1. Appendix of the manuscript "On internal fuzzy cluster validity indices for soft subspace clustering of high-dimensional datasets".

1.1. Fuzzy cluster validity indices

Table 1: Fuzzy CVIs that only use **U**

Index	Best value	Definition
PC	$max_{1\leq i\leq c}$	$\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2$
MPC	$max_{1\leq i\leq c}$	$1-\frac{c}{c-1}(1-PC)$
IPC(U_c^a)	$max_{1\leq i\leq c}$	$100\left[\frac{PC(U_{c-1})-PC(U_c)}{PC(U_{c-1})}-\frac{PC(U_c)-PC(U_{c+1})}{PC(U_c)}\right]$
PE	$min_{1\leq i\leq c}$	$-\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log_a(u_{ik})$
PEB	$min_{1\leq i\leq c}$	$\frac{PE}{\log_a c}$
NPE	$min_{1\leq i\leq c}$	$\frac{nPE}{n-c}$
KYI	$min_{1\leq i\leq c}$	$\frac{2}{c(c-1)}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\sum_{k=1}^{n}[c\times\min(u_{ik},u_{jk})\times h(x_k)^b]$
P	$max_{1\leq i\leq c}$	$\frac{1}{n}\sum_{k=1}^{n}\max_i(u_{ik})-\frac{1}{K}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\left[\frac{1}{n}\sum_{k=1}^{n}\min(u_{ik},u_{jk})\right]$
MPO	$max_{1\leq i\leq c}$	$\left[\left(\frac{c+1}{c-1}\right)^{1/2}\times\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^2}{\min_{1\leq i\leq c}\left\{\sum_{k=1}^{n}u_{ik}^2\right\}}\right]-\left[\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}O_{ijk}(\mathbf{U})^c\right]$

[a] U_c is a fuzzy partition with c clusters

[b] $h(x_k)=-\sum_{i=1}^{c}u_{ik}\log_a u_{ik}$

[c] $O_{ijk}(\mathbf{U})=\begin{cases}1-|u_{ik}-u_{jk}| & \text{if } |u_{ik}-u_{jk}|\geq T_o\\ 0 & \text{otherwise}\end{cases}$

Table 2: Fuzzy CVIs that use **U, V, and X**

Index	Best value	Definition
FS	$min_{1\leq i\leq c}$	$\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}[d(x_k,v_i)^2 - d(v_i,\bar{v})^2]$
XB	$min_{1\leq i\leq c}$	$\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d(x_k,v_i)^2}{n \min_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2}$
K	$min_{1\leq i\leq c}$	$\frac{\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{2} d(x_k,v_i)^2 + \frac{1}{c}\sum_{i=1}^{c} d(v_i,\bar{v})^2}{\min_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2}$
T	$min_{1\leq i\leq c}$	$\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{2} d(x_k,v_i)^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{\substack{j=1 \\ j\neq i}}^{c} d(v_i,v_j)^2}{\min_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2 + 1/c}$
SC	$max_{1\leq i\leq c}$	$\left[\frac{\sum_{i=1}^{c} d(v_i,\bar{v})^2/c}{\sum_{i=1}^{c}\left(\sum_{k=1}^{n} u_{ik}^{m} d(x_k,v_i)^2 / \sum_{k=1}^{n} u_{ik}\right)}\right] - \left[\frac{\sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\left(\frac{\sum_{k=1}^{n}(\min(u_{ik},u_{lk}))^2}{\sum_{k=1}^{n}\min(u_{ik},u_{lk})}\right)}{\sum_{k=1}^{n}(\max_{1\leq i\leq c} u_{ik})^2 / \sum_{k=1}^{n}\max_{1\leq i\leq c} u_{ik}}\right]$
PBMF	$max_{1\leq i\leq c}$	$\left(\frac{1}{c} \times \frac{\sum_{k=1}^{n} d(x_k,\bar{v})}{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d(x_k,v_i)} \times \max_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)\right)^2$
PCAES	$max_{1\leq i\leq c}$	$\sum_{i=1}^{c}\left[\frac{\sum_{k=1}^{n} u_{ik}^{2}}{min_{1\leq i\leq c}\left\{\sum_{k=1}^{n} u_{ik}^{2}\right\}} - \exp\left(\frac{-\min_{i\neq j} d(v_i,v_j)^2}{\sum_{i=1}^{c} d(v_i,\bar{v})^2/c}\right)\right]$
SF	$max_{1\leq i\leq c}$	$\frac{\sum_{k=1}^{n}(\max_{1\leq i\leq c} u_{ik} - \max_{1\leq j\leq c, j\neq i} u_{jk})^{\alpha} S_{x_k}(\mathbf{U};\,\mathbf{X})^{a}}{\sum_{k=1}^{n}(\max_{1\leq i\leq c} u_{ik} - \max_{1\leq j\leq c, j\neq i} u_{jk})^{\alpha}}$
WLI	$min_{1\leq i\leq c}$	$\frac{\sum_{i=1}^{c}\left(\frac{\sum_{k=1}^{n} u_{ik}^{2} d(x_k,v_i)^2}{\sum_{k=1}^{n} u_{ik}}\right)}{\min_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2 + median_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2}$
HF	$min_{1\leq i\leq c}$	$\frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d(x_k,v_i)^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{\substack{j=1 \\ j\neq i}}^{c} d(v_i,v_j)^2}{\frac{n}{2c}(\min_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2 + median_{1\leq i,j\leq c, i\neq j} d(v_i,v_j)^2)}$

[a] $S_{x_k}(\mathbf{U};\,\mathbf{X}) = \frac{b_{ik}-a_{ik}}{\max(a_{ik},b_{ik})}$

1.2. Experimental setup

1.2.1. Synthetic datasets

The properties of the eight Gaussian mixture datasets are summarized in Table 3.

Table 3: Gaussian mixture datasets

Dataset	p	n	c_{opt}
Gaussian.k8	657	81	8
Gaussian.k2	4,232	181	2
Gaussian.k7	4,514	128	7
Gaussian.k9	5,041	108	9
Gaussian.k6	5,176	143	6
Gaussian.k5	6,203	130	5
Gaussian.k4	6,615	168	4
Gaussian.k3	7,959	75	3

1.2.2. Text

We use the subset versions of NSF [1], Hitech, Opinosis[2] and 20Newsgroups[3] textual collections made in [1]. The nine high-dimensional real text datasets used in this work are summarized in Table 4.

Table 4: Text collections

Dataset	p	n	c_{opt}	Domain
CSTR	1725	299	4	Scientific
NSF	2743	1600	16	Scientific
SyskillWebert	4339	333	4	Web pages
Hitech	6593	600	6	Newspaper
WAP	8049	1560	20	Web pages
Irish-Sentiment	8658	1,660	3	Sentiment Analysis
Opinosis	10784	51	3	Sentiment Analysis
20News groups	11015	2,000	4	E-mails
La1s	13195	3,204	6	News articles

1.2.3. DNA microarray

The gene expression datasets and their properties are summarized in Table 5.

[1] `http://archive.ics.uci.edu/ml/databases/nsfabs/`
[2] `http://kavita-ganesan.com/opinosis-opinion-dataset`
[3] `http://qwone.com/~jason/20Newsgroups/`

Table 5: DNA Microarray datasets

Dataset	p	n	c_{opt}	Domain
Sorlie	456	85	5	Breast Cancer
Christensen	1,413	217	3	N/A
Alon	2,000	62	2	Colon Cancer
Khan	2,308	63	4	Small Round Blue Cell Tumors
Gravier	2,905	168	2	Breast Cancer
Su	5,563	102	4	N/A
Pomeroy	7,128	60	2	CNS Tumor
Golub	7,129	72	2/3	Leukemia
West	7,129	49	2	Breast Cancer
Shipp	7,129	77	2	Lymphoma
Subramanian	10,100	50	2	N/A
Gordon	12,533	181	2	Lung Cancer
Singh	12,600	102	2	Prostate Cancer
Chiaretti	12,625	128	2/6	Leukemia
Tian	12,625	173	2	Myeloma
Yeoh	12,625	248	6	Leukemia
Chin	22,215	118	2	Breast Cancer
Borovecki	22,283	31	2	Huntington's Disease
Burczynski	22,283	127	3	Crohn's Disease
Chowdary	22,283	104	2	Breast Cancer

1.3. Similarity and distance measures

The cosine dissimilarity was defined as $1 - sim(x_k, v_i)$ where $sim(x_k, v_i)$ is the similarity between an object x_k to the cluster center v_i. The weighted version of cosine was defined as

$$wsim(x_k, v_i) = w\cos(\theta) = \frac{x_k \cdot w_i v_i}{\sqrt{\sum_{l=1}^{p} x_{kl}^2}\sqrt{\sum_{l=1}^{p} w_{il} v_{il}^2}}. \quad (1)$$

1.3.1. Methodology Process

The parameter values of m, stop criteria (threshold ε, maximum number of iterations t_{max}), and parameters γ of EWFCM and ESSC and η of ESSC are shown in Algorithm 1 that summarizes the clustering process in this work.

Algorithm 1 Clustering process for a given dataset $\mathbf{X} \subset \mathbb{R}^{n \times p}$

Input: d = Cosine, $f = 0.25, 0.50, 0.75$, Manhattan, Euclidean
wd = Weighted cosine, $f = 0.25, 0.50, 0.75$, Manhattan, Euclidean
$c = c_{min} = 2$, $\varepsilon = 10^{-5}$, and $t_{max} = 1000$

```
foreach m ∈ {1.5, 1.8, 2.0, 2.2, 2.5} do
    while c ≤ c_max do
        Randomly initialize U^(0);
        foreach d do
            FCM(X, c, m, U^(0), ε, t_max)
        end
        foreach wd do
            SCAD1(X, c, m, U^(0), ε, t_max)
            foreach γ ∈ {0.001, 0.01, 0.1} do
                EWFCM(X, c, m, U^(0), ε, t_max, γ)
            end
            foreach γ ∈ {10, 100, 1000} do
                foreach η ∈ {0.3, 0.6, 0.9} do
                    ESSC(X, c, m, U^(0), ε, t_max, γ, η)
                end
            end
        end
        c = c + 1
    end
end
```

Output: $\mathbf{U}^{(t)}$, $\mathbf{V}^{(t)}$, $\mathbf{W}^{(t)}$ of each algorithm.

1.4. Results and Discussion

The next tables summarize which distance measure and value of the fuzzy exponent, for each clustering algorithm, that the CVIs had a high selection rate of c_{min}, c_{opt}, or c_{max} as the optimal c. In these tables, the Manhattan and Euclidean distance functions, in the *d/wd* column, are represented by their respective exponents $f = 1$ and $f = 2$.

We show in Table 6 that the FCVIs more incorrectly recognized the optimal number of clusters when it was used the cosine dissimilarity and weighting fuzzy exponent $m = 2.5$. On the other hand, the CVIs had their best results, by selecting $c = c_{opt}$, for FDM of order $f = 0.50$ and $m = 1.8$.

Table 6: Distance measure and fuzzy exponent m with high validation rates for the Gaussian datasets

Algorithm	c_{min}		c_{opt}		c_{max}	
	d/wd	m	d/wd	m	d/wd	m
FCM	2	2.5	1	1.5	cos	2.5
SCAD1	w cos	2.5	1	1.5	w cos	2.5
EWFCM ($\gamma = 0.1$)	0.25	1.5	0.25-2[a]	2.5	w cos	2.5
ESSC ($\gamma = 10, \eta = 0.3$)	w cos	2.5	0.50	1.8	w cos	2.5
ESSC ($\gamma = 100, \eta = 0.3$)	w cos	2.5	0.25	1.8	w cos	2.5
ESSC ($\gamma = 1000, \eta = 0.3$)	w cos	2.5	0.50	1.8	w cos	2.5

[a]The distance measures with exponent f from 0.25 until 2 ($f = 0.25, 0.50, 0.75$, Manhattan, Euclidean) were selected.

In Table 7, we show that the FCVIs more incorrectly recognized the optimal number of clusters, by selecting $c = c_{min}$, when it was used the Manhattan distance ($f = 1$) and weighting fuzzy exponent $m = 1.5$, or by selecting $c = c_{max}$, when the cosine dissimilarity and $m = 2.5$ were used. Once again, the CVIs had their best results for FDM of order $f = 0.50$, and $m = \{1.8, 2.0\}$ were used to cluster text collections. Besides the cosine dissimilarity does not have the best results, as it was expected, when used to measure the dissimilarity between two documents, it was the measure that the CVIs more incorrectly recognized the optimal number of clusters c as c_{max}.

Table 7: Distance measure and fuzzy exponent m with high validation rates for the text collections

Algorithm	c_{min}		c_{opt}		c_{max}	
	d/wd	m	d/wd	m	d/wd	m
FCM	1	1.5, 2.2	0.5	2.0	cos, 0.25, 2	2.5
SCAD1	1	1.5, 2.2	0.5	2.0	w cos	2.5
EWFCM ($\gamma = 0.01$)	1	1.5	0.5	2.2	w cos	2.5
ESSC ($\gamma = 10, \eta = 0.3$)	1	1.5	0.5	2.5	w cos	2.5
ESSC ($\gamma = 100, \eta = 0.3$)	1, 2	1.5	0.25, 0.75	1.8	w cos	2.5
ESSC ($\gamma = 1000, \eta = 0.3$)	1, 2	1.5, 2.2	0.25	1.5, 1.8	w cos	2.5

We show in Table 8 that the FCVIs more incorrectly recognized the optimal number of clusters when it was used the cosine dissimilarity and weighting fuzzy exponent $m = 2.0$ ($c = c_{min}$) and $m = 2.2$ ($c = c_{max}$). Different from the Gaussian and text collections, the CVIs had their best results when the conventional Euclidean distance and $m = \{1.5, 2.0\}$ were used to cluster high-dimensional microarray

data. In the next section, we present a summary of this work and its final considerations.

Table 8: Distance measure and fuzzy exponent m with high validation rates for the microarray datasets

Algorithm	c_{min}		c_{opt}		c_{max}	
	d/wd	m	d/wd	m	d/wd	m
FCM	cos, 0.5, 1	1.8	0.5, 1	1.5	cos	2.2, 2.5
SCAD1	w cos, 0.75	1.5, 2.0	w cos, 0.25, 0.5	1.5	w cos	2.2, 2.5
EWFCM ($\gamma = 0.1$)	w cos, 1	1.8, 2.2	0.25, 2	1.5	w cos	2.5
ESSC ($\gamma = 10, \eta = 0.3$)	0.5, 2	2.0, 2.2	0.5, 2	2.0	w cos	2.2
ESSC ($\gamma = 100, \eta = 0.3$)	w cos, 2	2.0	2	2.0	w cos	2.2
ESSC ($\gamma = 1000, \eta = 0.3$)	2	2.0	2	2.0	w cos	2.2

Bibliography

[1] T. M. Nogueira, S. O. Rezende, H. A. Camargo, Flexible document organization: Comparing fuzzy and possibilistic approaches, in: 2015 IEEE International Conference on Fuzzy Systems, 2015, pp. 1–8.