combining all the results we get, with probability atleast $1-\delta$

$$\frac{1}{2ndb}\left\|\boldsymbol{V}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}+\gamma'\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\hat{\boldsymbol{\beta}}\right\|_2^2$$
$$\leq M^2\left(\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)+\frac{1}{n}M^2$$
$$+\gamma'\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_2+\gamma'\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2$$

Upon rearranging the terms, we obtain that with probability atleast $1-\delta$,

$$\frac{1}{2ndb}\left\|\boldsymbol{V}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\boldsymbol{V}_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}$$
$$\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)$$
$$+\gamma'\left[\sum_{(i,j)\in\mathcal{E}}\left(\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\hat{\boldsymbol{\beta}}\right\|_2-\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\hat{\boldsymbol{\beta}}\right\|_2^2\right)\right.$$
$$\left.+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2\right]$$
$$\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)$$
$$+\gamma'\left[\frac{|\mathcal{E}|}{4}+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{Z}_{\mathcal{C}(i,j)}\boldsymbol{\beta}\right\|_2^2\right]$$
$$\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)$$
$$+\gamma'\left[\frac{|\mathcal{E}|}{4}+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2^2\right]$$

□

A.3 Proof of Corollary 1

Corollary 1. *Suppose* $\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2\leq C$, *for all* $1\leq i,j\leq n$, *for some constant* C, $|\mathcal{E}|\leq kn$ *and* $\gamma'\geq\frac{M}{ndb\sqrt{n}}$. *If* $d=o(n)$, *then* $\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}\xrightarrow{p}0$ *as* $n,d\to\infty$.

Proof. For any fixed δ, we know from Theorem 1 that with probability atleast $1-\delta$.

$$
\begin{aligned}
&\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{|\mathcal{E}|}{4}+\gamma'\left[\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2^2\right]\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{|\mathcal{E}|}{4}+\left(C+C^2\right)\gamma'|\mathcal{E}|\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{kn}{4}+\left(C+C^2\right)\gamma' kn\to 0\quad\text{as}\quad n,d\to\infty.
\end{aligned}
$$

Thus, for any fixed $\epsilon>0$, $P\left(\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}>\epsilon\right)\leq\delta$ as $n,p\to\infty$.

Hence, $\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}\xrightarrow{p}0$.

□

A.4 Proof of Corollary 2

Corollary 2. *Suppose* $\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2\leq C$, *for all* $1\leq i,j\leq n$, *for some constant* C, $|\mathcal{E}|\leq kn$ *and* $\gamma'\geq\frac{M}{ndb\sqrt{n}}$. *Then* $\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}=O\left(\frac{1}{\sqrt{n}}\right)$.

Proof. For any fixed δ, we know from Theorem 1 that with probability atleast $1-\delta$.

$$
\begin{aligned}
&\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{|\mathcal{E}|}{4}+\gamma'\left[\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2+\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{D}_{\mathcal{C}(i,j)}\boldsymbol{u}\right\|_2^2\right]\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{|\mathcal{E}|}{4}+\left(C+C^2\right)\gamma'|\mathcal{E}|\\
&\leq M^2\left(\frac{1}{n}+\frac{d}{\sqrt{ndb}}+c\frac{1}{b\sqrt{nd}}\sqrt{\log\left(\frac{1}{\delta}\right)}+c\frac{\log\left(\frac{1}{\delta}\right)}{ndb}\right)\\
&\quad+\gamma'\frac{kn}{4}+\left(C+C^2\right)\gamma' kn\\
&=O\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

Thus, $\sqrt{n}\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}=O(1)$. This implies that there exists a constant ϵ such that, $P\left(\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}\leq\epsilon\right)\geq 1-\delta$, for all $n\in\mathbb{N}$. Hence, $\sqrt{n}\frac{1}{2ndb}\|\hat{\boldsymbol{u}}-\boldsymbol{u}\|^2_{\hat{\boldsymbol{I}}_{B_{l_t}}}$ is tight.

□

A.5 Derivation of Time Complexity

The main loop in Algorithm 1 performs N iterations. It is enough to calculate the complexity of each step inside this loop. Constructing a random partition of l buckets of b points takes $O(lb)=O(n)$ steps. Calculating $MoM_B(\mathbf{U})$ for each bucket takes $O(bd)$ steps. Therefore, finding the median bucket B_{l_t}, takes $O(lbd)=O(nd)$ steps. Calculating each g_i requires checking whether the index i is in B_{l_t}, which takes at most $O(b)$ checks, evaluating $(\mathbf{u}_i^{(t)}-\mathbf{x}_i)$, which takes $O(d)$ constant time evaluations, and evaluating $\sum_j w_{ij}(\mathbf{u}_i^{(t)}-\mathbf{u}_j^{(t)})\mathbb{1}(||\mathbf{u}_i^{(t)}-\mathbf{u}_j^{(t)}||_2^2<\mu)$, which takes $O(kd)$ constant-time evaluations. Therefore, calculating all g_i's require at most $O(n(b+d+kd))=O(nkd)$ steps. Calculating each m_i, v_i and u_i takes $O(d)$ constant-time evaluations. Therefore, calculating all m_i's, v_i's, and u_i's require $O(nd)$ constant-time evaluations. Hence, the per-iteration complexity of the main loop is $O(nkd)$. Therefore, the complexity of Algorithm 1 is $O(Nnkd)$.

A.6 Description of datasets

Table 1 contains the name and description of all 18 datasets that were used in the experiments.

A.7 Synthetic Data

In this section, we get empirical results on 3 simulated datasets. We contaminate every dataset by adding noise points following a specific uniform distribution. The contamination levels are: 0%, 5%, 10%, 15% and 20%.

- **Blobs**: 3 blobs of data points containing 500 points each simulated from bi-variate Gaussian distribution with different means and covariance matrix as the $\mathbf{I}_{2\times 2}$. Noise is simulated uniformly from the smallest enclosing axis-parallel rectangle.
- **Circles**: This dataset contains a large circle of radius 1 (consisting of 350 points) containing a smaller circle of radius 0.25 (consisting of 350 points) in two-dimensional space, forming a non-linearly separable pattern. Each point in the dataset is assigned to one of two classes, representing the respective circle to which it belongs. We artificially contaminate the dataset by introducing random

	Datasets	Type	No. of Data-points	Input Dimension	No. of Clusters
1.	Iris	Real	150	4	3
2.	Newthyroid	Real	215	5	3
3.	Ecoli	Real	336	7	8
4.	Wisconsin	Real	683	9	2
5.	Wine	Real	178	13	3
6.	Zoo	Real	101	16	7
7.	Dermatology	Real	358	34	6
8.	Brain	Real	42	5597	5
9.	Lung	Real	203	3312	5
10.	Lymphoma (bio)	Real	96	4026	9
11.	Coil 20	Real	1440	1024	20
12.	Wdbc	Real	569	30	2
13.	Lung-discrete	Real	73	325	7
14.	ORLRaws10p	Real	100	10304	10
15.	Lymphoma (microarray)	Real	62	4026	3
16.	Blobs	Simulated	1500	2	3
17.	Circles	Simulated	700	2	2
18.	Moons	Simulated	1500	2	2

Table 1: List of Datasets

Dataset	Index	KM	MKM	CC	RCC	RConv	RBKM	COMET
Iris (k = 3)	k^*	3.21±1.01	2.91±0.79	2.57±0.51	147.90±0.70	3.71±0.47	2.90±0.80	3.21±0.43
	ARI	0.56±0.06$^{\sim}$	**0.59±0.06**	0.55±0.01†	0.001±0.00†	0.46±0.06†	**0.59±0.06**	0.55±0.038$^{\sim}$
	AMI	0.66±0.05$^{\sim}$	0.67±0.04$^{\sim}$	**0.71±0.01**	0.006±0.002†	0.61±0.03†	0.67±0.04$^{\sim}$	0.65±0.03†
Ecoli (k = 8)	k^*	3.20±2.24	1.92±1.82	34.21±1.25	334.80±2.56	16.86±2.66	2.00±0.00	10.6±1.95
	ARI	0.27±0.23†	0.10±0.17†	0.47±0.06$^{\sim}$	0.00±0.00†	**0.51±0.04**	0.04±0.01†	0.46±0.04$\sim$
	AMI	0.25±0.21†	0.09±0.16†	**0.46±0.02**	0.002±0.004†	0.46±0.01$^{\sim}$	0.09±0.01†	0.43±0.02$^{\sim}$
Zoo (k = 7)	k^*	4.43±2.39	4.37±1.87	16.78±2.26	66.93±3.95	6.00±0.88	2.00±0.00	6.57±0.85
	ARI	0.53±0.22†	0.64±0.23†	0.71±0.10†	0.11±0.02†	**0.89±0.09**	0.03±0.04†	0.85±0.02$^{\sim}$
	AMI	0.61±0.21†	0.69±0.21†	0.75±0.05†	0.27±0.03†	0.84±0.06$^{\sim}$	0.05±0.02†	**0.85±0.02**
Brain (k = 5)	k^*	5±1.92	5±1.49	18±0.00	42±0.00	5±0.36	2±0.50	4±0.00
	ARI	0.26±0.10†	0.26±0.10†	0.64±0.02$^{\sim}$	0.00±0.00†	0.56±0.06†	0.016±0.02†	**0.66±0.03**
	AMI	0.33±0.10†	0.27±0.11†	0.62±0.03†	0.00±0.00†	0.62±0.05†	0.03±0.04†	**0.72±0.03**
Lung (k = 5)	k^*	4.21±1.63	4.41±1.64	23.86±0.36	7.40±2.16	3.00±0.00	1.04±0.20	5.00±0.00
	ARI	0.39±0.20†	**0.53±0.19**	0.35±0.004†	0.31±0.13†	0.35±0.003†	0.00±0.00†	0.36±0.003†
	AMI	0.49±0.20†	**0.55±0.17**	0.35±0.001†	0.22±0.09†	0.37±0.001†	0.00±0.00†	0.54±0.01$^{\sim}$
Lymphoma(bio) (k = 9)	k^*	4.28±1.61	6.32±1.41	96±0.00	96±0.00	6.57±0.51	2±0.00	6.93±0.82
	ARI	0.37±0.12†	**0.49±0.09**	0.00±0.00†	0.00±0.00†	0.26±0.09†	0.06±0.06†	0.41±0.02†
	AMI	0.47±0.08†	**0.56±0.07**	0.00±0.00†	0.00±0.00†	0.45±0.08†	0.09±0.06†	0.51±0.01$^{\sim}$
Coil20 (k = 20)	k^*	8.66±3.72	5.62±1.69	90.07±5.01	1440±0.00	20±0.00	1.31±0.47	19±0.00
	ARI	0.30±0.11†	0.21±0.05†	0.69±0.001†	0.00±0.00†	**0.82±0.00**	0.00±0.00†	0.80±0.00†
	AMI	0.58±0.09†	0.50±0.07†	0.86±0.00†	0.00±0.00†	**0.93±0.00**	0.00±0.00†	0.92±0.00†
Wdbc (k = 2)	k^*	1±0.00	1±0.00	136.43±5.65	569±0.00	6.78±0.58	2±0.00	2.21±0.42
	ARI	0.00±0.00†	0.00±0.00†	**0.38±0.03**	0.00±0.00†	0.09±0.01†	0.001±0.002†	0.17±0.31†
	AMI	0.00±0.00†	0.00±0.00†	**0.24±0.01**	0.00±0.00†	0.08±0.01†	0.01±0.003†	0.14±0.26†
Lymphoma (Microarray) (k = 3)	k^*	2.44±1.63	2.23±1.48	3±0.00	3±0.85	2.43±0.51	1.06±0.25	3±0.00
	ARI	0.22±0.27†	0.22±0.28†	0.79±0.00†	**0.86±0.03**	0.33±0.42†	0.00±0.011†	0.79±0.00†
	AMI	0.26±0.29†	0.27±0.31†	0.71±0.00†	**0.81±0.03**	0.29±0.37†	0.00±0.001†	0.71±0.00†

Table 2: Results for Real Life Datasets

† : significantly different from the best performing algorithm , $^{\sim}$: statistically same as the best performing algorithm .

points in the middle-annulus to understand which algorithm best separates the two clusters.

- **Moons**: The two moons dataset is a synthetic dataset consisting of two interleaving crescent-shaped clusters resembling two half-moons. The data contains 1500 two-dimensional points, and the two clusters are non-linearly separable. Noise is simulated uniformly from the smallest enclosing axis-parallel rectangle.

Discussion: As shown in Figure 3, 4, 5 and 6, clustering results vary significantly across algorithms. Our proposed algorithm, **COMET**, excels in identifying true clusters, maintaining high ARI values even as noise increases, demonstrat-

Index	Noise	KM	MKM	CC	RCC	RConv	RBKM	COMET
ARI	0	0.52±0.36†	0.55±0.39†	0.79±0.00†	0.01±0.00†	0.85±0.00†	0.002±0.001†	**0.88±0.00**
	5	0.58±0.36†	0.65±0.33†	0.82±0.01†	0.01±0.01†	0.84±0.00†	0.08±0.03†	**0.87±0.00**
	10	0.52±0.35†	0.47±0.39†	0.81±0.01†	0.01±0.00†	0.85±0.03†	0.15±0.06†	**0.87±0.01**
	15	0.66±0.29†	0.54±0.38†	0.80±0.02†	0.01±0.01†	0.84±0.02†	0.24±0.05†	**0.86±0.01**
	20	0.40±0.39†	0.53±0.38†	0.82±0.03†	0.01±0.01†	0.86±0.03†	0.27±0.11†	**0.88±0.02**
AMI	0	0.47±0.32†	0.47±0.34†	0.66±0.00†	0.08±0.00†	0.73±0.00†	0.002±0.001†	**0.80±0.00**
	5	0.52±0.32†	0.57±0.29†	0.69±0.01†	0.06±0.001†	0.73±0.01†	0.12±0.02†	**0.76±0.02**
	10	0.48±0.31†	0.41±0.34†	0.67±0.01†	0.07±0.003†	0.75±0.03†	0.19±0.05†	**0.76±0.01**
	15	0.60±0.25†	0.47±0.33†	0.67±0.02†	0.08±0.005†	0.72±0.02$^{\sim}$	0.27±0.03†	**0.75±0.03**
	20	0.37±0.35†	0.46±0.33†	0.68±0.02†	0.07±0.01†	0.76±0.03$^{\sim}$	0.31±0.06†	**0.77±0.02**
k^*	0	2.09±0.85	1.92±0.83	15±0.00	462±0.00	3±0.00	2±0.00	**3±0.00**
	5	2.13±0.45	2.2±0.81	15±1.86	508±7.12	3±0.56	2±0.00	**2±0.00**
	10	2.25±0.63	2.00±0.82	15±1.86	477±9.06	2±1.04	2±0.00	**3±0.00**
	15	2.47±0.62	2.04±0.92	15±1.42	457±±8.57	3±0.97	2±0.00	**3±0.00**
	20	1.93±0.69	2.12±0.83	15±1.42	481±9.21	3±0.97	2±0.00	**2±0.00**

Table 3: Performance of Different Algorithms on **Wisconsin** on different Noise levels
† : significantly different from the best performing algorithm, $^{\sim}$: statistically same as the best performing algorithm .

ing robustness. In contrast, $k-$means and MoM $k-$means struggle to detect the underlying structure which is a limitation of the $k-$means framework. Convex Clustering performs similarly to COMET in identifying cluster structure and count, but struggles with noise, whereas COMET effectively isolates noise, showcasing its superior performance in noisy environments.

A.8 Real-life Data

Here we have provided the performance of the selected algorithms on the real-life datasets mentioned earlier in Table 1, which are not included in Table 2 in the main paper. Here, k^* refers to an estimated number of clusters. The actual number of clusters is indicated as k. Here the standard deviation is that of the performance measure, not of the mean statistic.

Observe in Table 2, that even though in some datasets like Zoo, Coil20, Lymphoma (Microarray), etc., COMET is not the best performing algorithm, still it gives performance close to the respective best performing algorithms.

A.9 Case study on Wisconsin Dataset

We evaluate our algorithm's performance on the Wisconsin dataset, available, and compare it with other algorithms. The dataset contains 699 cases from a study on breast cancer patients, with 9 categorical features and two classes: Benign and Malignant. After refining the dataset to exclude missing values, we work with 683 instances.

The case study focuses on how our algorithm performs in the presence of noise, which is common in real-life clustering tasks. We introduce varying levels of uniform noise and test the algorithms at 0%, 5%, 10%, 15%, and 20% noise. For each noise level, we add $\lfloor np \rfloor$ new "noise" datapoints uniformly within the minimum hypercube containing the original data. The results, shown in Table 3 and Figure 1, reveal that COMET outperforms all algorithms, achieving the highest ARI and AMI across noise levels. Here the standard deviation is that of the performance measure, not of the mean statistic. Convex clustering and Robust Convex clustering also perform well, but slightly worse than COMET. $k-$means and MoM $k-$means are similar but much less effective, while RBKM and RCC perform poorly, with RBKM improving as noise increases. The *t*-SNE plots showing the clustering results of various algorithms on this datsets are provided in the appendix (see A.13). *t*-SNE Plots for performance of various algorithms on Wisconsin dataset is given in Figure 7.

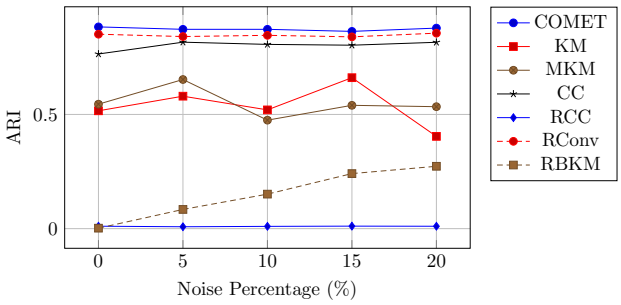


Figure 1: Line plot for performance of different algorithms on **Wisconsin**

A.10 *t*-SNE Plots for Wisconsin and Brain dataset

t-SNE plots for the clustering results on the Brain dataset using various clustering algorithms are given in Figure 8.

A.11 Wilcoxon-Rank Sum Test

To assess whether the ARI and AMI scores produced by our algorithm are *significantly higher* than those of selected baseline clustering methods, we employ the Wilcoxon Rank-Sum test. The ARI and AMI scores are computed for various algorithms on different 10% contaminated datasets, and the corresponding p-values of the Wilcoxon Rank-Sum test are estimated using *Monte Carlo simulation*. The results are presented in Table 4 and 5. For any value in the table with $p \leq 0.05$, we consider the difference to be *statistically sig-*

Dataset	CC	RCC	RConv	KM	MKM	RBKM
Iris	0.9409	2.32E-06	0.000385	0.97549	0.99922	0.998
Newthyroid	3.35E-06	1.96E-06	1.13E-06	1.61E-09	1.42E-09	1.48E-09
Ecoli	0.7399	1.37E-06	0.9967	2.52E-09	2.52E-09	1.31E-09
Wisconsin	3.35E-06	3.72E-06	1.78E-06	2.79E-05	1.41E-07	1.27E-09
Wine	0.0001	4.49E-07	1.64E-05	0.18801	0.001763	1.63E-09
Zoo	3.88E-05	2.18E-06	0.9984	5.96E-09	8.90E-07	1.49E-09
Dermatology	3.35E-06	3.62E-07	2.6E-05	1.05E-05	2.07E-07	1.61E-09
Brain	0.9993	2.41E-07	4.96E-05	1.98E09	1.72E-09	1.55E-10
Lung	3.35E-06	0.0159	5.28E-07	0.913832	0.9999	5.29E-17
Lymphoma	3.35E-06	4.44E-07	1.31E-06	0.03995	0.9997	1.56E-09
Coil-20	3.55E-06	1.17E-07	0.999	1.74E-09	1.73E-09	5.35E-15
WDBC	0.9949	4.46E-07	0.9895	9.09E-21	9.09E-21	0.00126
Lung-discrete	3.35E-06	3.50E-06	9.7E-07	2.04E-09	6.03E-09	1.27E-09
ORLRaw10P	3.35E-06	1.17E-07	3.55E-07	1.65E-09	1.63E-09	1.21E-09
Lymphoma (micro)	0.5	0.9999	0.000241	1.67E-08	9.58E-08	5.7E-16

Table 4: COMET vs Other Algorithms (ARI values only)

Dataset	CC	RCC	RConv	KM	MKM	RBKM
Iris	0.9999	2.32E-06	0.00029	0.58855	0.9992	0.9669
Newthyroid	3.35E-06	1.96E-06	7.47E-05	1.61E-09	1.42E-09	1.5E-09
Ecoli	0.9999	1.37E-06	0.9997	2.52E-09	2.52E-09	1.31E-09
Wisconsin	3.35E-06	3.72E-06	1.78E-06	0.004352	9.95E-07	1.27E-09
Wine	6.85E-06	4.49E-07	1.64E-05	0.17594	0.000765	1.63E-09
Zoo	5.15E-06	2.18E-06	0.4583	3.24E-09	1.44E-07	1.49E-09
Dermatology	3.35E-06	3.62E-07	2.6E-05	0.000254	1.65E-06	1.61E-09
Brain	3.35E-06	2.41E-07	4.96E-05	1.73E-09	1.72E-09	1.55E-10
Lung	3.35E-06	1.87E-06	5.28E-07	0.83273	0.9847	5.3E-17
Lymphoma	3.35E-06	4.44E-07	0.0921	0.0508	0.9996	1.6E-09
Coil-20	3.54E-06	1.17E-07	0.999	1.74E-09	1.73E-09	5.6E-15
WDBC	0.9949	4.46E-07	0.994	1.79E-11	9.09E-21	0.8603
Lung-discrete	3.35E-06	3.50E-06	9.7E-07	2.34E-09	2.51E-07	1.27E-09
ORLRaw10P	3.35E-06	1.17E-07	3.55E-07	1.65E-09	1.63E-09	1.21E-09
Lymphoma (micro)	0.5	0.9999	0.000241	1.06E-07	1.28E-06	5.7E-16

Table 5: COMET vs Other Algorithms (AMI values only)

nificant, indicating that our algorithm produces higher ARI and AMI scores under the tested conditions.

Based on the values in the Table 4 and 5, we observe that the performance measures for our algorithm are significantly higher than those for other algorithms. Our closest competitor is RConv, which performs better than us on 4 datasets out of the 15 we have tested on.

A.12 Ablation Study

Gamma We will do a sensitivity analysis of the hyperparameter γ on the performance of our algorithm For the convex clustering cost function, once a graph is created using $k-$NN for some k, the u_i's converge to the mean of the connected components if the value of γ is high. During our test runs, we also took a large value of γ, say 50000. We observe that, for our algorithm, the value of γ doesn't influence the final result, provided it is large (say, $\geq$ 1000). Refer to the Figure 2 for the plot.

Ablation study on Wisconsin Dataset We will study the fluctuations in our performance measures with varying hyperparameters. We will turn to the **Wisconsin Breast Cancer** dataset for our ablation studies. We have shown in section A.12 that the hyperparameter γ does not have much influence on the final clustering of the dataset. Two hyperparameters, k, which is a hyperparameter for the $k-$NN graph structure, and μ, require tuning to achieve optimal performance of our algorithm. For our experiments, we are varying k in $\{24, 27, 30, 33, 36\}$, μ in $[4, 17]$ and p(noise level) in $\{0\%, 5\%, 10\%, 15\%, 20\%\}$. For each pair (p,k), we are

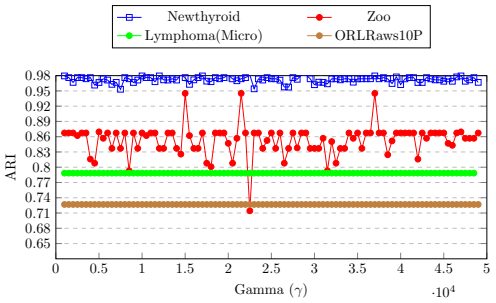


Figure 2: ARI for different values of γ after adding 10% noise

varing μ and reporting the mean of ARI and AMI. For each noise level, we observe that both the AMI and ARI values increase stochastically with μ and converge to a value. The fluctuations in ARI increase gradually with noise level for every value of k. Within each noise level however, ARI is most stable for the mid-range of k, which is 27 and 30, indicating that values of k from 27 to 30 have higher stabilities. Same observation can be made for AMI as well. All the fluctuations can be attributed to the randomness in adding noise and the optimization procedure of our objective function. Refer to the Figures 9 and 10 for the visuals of the tuning process.

Ablation study on Newthyroid Dataset We will turn to the Newthyroid dataset for our ablation studies. We have shown in section A.12 that the hyperparameter γ does not have much influence on the final clustering of the dataset. Two hyperparameters, k, which is a hyperparameter for the $k-$NN graph structure, and μ, require tuning to achieve optimal performance of our algorithm. For our experiments, we are varying k in $\{31, 38, 45, 52, 59\}$, μ in $[12, 110]$ and p(noise level) in $\{0\%, 5\%, 10\%, 15\%, 20\%\}$. For each pair ($p$,$k$), we are varying μ and reporting the mean of ARI and AMI. The ARI is almost always between 0.93 and 0.98 and very rarely drops to 0.92 at higher noise levels. The AMI values are also between 0.82 and 0.92. The Here we did not notice much fluctuations for both ARI and AMI values even for 20% noise level. All the fluctuations can be attributed to the randomness of adding noise and the optimization procedure of our objective function. The fluctuations increase with increasing p, however the increase is only slight. The fluctuations decrease while choosing higher values of k (such as 52 in our case). The effect of changing μ is also not apparent from the plots indicating that there are well-separated clusters in this dataset. However, for much smaller values of μ (say below 10), the ARI and AMI values are expected to drop. For a detailed graphical illustration of the tuning process, please refer to the Figures 11 and 12.

A.13 Plots

In this section, we provide the plots for results on various studies conducted.

Plots for Synthetic Datasets This section contains plots based on the synthetic data described in the section A.7. Figures 3, 4 and 5 represent a comparison of COMET with other SOTA algorithms for different noise levels on different datasets. Figure 6 gives a visual representation of the outputs of different algorithms at 10% noise. It is clearly visible that COMET captures the clustering pattern well and also properly classifies noise.

***t*-SNE Plots** This section contains the *t*-SNE plots for all the algorithms run on Wisconsin (Figure 7) and Brain (Figure 8). The *t*-SNE plots help in understanding the clustering pattern picked by various algorithms better for high dimensional datasets.

Ablation Plots Plots for ablation study as described in A.12 and A.12 are present in this section. The figures 9 and 10 correspond to the variation in ARI and AMI respectively for Wisconsin dataset. The figures 11 and 12 correspond to the variation in ARI and AMI respectively for NewThyroid dataset.

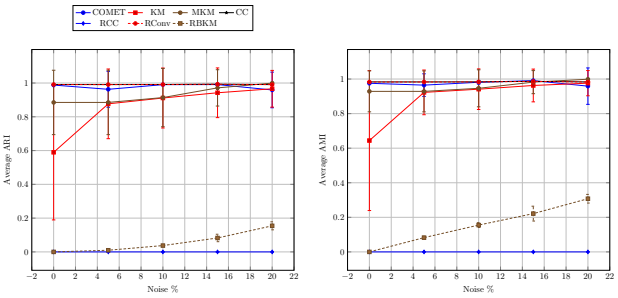


Figure 3: Performance of Algorithms on Blobs Dataset (ARI and AMI Values)

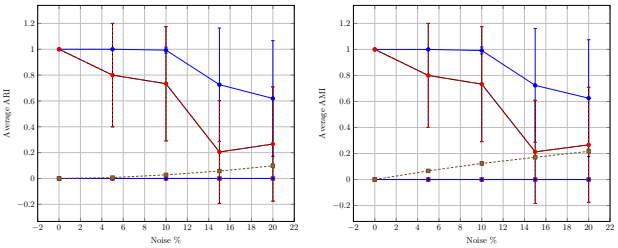


Figure 4: Performance of Algorithms on Circles Datasets (ARI and AMI Values)

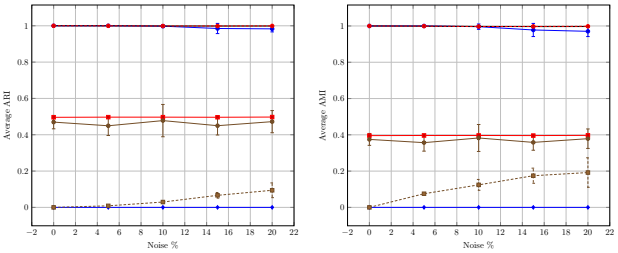


Figure 5: Performance of Algorithms on Moons Dataset (ARI and AMI Values)

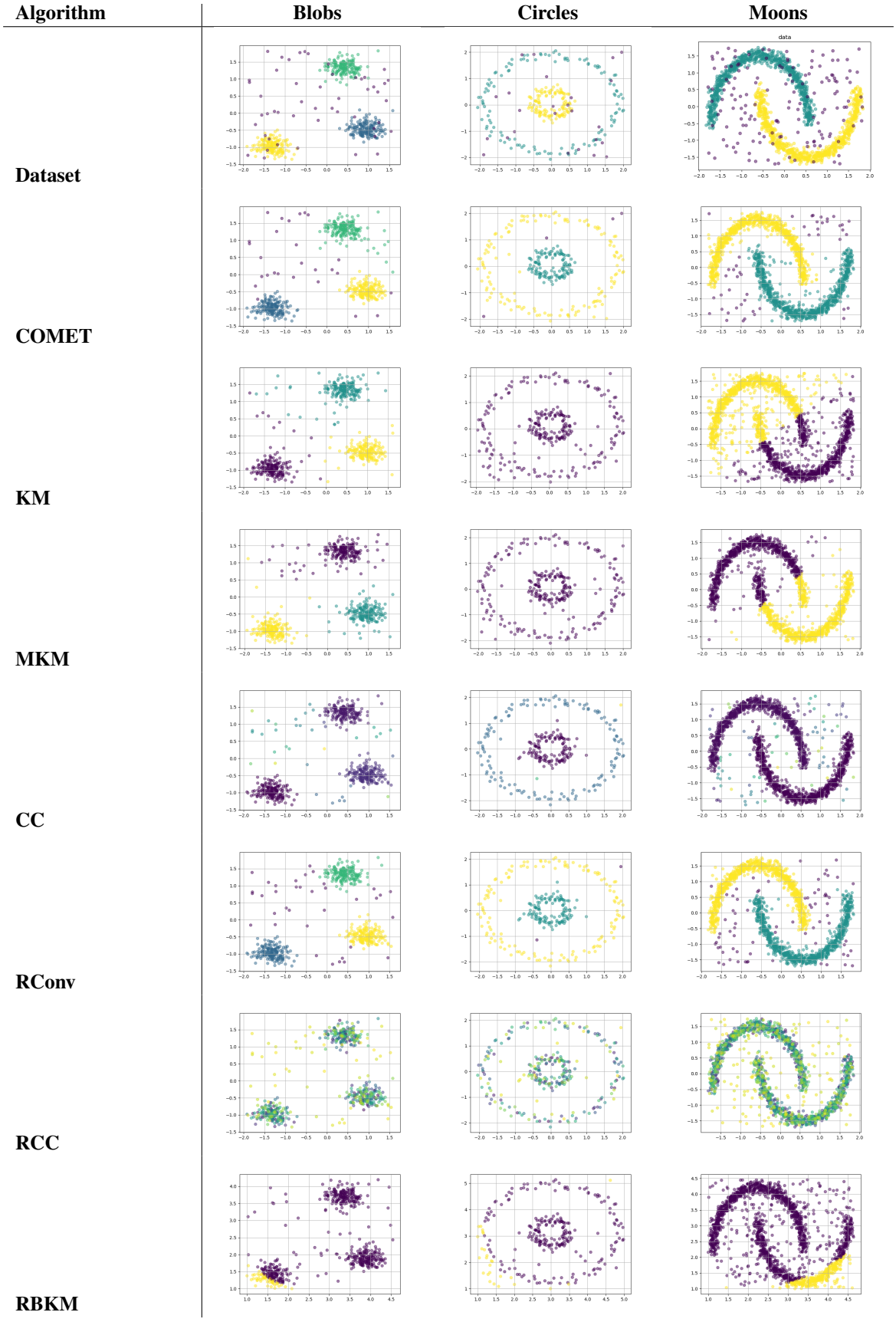
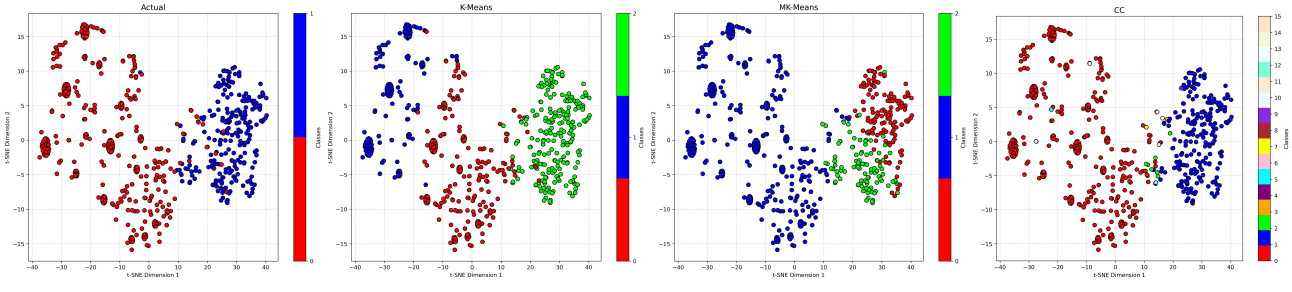
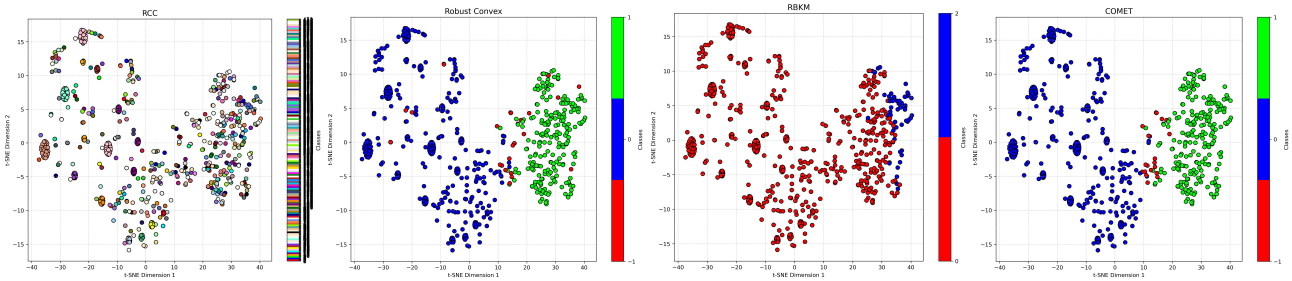


Figure 6: Clustering results for 7 algorithms across 3 datasets. Each row corresponds to an algorithm, and each column corresponds to a dataset.

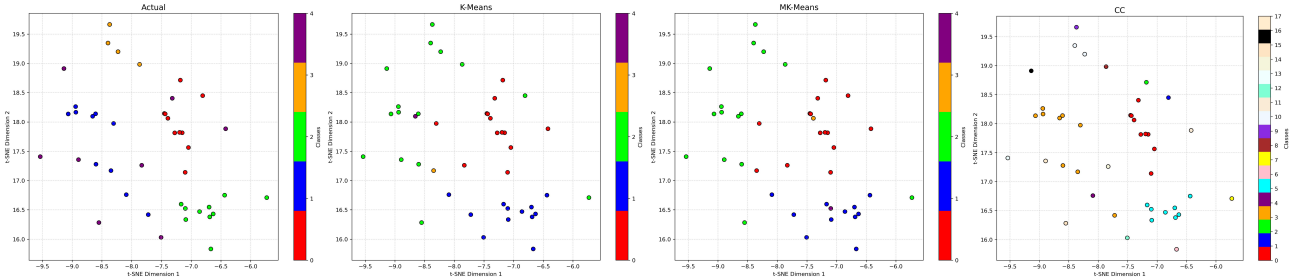


(a) Dataset (b) KM (c) MKM (d) CC

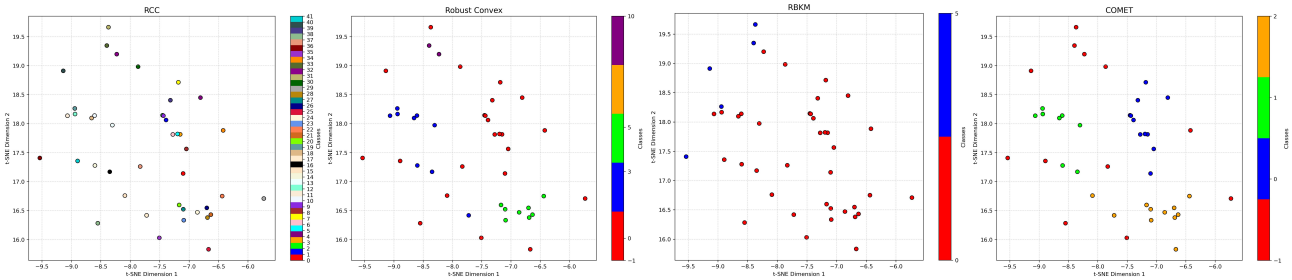


(e) RCC (f) RConv (g) RBKM (h) COMET

Figure 7: *t*-SNE plot of the Wisconsin dataset after clustering under various algorithms at 10% noise.

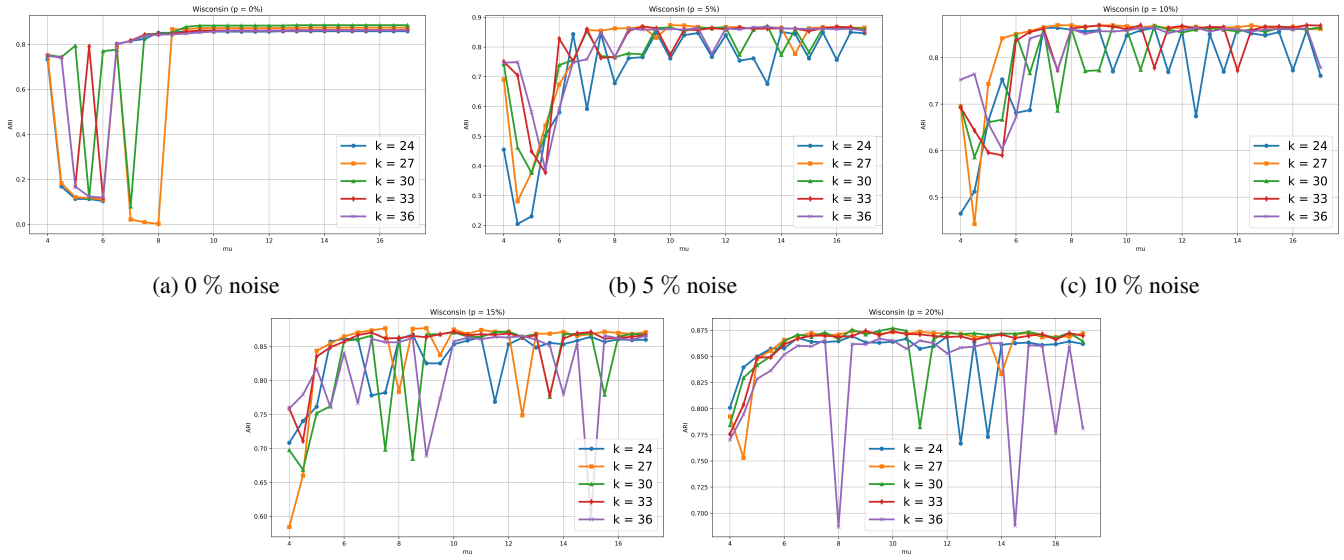


(a) Dataset (b) KM (c) MKM (d) CC



(e) RCC (f) RConv (g) RBKM (h) COMET

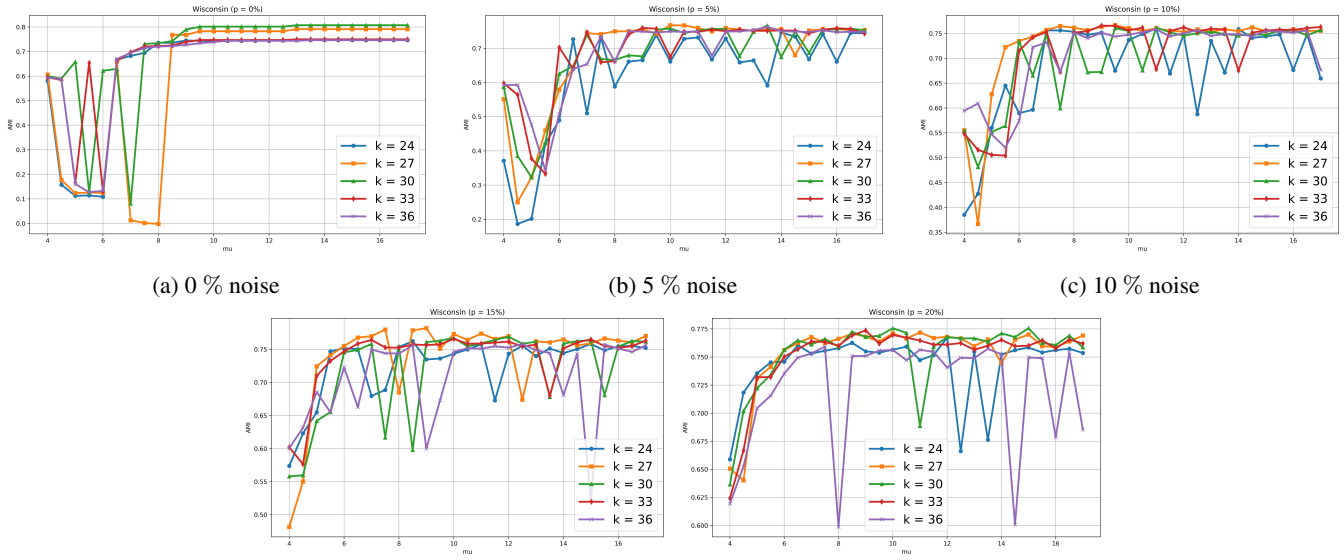
Figure 8: *t*-SNE plot of the Brain dataset after clustering under various algorithms.



(a) 0 % noise (b) 5 % noise (c) 10 % noise

(d) 15 % noise (e) 20 % noise

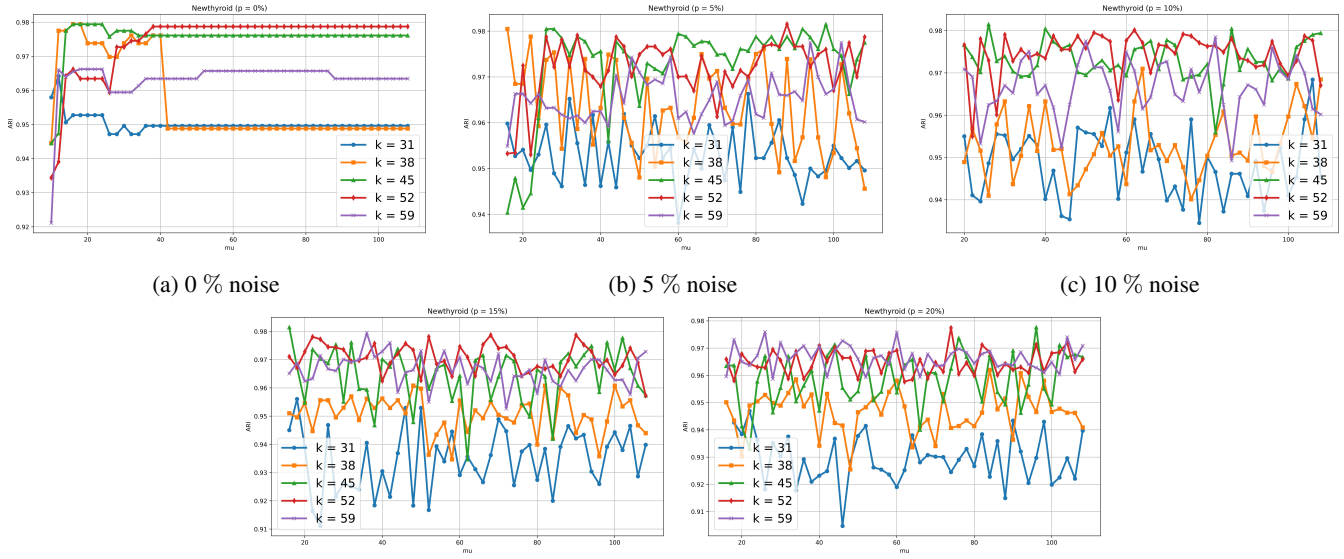
Figure 9: Ablation Studies of the ARI Values obtained from the Wisconsin Breast Cancer Dataset. Each subfigure corresponds to a different level of noise introduced into the dataset.



(a) 0 % noise (b) 5 % noise (c) 10 % noise

(d) 15 % noise (e) 20 % noise

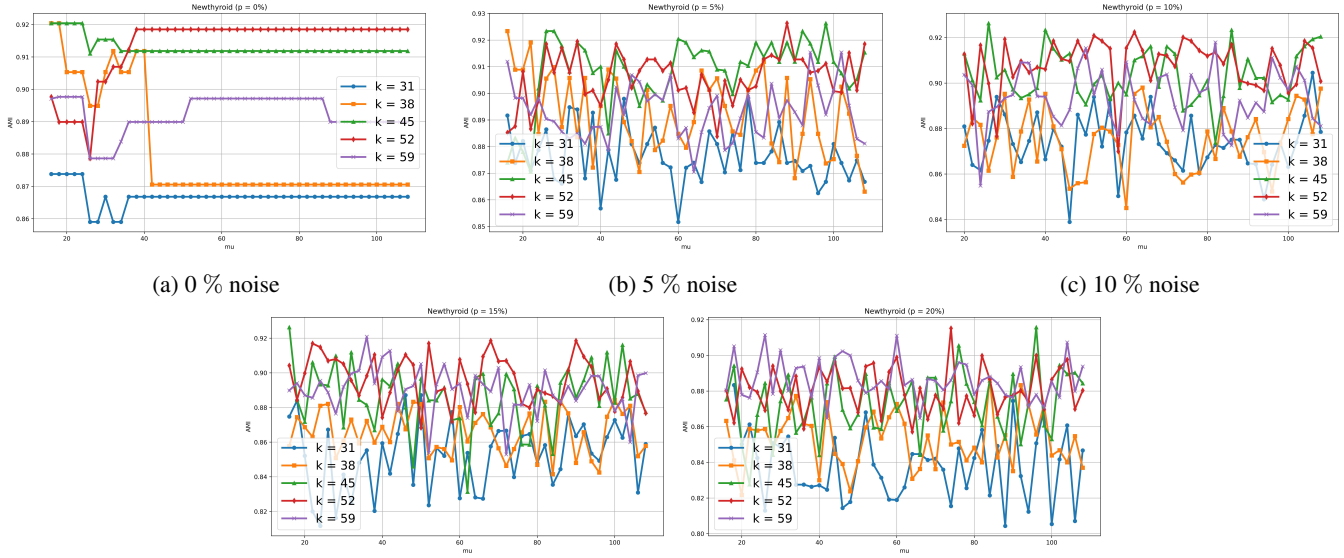
Figure 10: Ablation Studies of the AMI Values obtained from the Wisconsin Breast Cancer Dataset. Each subfigure corresponds to a different level of noise introduced into the dataset.



(a) 0 % noise (b) 5 % noise (c) 10 % noise

(d) 15 % noise (e) 20 % noise

Figure 11: Ablation Studies of the ARI Values obtained from the NewThyroid Dataset. Each subfigure corresponds to a different level of noise introduced into the dataset. (Note: The graphs may seem to have high variability, but the values in the y-axis shown are within 0.9 to 1)



(a) 0 % noise (b) 5 % noise (c) 10 % noise

(d) 15 % noise (e) 20 % noise

Figure 12: Ablation Studies of the AMI Values obtained from the NewThyroid Dataset. Each subfigure corresponds to a different level of noise introduced into the dataset. (Note: the range in the y-axis shown in the graphs is 0.8 to 1)