Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering

Kamil Kmita, Katarzyna Kaczmarek-Majer and Olgierd Hryniewicz

***Abstract*—Controlling the impact of partial supervision on the outcomes of modeling is of uttermost importance in semi-supervised fuzzy clustering. Semi-Supervised Fuzzy C-Means (SSFCMeans), a specific model we consider, uses a single hyperparameter called a scaling factor α to weigh the impact of partially labeled data. This concept became widespread and was reused directly in many works building on SSFCMeans, or even applied to other fuzzy clustering algorithms such as Possibilistic C-Means. However, none of the works challenged the original interpretation of α which suggests that the impact of partial supervision is directly proportional to the scaling factor. We fill the above research gap and thoroughly analyze this relationship. We provide novel explanations of the scaling factor α in terms of the key element of fuzzy clustering - the membership values. We prove that the impact of partial supervision is a non-linear function of α. Our approach is rooted in the explainability framework, which distinguishes interpretation from an explanation and treats the latter as superior. Explaining the scaling factor leads to an explainable impact of partial supervision and enables greater control of it. Finally, built on the novel explanations, we propose a unified, analytically justified framework for selecting the value of the hyperparameter α that is based on the cross-validation approach. We illustrate that the proposed framework enables an extensive analysis of the impact of partial supervision in SSFCMeans with a simulation experiment.**

***Index Terms*—Semi-supervised learning, Fuzzy clustering, Explainable artificial intelligence, Partial supervision, Fuzzy C-Means, Possibilistic C-Means.**

I. Introduction

SEMI-SUPERVISED learning (SSL) is often said to be "halfway between supervised and unsupervised learning" [1, p. 2]. Considering such a description, it does matter from which end we look at the problem: the unsupervised or the supervised one. Let us thus consider fuzzy clustering, one of the major unsupervised learning tasks. The aim is to group N unlabeled observations into c subgroups (clusters) so that observations in the same cluster are similar to each other while being dissimilar to observations from the other clusters. The unsupervised problem becomes semi-supervised when new information about a part of M observations out of all N observations ($M < N$) is obtained. This additional information is hence called *partial supervision*. In our scenario, it is given in the form of a label $y \in \{y_1, \ldots, y_c\}$ denoting the class to which an observation belongs.

K. Kmita, K. Kaczmarek-Majer and O. Hryniewicz are with the Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland (e-mail: kmita@ibspan.waw.pl; k.kaczmarek@ibspan.waw.pl; hryniewi@ibspan.waw.pl) *(Corresponding author: Kamil Kmita)*.

Kamil Kmita and Katarzyna Kaczmarek-Majer received funding from Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project: "Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)" http://bipolar.ibspan.waw.pl.

The class of semi-supervised fuzzy clustering models adapted to handle this type of partial supervision that we regard (i) is based on the partitioning approach where the number of clusters $c \geq 2$ is fixed, and (ii) defines similarity as a distance between observations and clusters' prototypes measured by a metric d. These models are thus referred to as to *distance-based semi-supervised fuzzy clustering* models (SSFC in short) in the literature [2].

The fundamental design choice of any SSFC model is how to manage the impact of partial supervision on the results of clustering: estimated degrees of memberships and clusters' prototypes. One technique of controlling the impact of partial supervision that we call the *additive combination* was introduced in [3]. It relies on a special construction of the associated objective function that combines two components in an additive manner: the unsupervised one and the supervised one. Pedrycz and Waletzky [3] proposed the additive combination as an element of the Semi-Supervised Fuzzy C-Means (SSFCMeans) model, the adaptation of the famous unsupervised Fuzzy C-Means (FCM) described in [4].

Works [5]–[13] extended SSFCMeans in different ways and modified the mechanism of handling partial supervision to various extents, but did not change the core idea of additive combination nor its interpretation. Works [14]–[17] wrapped SSFCMeans to analyze data streams, primarily in the problem of monitoring bipolar disorder. Works [18]–[22] explored safe semi-supervised clustering aiming at handling mislabeled instances (label errors). Kmita et al. [23] developed a procedure to estimate the uncertainty of labels resulting from an indirect annotation process. Last but not least, the very idea of the additive combination was applied to unsupervised fuzzy clustering models alternative to Fuzzy C-Means. These include Possibilistic C-Means (PCM) proposed in [24], and a mixture of FCM and PCM called Possibilistic Fuzzy C-Means (PFCM) [25]. The core unsupervised models, Fuzzy C-Means and Possibilistic C-Means differ in the implementation and interpretation of the soft assignment mechanism. PCM, just like FCM, was studied and modified by many researchers, including [26] who proposed Repulsive PCM. A semi-supervised version of Repulsive PCM was proposed in [27], and a semi-supervised adaptation of PFCM was described in [28].

All the abovementioned SSFC models share the same way of controlling the impact of partial supervision formulated originally in [3], although it may not be phrased directly (as different naming conventions are used). This impact is controlled with a single hyperparameter of the algorithm that we call a *scaling factor* after [3] and denote it with α. Pedrycz

and Waletzky described the role of this hyperparameter α "(...) is to maintain a balance between the supervised and unsupervised component within the optimization mechanism" [3, p. 789]. They did not quantify the impact nor discuss the relationship between the value of α and the key outcome of the SSFCMeans algorithm: the degrees of membership, and none of the positions in the literature that reused the additive combination technique in the sense of [3] explained this relationship either.

The main contribution of this work is to comprehensively explain the role of the scaling factor α in SSFC because the existing descriptions can be treated only as interpretations of it. The distinction between these two terms is receiving close attention in statistical learning [29]–[31], with an explanation perceived as superior to an interpretation. We also postulate to unambiguously quantify the impact of partial supervision in the form of a function of α denoted as IPS(α).

Explainable models are especially important in healthcare data modeling, and are often referred to as eXplainable AI (XAI); such models enable the comprehension of the reasoning underlying the predictions that they produce. The motivation for this work arose also from the previous work of the authors [23], where a procedure called Confidence Path Regularization was proposed. This procedure wrapped the SSFCMeans model to estimate label uncertainty in the semi-supervised problem of monitoring the health status of patients diagnosed with bipolar disorder. The scaling factor α is of key importance for this procedure, and considerations on the topic of Confidence Path Regularization led to the conclusion that existing descriptions of α were not sufficient for the improvements of the whole procedure.

In this article, we fill the identified research gap and explain the impact of partial supervision in two core SSFC models. Firstly, we study the aforementioned SSFCMeans model. The second model we investigate is called Semi-Supervised Possibilistic C-Means (SSPCMeans). We create it by applying the additive combination technique to introduce partial supervision to the classical PCM. SSFCMeans and SSPCMeans differ in the implementation of the soft assignment, hence the explanation of the scaling factor will differ as well. Our explanations apply to any model extending either SSFCMeans or SSPCMeans.

The structure of this article is as follows. In Section II, we discuss preliminaries of semi-supervised fuzzy clustering. We present the additive combination technique, SSFCMeans, and SSPCMeans models in detail. In Section III, we formalize a difference between an interpretation and an explanation and provide two novel explanations of the scaling factor α. Section IV is focused on the practical considerations stemming from the novel quantification of the impact of partial supervision. Finally, Section V concludes the article.

II. Semi-Supervised Fuzzy Clustering Preliminaries

We now introduce basic definitions related to the semi-supervised fuzzy clustering. Let j denote any observation (unsupervised or supervised), $j = 1, \ldots, N$, and k denote a given cluster, $k = 1, \ldots, c$. In addition to these indices, partial supervision requires to distinguish between (i) supervised observations indexed by $i = 1, \ldots, M$, and (ii) unsupervised observations indexed by $h = 1, \ldots, H$. A jth observation is represented by a $p-$dimensional feature vector $\mathbf{x}_j \in \mathbb{R}^p$, and a kth cluster is represented by a $p-$dimensional vector $\mathbf{v}_k \in \mathbb{R}^p$ called a prototype of the cluster. In the remainder of this article, d means the Euclidean distance.

The soft assignment of jth observation to kth cluster is usually expressed by a membership $u_{jk} \in [0, 1]$. This convention is used in Fuzzy C-Means and all models building on it. However, Possibilistic C-Means uses a typicality $t_{jk} \in [0, 1]$ convention to stress the fact that the interpretation of the soft assignment in PCM differs from the one used in FCM. For a cohesive presentation, we express a general concept of the soft assignment common for all SSFC models by memberships u_{jk} when the specific details do not affect the overall reasoning.

The partial information itself is expressed in the form of a prior memberships matrix $F = [f_{jk}]$ of the same dimension as memberships matrix U. Every cluster represented by a specific column of matrix F must be arbitrarily associated with a single class. To this end, we create a $c-$tuple $Y = \langle y_1, \ldots, y_c \rangle$ out of the set $\{y_1, \ldots, y_c\}$ and associate kth column in F with kth label y_k from Y. We define f_{jk} as binary entries such that $f_{jk} = 1$ only if jth observation is known to belong to kth cluster (associated with the y_k label); otherwise f_{jk} is equal to 0. An unsupervised observation h has all prior memberships $f_{hk} = 0 \; \forall k$. Frequently, an auxiliary variable b_j is used; $b_j = 1$ iff jth observation is supervised. In our scenario, $b_j = \sum_{k=1}^{c} f_{jk}$, hence one could question if this variable is indeed necessary. The choice whether to use b_j and how to include it in the model is a matter of subtle consequences that we discuss introducing relevant models in the remainder of this Section.

With partial supervision introduced in SSFC, a need occurs for supervised observations to distinguish between their membership degrees to "unsupervised" and "supervised" cluster. By "supervised cluster" we mean "the cluster associated with the class y_k that the observation is known to belong to". To retrieve this cluster, we define a function $s(i) \in \{1, \ldots, c\}$ that selects the index of the cluster associated with the ith supervised observation's class, i.e., $f_{ik} = 1$ iff $k = s(i)$.

Further on, we discuss 3 distinct types of memberships

- u_{hk} membership of an unsupervised observation h to any cluster k,
- $u_{i,k\neq s(i)}$ membership of a supervised observation i to any non-supervised cluster,
- $u_{i,s(i)}$ membership of a supervised observation i to the supervised cluster $s(i)$.

A. The additive combination technique

Since semi-supervised fuzzy clustering models modify the core unsupervised fuzzy clustering models, we introduce the additive combination technique by first considering a general form of unsupervised fuzzy clustering model parameterized by hyperparameters gathered in Θ. The optimization problem is formulated as

$$\arg\min_{U,V} \quad Q(U,V;X,\Theta) \tag{1}$$

where Q is the objective function, $U = [u_{jk}]_{N\times c}$ is a memberships matrix, $V = [\mathbf{v}_k]_{c\times p}$ is a prototypes matrix, and $X = [\mathbf{x}_j]_{N\times p}$ is a features matrix.

SSFC models adapt the minimization problem from (1) by combining the unsupervised objective function Q with its counterpart Q^S which incorporates the partial supervision (hence the name *additive combination*), arriving at

$$J(U,V;X,F,\Theta) = \underbrace{Q(U,V;X,\Theta)}_{\text{unsupervised comp.}} + \alpha \cdot \underbrace{Q^S(U,V;X,F,\Theta)}_{\text{supervised comp.}} \tag{2}$$

where $\alpha > 0$ is the scaling factor that controls the impact of partial supervision.

Specific hyperparameters gathered in Θ differ by models, yet one hyperparameter is common for all of them. It is the "fuzzifier" $m > 1$ that controls the fuzziness of the soft assignments. Bezdek et al. [4, p. 70] described it as "the larger m is, the fuzzier are the membership assignments". In this article, we use the specific value $m = 2$. The justification is provided in [3, p. 789]: any value of $m \neq 2$ would result in a situation where the variables optimized were linked together in the form of a polynomial and numerical procedures would be needed to solve its roots.

In general, finding optimal $(U^\star, V^\star)$ per (1) is intractable and approximation algorithms are often used. A typical optimization procedure for fuzzy clustering is described in [32]. It relies on fixing one variable and optimizing the other at a time. Such an iterative procedure is performed until a convergence criterion is met. The formulae for two variables $\hat{U}$ and $\hat{V}$ are obtained by studying first-order necessary conditions for a global minimizer $(U^\star, V^\star)$ of a respective objective function. SSFC models can follow the same optimization procedure as long as functions $J(U) = J(U;V,X,F,\Theta)$ and $J(V) = J(V;U,X,F,\Theta)$ remain convex. Indeed, this is the case for the functions J_{SSFCM} and J_{SSPCM} introduced in the remainder of this Section.

The two models we discuss, SSFCMeans and SSPCMeans, draw heavily from those introduced in [3] and [27], respectively. Our subtle yet important modifications are discussed in Subsection II-B and Subsection II-C. Whenever we present original equations from the referenced articles, we adapt them to follow the nomenclature introduced in this Section. We annotate all formulae from [3] with subscript (or superscript) $P97$.

B. Semi-Supervised Fuzzy C-Means

The objective function $J_{\text{SSFCM}}(U,V;X,F,\Theta)$ proposed in this article has a form

$$J_{\text{SSFCM}} = \sum_{k=1}^{c}\sum_{j=1}^{N} u_{jk}^2 d_{jk}^2 + \alpha \sum_{k=1}^{c}\sum_{j=1}^{N} b_j (u_{jk} - f_{jk})^2 d_{jk}^2, \tag{3}$$

where the first component corresponds to the objective function Q_{FCM} of the classical unsupervised Fuzzy C-Means model described in [32]. The minimization problem to solve is thus

$$\arg\min_{U,V} \quad J_{\text{SSFCM}}(U,V;X,F,\Theta) \tag{4a}$$

$$\text{s.t.} \quad \sum_{k=1}^{c} u_{jk} = 1 \quad \forall j = 1,\dots,N, \tag{4b}$$

$$0 < \sum_{j=1}^{N} u_{jk} < N \quad \forall k = 1,\dots,c, \tag{4c}$$

$$u_{jk} \in [0,1], \tag{4d}$$

where constraints (4b), (4c), (4d) are the same as in unsupervised FCM. Below we present the objective function J_{P97} from [3, Eq. (2)]

$$J_{\text{P97}} = \sum_{k=1}^{c}\sum_{j=1}^{N} u_{jk}^2 d_{jk}^2 + \alpha \sum_{k=1}^{c}\sum_{j=1}^{N} (u_{jk} - b_j \cdot f_{jk})^2 d_{jk}^2. \tag{5}$$

As opposed to J_{SSFCM} (3) proposed in this article, it was only f_{jk} that was multiplied by b_j, not the entire expression $(u_{jk} - f_{jk})^2$. Pedrycz and Waletzky [3] presented in detail a solution to (4) w.r.t U - but using J_{P97}, not J_{SSFCM}.

We applied the same analysis as presented in [3], but for the objective function J_{SSFCM} (3) proposed in this article, obtaining the formula for the optimal membership

$$\hat{u}_{jk} = \frac{1 + \alpha\big(b_j - b_j \sum_{g=1}^{c} f_{jg}\big)}{1 + \alpha b_j} e_{jk} + \frac{\alpha b_j}{1 + \alpha b_j} f_{jk}. \tag{6}$$

We do not present full derivation, referring the reader interested in details to [3]. An important part of (6) is

$$e(\mathbf{x}_j, V, k) = e_{jk} = \frac{1}{\sum_{g=1}^{c} d_{jk}^2 / d_{jg}^2} = \left(\sum_{g=1}^{c} \frac{d^2(\mathbf{x}_j, \mathbf{v}_k)}{d^2(\mathbf{x}_j, \mathbf{v}_g)} \right)^{-1} \tag{7}$$

that we call the *data evidence*. Note that the data evidence e_{jk} is a function of the feature's vector $\mathbf{x}_j$, the prototypes, and the index of the cluster considered. Consequently, the membership u_{jk} (6) is also a function $u_{jk} = u(\mathbf{x}_j, V, k, \alpha)$, but the notation u_{jk} is used for brevity.

Let us now apply the generic formula from (6) to distinct types of the membership $\hat{u}_{hk}$, $\hat{u}_{i,k\neq s(i)}$, $\hat{u}_{i,s(i)}$. First, consider an unsupervised observation h. In such case, $b_h = 0$ and $f_{hg} = 0\ \forall g$. Then, (6) simplifies to

$$\hat{u}_{hk} = \frac{1 + \alpha(0 - 0)}{1 + \alpha \cdot 0} \cdot e_{hk} + \frac{\alpha}{1 + \alpha \cdot 0} \cdot 0 = e_{hk}. \tag{8}$$

For the unsupervised observation, there is no direct impact of partial supervision on the value of the membership. It depends only on the data evidence, just as in FCM [32, p. 66].

Investigating ith supervised observation and its memberships, we first consider a degree of membership to any non-supervised cluster $k \neq s(i)$

$$\hat{u}_{i,k\neq s(i)} = \frac{1}{1+\alpha} \cdot e_{i,k\neq s(i)}. \tag{9}$$

The data evidence $e_{i,k\neq s(i)}$ is decreased by the factor of $\frac{1}{1+\alpha}$. It is a desired result of the partial supervision mechanism. Even if the data evidence were to support the belonging of the ith observation to the $k \neq s(i)$ cluster, the additional information we possess would decrease this membership.

The equations above clarify the mechanism of the SSFCMeans, but it is the membership of the supervised observation i to the supervised cluster $s(i)$ that is of major interest

$$\hat{u}_{i,s(i)} = \frac{1}{1+\alpha} \cdot e_{i,s(i)} + \frac{\alpha}{1+\alpha}. \tag{10}$$

We can observe that it includes a data-invariant component $\frac{\alpha}{(1+\alpha)}$ that depends only on the value of the scaling factor.

We now recall the formula for the optimal membership $\hat{u}_{jk}^{\text{P97}}$ presented in [3, p. 789] without equation number as

$$\hat{u}_{jk}^{\text{P97}} = \frac{1+\alpha\big(1-b_j\sum_{g=1}^{c} f_{jg}\big)}{1+\alpha} e_{jk} + \frac{\alpha}{1+\alpha}\big(b_j f_{jk}\big). \tag{11}$$

While (11) differs from (6), the distinct types of memberships $\hat{u}_{hk}^{\text{P97}}$, $\hat{u}_{i,k\neq s(i)}^{\text{P97}}$ $\hat{u}_{i,s(i)}^{\text{P97}}$ do not differ from their counterparts derived from J_{SSFCM} and presented in (8), (9), (10). We leave the simple calculus confirming this statement to the reader and state that the difference between J_{SSFCM} and J_{P97} does not result in different estimated memberships. However, this difference between the objective functions affects estimated prototypes $\hat{V}$. Let us note that [3] did not permit partial supervision to influence clusters' prototypes, associating $\hat{V}^{\text{P97}}$ with formulae from unsupervised FCM. Therefore, to present the effect of treating b_j differently in J_{SSFCM} and J_{P97}, we derive $\hat{V}$ from scratch. Let us define

$$\begin{aligned} J_{\text{SSFCM}}(\mathbf{v}_k) &= \sum_{k=1}^{c}\sum_{j=1}^{N}\Big(u_{jk}^2 + \alpha b_j (u_{jk}-f_{jk})^2\Big)\|\mathbf{x}_j - \mathbf{v}_k\|^2 \\ &= \sum_{k=1}^{c}\sum_{j=1}^{N}\phi_{jk}\|\mathbf{x}_j - \mathbf{v}_k\|^2, \end{aligned} \tag{12}$$

where $\phi_{jk} = u_{jk}^2 + \alpha b_j (u_{jk}-f_{jk})^2$ is called an *individual contribution*. We now find the stationary point of $J_{\text{SSFCM}}(\mathbf{v}_k)$, by setting $\partial J_{\text{SSFCM}}(\mathbf{v}_k)/\partial \mathbf{v}_k = -2\sum_{j=1}^{N}\phi_{jk}(\mathbf{x}_j - \mathbf{v}_k)$ to $\mathbf{0}$, and obtain

$$\hat{\mathbf{v}}_k = \frac{\sum_{j=1}^{N}(\phi_{jk}\cdot\mathbf{x}_j)}{\sum_{j=1}^{N}\phi_{jk}}. \tag{13}$$

Optimizing $J_{\text{P97}}(\mathbf{v}_k)$, one would arrive at the similar equation to (13), but instead of individual contributions ϕ_{jk}, there would be $\omega_{jk} = u_{jk}^2 + \alpha\cdot(u_{jk} - b_j\cdot f_{jk})^2$.

Let us compare the form of individual contribution ϕ_{jk} and ω_{jk} in 3 distinct types of the soft assignment:

$$\phi_{hk} = u_{hk}^2, \tag{14a}$$
$$\phi_{i,k\neq s(i)} = (1+\alpha)u_{i,k\neq s(i)}^2, \tag{14b}$$
$$\phi_{i,s(i)} = u_{i,s(i)}^2 + \alpha(u_{i,s(i)}-1)^2, \tag{14c}$$

and

$$\omega_{hk} = (1+\alpha)u_{hk}^2, \tag{15a}$$
$$\omega_{i,k\neq s(i)} = (1+\alpha)u_{k\neq s(i)}^2, \tag{15b}$$
$$\omega_{i,s(i)} = u_{i,s(i)}^2 + \alpha(u_{i,s(i)}-1)^2. \tag{15c}$$

In the case of $\hat{\mathbf{v}}_k^{\text{P97}}$, the individual contribution of the unsupervised observation ω_{hk} is the same as the contribution of the supervised $\omega_{i,k\neq s(i)}$. It is undesired and does not occur in the case of ϕ_{hk} and $\phi_{i,k\neq s(i)}$. Note that in SSFCMeans, u_{hk} is not impacted by the scaling factor α in any way, and this is why we postulate the same for $\hat{\mathbf{v}}_k$ when considering a contribution of the unsupervised observation h.

C. Semi-Supervised Possibilistic C-Means

We now apply the additive combination technique from (2) to introduce partial supervision to PCM. The idea of PCM comes from a relaxation of the probabilistic constraint in FCM presented in (4b). To avoid a trivial solution where each membership was estimated to be 0, a special form of the objective function was proposed in [24]:

$$Q_{\text{PCM}}(T,V;X,\Theta) = \sum_{k=1}^{c}\sum_{j=1}^{N} t_{jk}^m d_{jk}^2 + \sum_{k=1}^{c}\gamma_k\sum_{j=1}^{N}(1-t_{jk})^m, \tag{16}$$

where $T = [t_{jk}]$ is a typicalities matrix, and vector $\Gamma = (\gamma_1,\ldots,\gamma_c)^T$ contains cluster-specific scalars $\gamma_k > 0$. Note that [24, p. 101] allowed $m \in (1,\infty)$, but recall that in this article we set $m = 2$.

The supervised component Q_{SSPCM}^S that we propose is the same as in [27]

$$Q_{\text{SSPCM}}^S(T,V;X,F,\Theta) = \sum_{k=1}^{c}\sum_{j=1}^{N} b_j\cdot(t_{jk}-f_{jk})^2\cdot d_{jk}^2. \tag{17}$$

Since we regard classical approaches in this article, we propose to combine (17) with (16) to obtain the objective function

$$\begin{aligned} J_{\text{SSPCM}}(T,V;X,F,\Theta) = Q_{\text{PCM}}(T,V;X,\Theta) \\ + \alpha\cdot Q_{\text{SSPCM}}^S(T,V;X,F,\Theta). \end{aligned} \tag{18}$$

The minimization problem is

$$\underset{T,V}{\arg\min} \quad J_{\text{SSPCM}}(T,V;X,F,\Theta) \tag{19a}$$
$$\text{s.t.} \quad 0 < \sum_{j=1}^{N} t_{jk} < N \quad \forall k = 1,\ldots,c, \tag{19b}$$
$$t_{jk} \in [0,1]. \tag{19c}$$

where the constraints (19b) and (19c) are the same as in the unsupervised PCM.

Compared with our approach, Antoine et al. [27] combined (17) with the objective function of Repulsive PCM, defined as

$$Q_{\text{RPCM}}(T, V; X, \Theta) = Q_{\text{PCM}} + \sum_{k=1}^{c} \eta_k \sum_{l \neq k} \frac{1}{\|\mathbf{v}_k - \mathbf{v}_l\|^2}. \quad (20)$$

However, since the objective functions (16) and (20) include t_{jk} in the same way, the formula for the optimal typicality in SSPCMeans is thus the same as derived in [27] and presents as

$$\hat{t}_{jk} = \frac{\gamma_k + \alpha \cdot b_j \cdot d_{jk}^2 \cdot f_{jk}}{\gamma_k + (\alpha \cdot b_j + 1) d_{jk}^2}. \quad (21)$$

Considering distinct types of $\hat{t}_{jk}$:

$$\hat{t}_{hk} = \frac{\gamma_k}{\gamma_k + d_{hk}^2}, \quad (22a)$$

$$\hat{t}_{i,k \neq s(i)} = \frac{\gamma_k}{\gamma_k + (\alpha + 1) \cdot d_{i,k \neq s(i)}^2}, \quad (22b)$$

$$\hat{t}_{i,s(i)} = \frac{\gamma_{s(i)} + \alpha \cdot d_{i,s(i)}^2}{\gamma_{s(i)} + (\alpha + 1) \cdot d_{i,s(i)}^2}. \quad (22c)$$

The optimal cluster's prototype in our SSPCMeans differs from [27], because (16) and (20) differ in treating V, and has a form

$$\hat{\mathbf{v}}_k = \frac{\sum_{j=1}^{N} \left(t_{jk}^2 + \alpha b_j (t_{jk} - f_{jk})^2 \right) \cdot \mathbf{x}_j}{\sum_{j=1}^{N} t_{jk}^2 + \alpha b_j (t_{jk} - f_{jk})^2}, \quad (23)$$

where derivation is analogous to the one presented for SSFCMeans in the previous subsection.

III. Explanations of the scaling factor α

Despite a wealth of literature spanning over 30 years on the topic of SSFC, surprisingly little attention was paid to the sound understanding of the scaling factor α. One of the main contributions of this article is that we systematically reviewed existing descriptions of the scaling factor [5]–[23], [27], [28] and concluded that these are highly alike and do not challenge the core meaning of the interpretation of the scaling factor α provided by Pedrycz and Waletzky in [3]. Therefore, we treat the interpretation from [3] as canonical and formulate

Interpretation 1: The role of the scaling factor α is to maintain a balance between the supervised and unsupervised components within the optimization mechanism.

A critical issue with Interpretation 1 is that the scaling factor α is considered only in the context of the objective function. We thus extend this definition and provide a discussion about a connection between the scaling factor and the outcome of the model, i.e. the estimated memberships matrix $\hat{U}$.

Furthermore, the descriptions of the scaling factor α in the literature are imprecise and inconsistent. Below we list selected citations that use naming conventions different than Interpretation 1:

- "α be proportional to the rate N/M" [3, p. 788];
- "α parameter is set in such a way that two terms of the objective function have the same importance" [33, p. 57];
- "β is the impact intensity of the semi-supervised component" [13, p. 671];
- "where λ is the ratio of labeled sample points in the data sample" [11, p. 135];
- "where λ_1 and λ_2 are the regularization parameters which control the tradeoff between FCM and SSFCM" [22, p. 387].

The terms "balance", "intensity" or "tradeoff" may implicate the proportional impact of the scaling factor α on the outcomes of the model, but do not have to. There are no clear statements about the functional character of the impact in the corresponding articles. Only Pedrycz and Waletzky [3] use the word "proportional" directly, but they use it to establish α as a function of the data (the number of labeled observations), not to discuss how much α impacts the outcomes of modeling (regardless of the data).

The above problems lead to inconsistent processes of selecting the value of the hyperparameter α that are not justified analytically. The importance of the scaling factor α is clearly seen in (2). Regardless of the functional form of Q or Q^S, the role of α is the same. It clearly impacts the estimated variables $\hat{U}$ and $\hat{V}$.

A. Differences between interpretation and explanation

To distinguish between an interpretation and an explanation, we propose 3 criteria that an explanation of the scaling factor α must satisfy: (C1) interpretability, (C2) completeness, (C3) quantification.

Any description that satisfies criterion (C1) and one more criterion (C1 or C2), but not all 3 criteria, is considered an *interpretation*. Gilpin et al. [30] provide two criteria for evaluating explanations: interpretability and completeness, which are referred to as (C1) and (C2) in this article. Criterion (C3) is our additional requirement specific for the scaling factor α: we want to express the impact of partial supervision as a function IPS(α).

Let us now elaborate on how to check criteria (C1) - (C3) for a given description of the scaling factor α. For (C1) *interpretability*, Broniatowski [29] states that "an interpretable model should provide users with a description of what a stimulus (a data point or model's output) means in context". Regarding SSFC, the scaling factor α is the stimulus we require to be put in a context. Moreover, it does not suffice to provide any context as an interpretable description should be "understandable to humans" [30].

(C2) *completeness* is satisfied when a description of the system's operation is accurate [30]. We associate this criterion with a proposition from [29] "an explanation of a model result is a description of how a model's outcomes came to be". Note that in the case of SSFC, the key outcome of the model is the estimated memberships matrix $\hat{U}$. Taking all the above into account, we require a complete description of the scaling factor

α to describe in an accurate way the relationship between α and $\hat{U}$.

Finally, criterion (C3) *quantification* stems from the need to numerically assess the difference between an impact of different α_1 and α_2, $\alpha_1 \neq \alpha_2$ values on the results of SSFC model. Explainable impact of partial supervision must associate a function IPS(α) that allows calculation of a difference IPS(α_1) - IPS(α_2).

B. Explanation of the scaling factor α in Semi-Supervised Fuzzy C-Means

It is clearly shown in (10) that for $\hat{u}_{i,s(i)}$, regardless of the data evidence, we are guaranteed that

$$\hat{u}_{i,s(i)} > \frac{\alpha}{1+\alpha}. \tag{24}$$

We propose to call the quantity $\frac{\alpha}{1+\alpha}$ the Absolute Lower Bound to stress its nature. To our knowledge, the Absolute Lower Bound has not been discussed in the literature so far even though it is a straightforward conclusion that stems from well-known equations and may significantly impact the outcomes of the model. Let us now formulate an explanation of the impact of partial supervision in SSFCMeans.

Explanation 1 (IPS in SSFCMeans): The scaling factor α quantifies the impact of partial supervision as IPS(α) = $\frac{\alpha}{1+\alpha}$, and establishes an Absolute Lower Bound for a membership of a supervised observation to the supervised cluster $u_{i,s(i)} >$ IPS(α).

C. Explanation of the scaling factor α in Semi-Supervised Possibilistic C-Means

Let us first consider an interpretation of the hyperparameter γ_k provided in [24].

Interpretation 2: The value of γ_k determines the distance at which the typicality value of a point in a cluster becomes 0.5.

It comes from the fact that if we consider a distance $d^2_{hk} := \gamma_k$, then for the typicality in unsupervised PCM (22)

$$t_{hk} = \frac{\gamma_k}{\gamma_k + d^2_{hk}} \overset{d^2_{hk}:=\gamma_k}{=} \frac{\gamma_k}{\gamma_k + \gamma_k} = 0.5. \tag{25}$$

With the aim of providing an explanation of the scaling factor α in Semi-Supervised Possibilistic C-Means, we will make a similar assumption and study the difference between: (I) a possibility of a supervised observation to the supervised cluster $t_{i,s(i)}$ from (22c) and (II) a possibility of unsupervised observation to any cluster t_{hk} from (22).

Let us consider arbitrary observation a and arbitrary cluster b. First, assume $t^{(I)}_{ab}$ is unsupervised typicality to any cluster, as in (22). We know that if we set $\gamma_b := d^2_{ab}$, then the typicality $t^{(I)}_{ab}$ is 0.5.

Suppose that we obtain the label of observation a so it becomes supervised, and $b = s(a)$ happens to be the supervised cluster. Therefore, the typicality value takes form from (22c), and assuming $d^2_{ab} = \gamma_b$

$$t^{(II)}_{ab} \overset{d^2_{ab}:=\gamma_b}{=} \frac{\gamma_b + \alpha\gamma_b}{\gamma_b + (\alpha+1)\gamma_b} = \frac{1+\alpha}{2+\alpha}. \tag{26}$$

TABLE I
A COMPARISON OF THE INTERPRETATION 1 OF THE SCALING FACTOR α WITH TWO NOVEL EXPLANATIONS PROPOSED.

description	criteria[1]		
	C1	C2	C3
Interpretation 1	+	−	±
Explanation 1 (SSFCMeans)	+	+	+
Explanation 2 (SSPCMeans)	+	+	+

[1] Convention: + means a criterion was met, − means it was not; ± denotes a partially met criterion.

Note that the only change includes the value of typicality $t^{(II)}_{ab}$: this is still the same observation a, the same cluster b, the same hyperparameter γ_b, and the same fixed distance d^2_{ab}.Therefore, we can quantify the impact of partial supervision

$$\text{IPS}(\alpha) = t^{(II)}_{ab} - t^{(I)}_{ab} = \frac{1+\alpha}{2+\alpha} - \frac{1}{2} = \frac{\alpha}{2(2+\alpha)}. \tag{27}$$

We can now propose an explanation of the scaling factor α.

Explanation 2 (IPS in SSPCMeans): In the supervised case, the scaling factor α increases the typicality of a supervised observation to the supervised cluster $t_{i,s(i)}$ by IPS(α)=$\frac{\alpha}{2(2+\alpha)}$ for the same distance $\gamma_{s(i)}$ at which the typicality in the unsupervised case was equal 0.5.

D. Checking the criteria

Table I contains a comparison of the Interpretation 1 with two new explanations of the scaling factor α proposed in this article with respect to criteria (C1)-(C3). First and foremost, all the descriptions considered in Table I meet the criterion (C1) *interpretability*. They put the role of the scaling factor α in a broader context of the model in a "human understandable" language.

Regarding the criterion (C2) *completeness*, let us recall that Interpretation 1 relates α with the objective function. The implicit statement "(...) a balance between the supervised and unsupervised component (...)" [3, p. 789] means in fact "a balance between the supervised and unsupervised component of the objective function". It is unclear from this interpretation how the outcome $\hat{U}$ of the SSFCMeans model came to be since Interpretation 1 does not relate α to the variable $\hat{u}_{jk}$. On the contrary, both Explanation 1 (SSFCMeans) and Explanation 2 (SSPCMeans) explain the scaling factor α in terms of its impact on the soft assignment variables by precise referral to the models' mechanisms. Explanation 1 relates IPS to the membership of a supervised observation to the supervised cluster $u_{i,s(i)}$, and Explanation 2 discusses IPS in terms of a difference between the supervised typicality $t_{i,s(i)}$ and the typicality as if the observation was treated as unsupervised.

Regarding the criterion (C3) *quantification*, Pedrycz and Waletzky [3] suggested that the value of α should be set to the rate $\frac{M}{N}$, relating it to the data. We enhance this proposition and express it in terms of the impact of partial supervision as a function $\text{IPS}_{\text{P97}}(\alpha) = \alpha$. Nonetheless, we show that the impact of partial supervision is not directly proportional to α.

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2024.3370768

IV. Practical considerations

In the preceding sections, our analyses have contributed to the establishment of theoretically sound explanations regarding the impact of partial supervision in semi-supervised fuzzy clustering. In the context of SSFCMeans and SSPCMeans models, this impact is regulated by the scaling factor α. Despite the provided explanations, practical questions arise: which values of α should be used? How can one empirically assess the impact of partial supervision when fitting the model to the data? In this Section, we build on the results from the preceding analyses and delve into these specific practical considerations.

The source code for reproducible simulations described in Section IV-B is publicly available on CodeOcean [34]. In the absence of open-source implementations of SSFCMeans, we implemented it in `R` language from scratch and made it publicly available on GitHub[1].

A. Constructing cross-validation grids

A standard practice for selecting the value of a hyperparameter of any model is to cross-validate (CV) it, i.e. to create a $K-$tuple of K different values to be checked (called a *grid*), fit a model for each value, and finally find the best model with respect to some criterion; the selected value of the hyperparameter is the one associated with the best model. In the SSFC domain, a common CV approach is to select a few α values that divide the search space roughly equally.

For instance, Bouchachia and Pedrycz [8] tried $\text{grid}_B = \langle 0.3, 0.5, 0.7, 0.9, 1\rangle$, whereas Antoine et al. [28] tried $\text{grid}_A = \langle 0.01, 0.05, 0.1, 0.5, 1\rangle$. These CV grids cover the space of α values, since they implicitly follow Interpretation 1 and the associated proportionality assumption that was expressed as $\text{IPS}_{\text{P97}}(\alpha) = \alpha$. Such a function has a significant analytical disadvantage: it is bounded only from below. Theoretically, using Interpretation 1, one could think about increasing the value of α infinitely, expecting that each increase in α will result in the directly proportional increase of the impact of partial supervision. In practice, none of the works reviewed in Section III analyzed this issue, and a maximum value of α considered in CV rarely exceeds 1 (as can be seen in grid_A and grid_B).

On the contrary, IPS(α) functions for both Explanation 1 and Explanation 2 do not suffer from such problems. They are non-linear, monotonically increasing functions of α bounded from up and below. Their properties enable an analytically justified procedure tailored to creating CV grids for α in SSFC. One can analyze the derivative IPS'(α) and decide on a point where the decrease in IPS becomes negligible. We call this point a β boundary. Fig. 1 presents IPS functions together with derivatives IPS'=$\partial\text{IPS}/\partial\alpha$ for SSFCMeans and SSPCMeans models. Fig. 2 contains the proposed Algorithm for selecting the α grid based on β boundary.

Let us construct an exemplary CV grid for the SSFCMeans model that we call grid_{IPS}. Examining Fig. 1, one can decide on a boundary IPS'=β beyond which the increase in IPS is negligible. Since in SSFCMeans IPS(α) = $\frac{\alpha}{1+\alpha}$, we arrive at the equation

$$\frac{\partial \text{IPS}}{\partial \alpha}(\alpha_\beta) = \frac{1}{(1+\alpha_\beta)^2} = \beta, \tag{28}$$

with α_β corresponding to the chosen boundary being

$$\alpha_\beta = \beta^{-1/2} - 1. \tag{29}$$

Let us set $\beta = 0.2$, and according to (29), a corresponding $\alpha_{0.2} \approx 1.24$. Further on, we calculate $\text{IPS}(\alpha_{0.2}) \approx 0.55$. For a $5-$fold CV, a single step is equal to $\frac{0.55}{5} = 0.11$, so that $\text{grid}_{\text{IPS}}\big(\text{IPS}(\alpha)\big) = \langle 0.11, 0.22, 0.33, 0.44, 0.55\rangle$. Translating this in terms of α, the final $\text{grid}_{\text{IPS}}(\alpha) = \langle 0.12, 0.28, 0.49, 0.79, 1.22\rangle$.

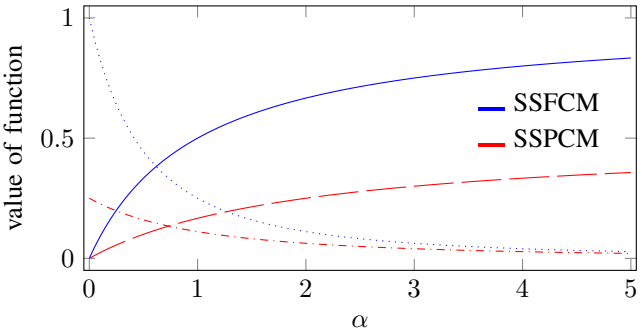

Fig. 1. The impact of partial supervision IPS(α) for $\alpha \in [0, 5]$ for both SSFCMeans and SSPCMeans. The IPS(α) for SSFCMeans is shown as a solid blue line, and the corresponding derivative is shown as a dotted blue line. The IPS(α) for SSPCMeans is shown as a dashed red line, and the corresponding derivative is shown as a red dash-dotted line.

1: choose a boundary value β beyond which IPS'(α) is treated as negligible,
2: retrieve α_β from the equation IPS'(α_β) = β,
3: calculate the value of IPS(α_β),
4: decide on the number of folds K in CV procedure,
5: calculate step_K = IPS(α_β) $\cdot \frac{1}{K}$,
6: derive $\text{grid}\big(\text{IPS}(\alpha)\big) = \langle \text{step}_K, 2 \cdot \text{step}_K, \ldots, K \cdot \text{step}_K\rangle$,
7: derive grid(α) by applying IPS^{-1} to each element of $\text{grid}\big(\text{IPS}(\alpha)\big)$.

Fig. 2. Algorithm for establishing cross-validation grid for α.

B. Empirical impact of partial supervision

When working with data, a need frequently occurs to ascertain how the introduction of partial supervision alters the outcomes of modeling a given dataset when contrasted with lack of supervision, or when the impact of partial supervision is reduced by a certain factor (e.g., twofold). We call it the analysis of the empirical impact of partial supervision, as it depends not only on the theoretical explanations but also on the specific data patterns. In the context of the SSFCMeans model, we postulate that the examination of the distribution of supervised memberships $\{u_{i,s(i)}\}_{i=1,\ldots,M}$ is not the optimal choice albeit an intuitive one. This is due to the combined

[1]https://github.com/ITPsychiatry/ssfclust

theoretical and *empirical* nature of $u_{i,s(i)}$ from (10). Denoting this membership in the functional convention, we obtain

$$\begin{aligned} u_{i,s(i)} &= u\big(\mathbf{x}_i, V, k = s(i), \alpha\big) \\ &= \frac{1}{1+\alpha} \cdot e\big(\mathbf{x}_i, V, k = s(i)\big) + \frac{\alpha}{1+\alpha}. \end{aligned} \tag{30}$$

It is the data evidence $e_{i,s(i)}$ that contains the truly empirical impact of the partial supervision, as it is the direct function of the data. Therefore, analysis of the distribution $\{e_{i,s(i)}\}$ enables a direct investigation of the extent to which the impact of partial supervision affected the prototypes, and consequently, the relative distances between the observation and these prototypes in a given model.

Let us now illustrate the above approach in a concrete data analysis scenario. We consider a 3-class semi-supervised problem, i.e. $Y = \langle y_1, y_2, y_3 \rangle$. The data is simulated in a nested loop. The outer loop consists of sampling 100 observations for each class from a two-dimensional Gaussian distribution $\mathcal{N}_2(\mu_k, \Sigma_k)$, where $\mu_1 = (5,5)^T$, $\mu_2 = (7,7)^T$, $\mu_3 = (9,9)^T$, and $\Sigma_1 = \Sigma_2 = \Sigma_3 = \text{diag}(5,5)$. Each kth distribution is associated with the kth class. Such a procedure yields spherical, overlapping clusters, which are hardly separable. An outcome of this outer loop is a features matrix $X_{[300,2]}$. Fig. 3 presents an example of such a matrix with colors and shapes denoting the classes of observations. An inner loop relies on randomly selecting 15% observations from each class that will remain supervised (leading to 45 observations treated as supervised in each simulated dataset). We performed 10 outer loops with 10 inner loops for each simulated X, arriving at 100 simulation runs.

We now build CV grids. Specifically, we compare a proposition from the literature $\text{grid}_B(\alpha) = \langle 0.3, 0.5, 0.7, 0.9, 1 \rangle$ [8] with $\text{grid}_{\text{IPS}}(\alpha) = \langle 0.12, 0.28, 0.49, 0.79, 1.22 \rangle$ that we constructed based on the Algorithm from Fig. 2 proposed in this work. The results for these grids are presented against a dense reference grid_{ref} composed of 50 α values dividing the interval $[0, 1.5]$ equally (the equivalent interval expressed in terms of IPS(α) is $[0, 0.6]$). Owing to grid_{ref}, we can observe a global pattern that one typically does not examine due to time and computational resource constraints.

Fig. 4 presents the summary of the results of fitting the SSFCMeans model to the data from each simulation run $r = 1, \ldots, 100$ for each α from the respective grid. We present the *total median* $\bar{e}(\alpha)$

$$\bar{e}(\alpha) = \text{Me}\big(\{e^{\alpha,r=1}_{1,s(1)}, e^{\alpha,r=2}_{1,s(1)}, \ldots, e^{\alpha,r=1}_{45,s(45)}, \ldots, e^{\alpha,r=100}_{45,s(45)}\}\big) \tag{31}$$

together with the interquartile range (IQR). For the range IPS(α) $\in [0, 0.25]$, $\bar{e}(\alpha)$ is growing approximately proportionally to $\frac{\alpha}{1+\alpha}$, which confirms the theoretical quantification of the impact of partial supervision. Starting at IPS(α) ≈ 0.25, the total median reaches the value of ≈ 0.55 and remains stable regardless of the increasing IPS(α). The exact results – in the form of pairs (IPS(α), $\bar{e}(\alpha)$) – present as: (0.12, 0.38), (0.28, 0.54), (0.49, 0.57), (0.79, 0.57), (1.22, 0.57) for grid_{IPS} and (0.3, 0.54), (0.5, 0.57), (0.7, 0.57), (0.9, 0.57), (1, 0.57) for grid_{B}.

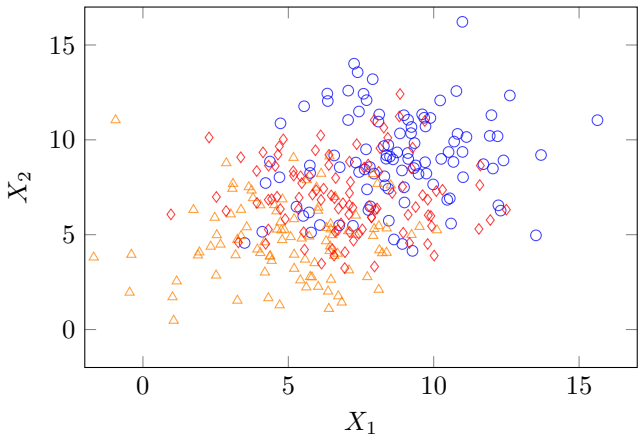

Fig. 3. An example of a single simulated features matrix $X_{[300,2]}$. The orange triangles represent data points belonging to class y_1, the red diamonds represent data points belonging to class y_2, and the blue circles represent data points belonging to class y_3.

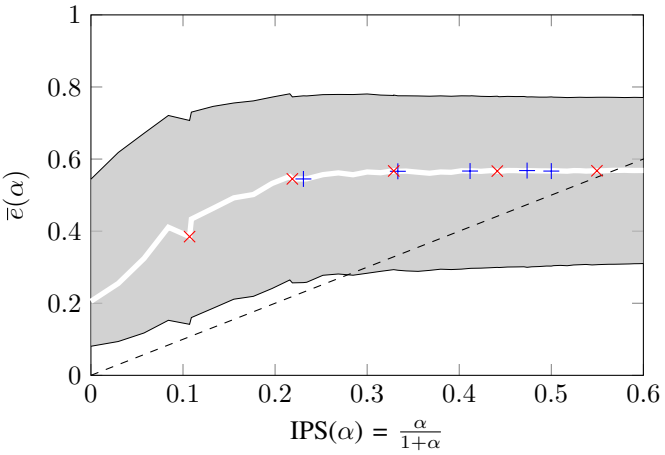

Fig. 4. Simulation results for $\bar{e}(\alpha)$ presented against IPS(α). Solid black lines represent Q1 and Q3 for grid_{ref}, the grey area represents IQR, and white line represents total median. Red crosses represent total medians for grid_{IPS}, and blue pluses represent total medians for grid_{B}. The black dotted line corresponds to IPS(α) = $\frac{\alpha}{1+\alpha}$.

The growth of $\bar{e}(\alpha)$ described above is associated with the increasing quality of true clusters' prototypes estimation. Table IV-B presents mean estimated prototypes coordinates $\hat{V}_1$ and $\hat{V}_2$ together with their standard deviations for models for 3 values of IPS(α): 0 denoting no supervision at all, 0.12 and 0.28 being two first entries from grid_{IPS} that enable to grasp the trends in simulation results described above. The total median $\bar{e}(\alpha)$ reaches a plateau at approximately IPS(α) = 0.28, since the SSFCMeans already identified the true clusters' prototypes. The model cannot result in higher median data evidence $e_{i,s(i)}$ despite the increasing impact of partial supervision due to the noise in the data. This is an example of the empirical impact of partial supervision deviating from the theoretical one.

Finally, let us note the differences between grid_{IPS} and grid_B. The former splits the IPS(α) space in equal intervals and allows to identify a changing trend in the behavior of $\bar{e}(\alpha)$ as compared with the latter, which covers a narrower interval of IPS(α). This specific simulation scenario confirms the need

TABLE II
MEAN ESTIMATED CLUSTERS' PROTOTYPES TOGETHER WITH THE STANDARD DEVIATIONS (IN BRACKETS) FOR SELECTED VALUES OF α

	cluster 1		cluster 2		cluster 3	
IPS(α)	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_1$	$\hat{V}_2$
0	7.29 (2.02)	7.16 (1.99)	6.86 (2.06)	7.06 (1.95)	7.02 (2.16)	6.71 (2.02)
0.12	5.71 (1.6)	5.71 (1.5)	7.05 (1.85)	6.89 (1.77)	8.35 (1.66)	8.32 (1.63)
0.28	4.78 (0.85)	4.93 (0.72)	7.19 (1.15)	6.82 (1.11)	9.16 (0.78)	9.15 (0.91)

TABLE III
DATA FOR EXEMPLARY CONFIDENCE PATH REGULARIZATION PROCEDURE, $\alpha = 2$

columns used in [23]				explainability approach	
reg_r	α_r	w_r	$\hat{u}^r_{i,s(i)}$	IPS(α_r)	$e^r_{i,s(i)}$
0.25	0.5	4	0.34	0.33	0.01
0.5	1	2	0.51	0.5	0.02
1	2	1	0.67	0.66	0.01

for the analytically justified creation of CV grids presented in the Algorithm from Fig. 2.

C. Estimating label uncertainty

In the previous subsection, we knew the process generating the data, hence the obtained labels y_i were certain. In practice, this process is typically unknown, therefore the certainty of the labels may be questioned. [3] proposed to handle this situation by incorporating a confidence factor $\text{conf}_j \in [0, 1]$ to the objective function of SSFCMeans. However, their approach requires to assess the uncertainty upfront. Frequently, such knowledge is not available, especially when the data annotation process is a complex one [23].

To overcome this problem, [23] proposed the Confidence Path Regularization (CPR) procedure to estimate the adjusted confidence factor $\text{conf}^\star_i$ from the data. CPR wraps SSFCMeans, implementing the *regularization* assumption: highly certain supervised observations should be consistently assigned high $u_{i,s(i)}$ across varying values of conf_i. A path of $r = 1, \ldots, R$ models is fitted, each decreasing the default α uniformly for all observations by $\text{conf}_i = \text{reg}_r \forall i$. The adjusted $\text{conf}^\star_i$ for ith observation is then obtained as a weighted summary of the memberships from R models

$$\text{conf}^\star_i = \frac{1}{\sum_{r=1}^{R} w_r} \cdot \sum_{r=1}^{R} u^r_{i,s(i)} w_r, \tag{32}$$

where weights w_r compensate for the decreased $\alpha_r = \alpha \cdot \text{reg}_r$. [23] proposed to use the proportionality rule, i.e., set $w_r = \frac{1}{\text{reg}_r}$. The first 4 columns of Table III contain exemplary data required to calculate $\text{conf}^\star_i$ in a CPR procedure composed of $R = 3$ steps.

Note that the above procedure is implicitly based on Interpretation 1 quantifying the impact of partial supervision as $\text{IPS}_{\text{P97}}(\alpha) = \alpha$, and hence may lead to inaccurate conclusions. For the example from Table III, the adjusted $\text{conf}^\star_i = 0.43$, and we conclude that this ith observation is not the most certain labeled observation, but definitely not the least certain one. However, if we focus on the information contained in 2 last columns of Table III, we clearly see that the above conclusion is inaccurate, as the data evidence is extremely low; this labeled observation should be thus considered as highly uncertain. This exemplary problem shows potential issues resulting from the use of incorrect quantification of the impact of partial supervision and motivates the introduction of explainability framework into the procedures such as CPR.

V. CONCLUSIONS

The scaling factor α weighs the impact of partial supervision in semi-supervised fuzzy clustering and thus has a substantial effect on the estimated memberships and clusters' prototypes. All the models building on the additive combination technique introduced in [3], ranging from semi-supervised adaptations of Possibilistic Fuzzy C-Means [28] to complex workflows that wrap the SSFCMeans model [23], share the same mechanism of regulating the impact of partial supervision by means of the scaling factor α.

We reviewed the existing interpretations of α and its relationship with the impact of partial supervision and concluded that these interpretations are imprecise. They lack completeness, since they interpret α only in terms of the objective function, not the membership degrees. They also suggest a directly proportional relationship between the impact of partial supervision on the memberships and the scaling factor, which we prove to be non-linear.

Therefore, in this article, we introduced model-specific explanations of the scaling factor α for both SSFCMeans and SSPCMeans that overcome the aforementioned limitations. They fulfill the three necessary criteria of an explanation (*interpretability*, *completeness* and *quantification*) that we proposed based on the discussions on the explainability framework [29]–[31]. Each explanation defines an associated function IPS(α) that quantifies the impact of partial supervision on the memberships.

The benefits of using our novel explanations are substantial. Not only do the explanations clarify the role of α, but also prove its impact to be a non-linear bounded function of α. This enables analytically justified procedures for selecting the value of α to use, such as building cross-validation grids based on IPS functions proposed in the Algorithm from Fig. 2. We also discussed the differences between theoretical and empirical impact of partial supervision, providing a simulation example to illustrate them.

Explanation 1 is of particular importance for procedures that estimate label uncertainty, such as Confidence Path Regularizatio [23]. The concepts of Absolute Lower Bound and data evidence encourage treating label uncertainty with respect to the ALB rather than to the nominal supervised membership.

Finally, further assessment of modeling the impact of partial supervision in the spirit of the additive combination technique remains open for future work. Firstly, Explanation 2 for

SSPCMeans requires a simulation or real-life data experiment that we performed for SSFCMeans only. Secondly, it seems a promising direction to assess if one could introduce custom flexibility into the shape of the Absolute Lower Bound curve $\frac{\alpha}{1+\alpha}$ from Explanation 1.

REFERENCES

[1] O. Chapelle, B. Schölkopf, and A. Zien, Eds., Semi-Supervised Learning, ser. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 2006.

[2] D. T. C. Lai and J. M. Garibaldi, "A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms," in 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011). Taipei, Taiwan: IEEE, Jun. 2011, pp. 1580–1586.

[3] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 27, no. 5, pp. 787–795, Oct. 1997.

[4] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, no. 2-3, pp. 191–203, Jan. 1984.

[5] C. Stutz and T. Runkler, "Classification and prediction of road traffic using application-specific fuzzy clustering," IEEE Transactions on Fuzzy Systems, vol. 10, no. 3, pp. 297–308, Jun. 2002.

[6] Chunfang Li, Lianzhong Liu, and Wenli Jiang, "Objective function of semi-supervised Fuzzy C-Means clustering algorithm," in 2008 6th IEEE International Conference on Industrial Informatics. Daejeon, South Korea: IEEE, Jul. 2008, pp. 737–742.

[7] E. Yasunori, H. Yukihiro, Y. Makito, and M. Sadaaki, "On semi-supervised fuzzy c-means clustering," in 2009 IEEE International Conference on Fuzzy Systems. Jeju Island, South Korea: IEEE, Aug. 2009, pp. 1119–1124.

[8] A. Bouchachia and W. Pedrycz, "Enhancement of fuzzy clustering by mechanisms of partial supervision," Fuzzy Sets and Systems, vol. 157, no. 13, pp. 1733–1759, Jul. 2006.

[9] ——, "A Semi-supervised Clustering Algorithm for Data Exploration," in Fuzzy Sets and Systems — IFSA 2003, J. G. Carbonell, J. Siekmann, T. Bilgiç, B. De Baets, and O. Kaynak, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2715, pp. 328–337.

[10] J. Gao, P.-N. Tan, and H. Cheng, "Semi-Supervised Clustering with Partial Background Information," in Proceedings of the 2006 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Apr. 2006, pp. 489–493.

[11] J. Xu, G. Feng, T. Zhao, X. Sun, and M. Zhu, "Remote sensing image classification based on semi-supervised adaptive interval type-2 fuzzy c-means algorithm," Computers & Geosciences, vol. 131, pp. 132–143, Oct. 2019.

[12] L. Liu and X.-J. Wu, "Semi-Supervised Possibilistic Fuzzy c-Means Clustering Algorithm on Maximized Central Distance," in Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013). China: Atlantis Press, 2013.

[13] F. Salehi, M. R. Keyvanpour, and A. Sharifi, "SMKFC-ER: Semi-supervised multiple kernel fuzzy clustering based on entropy and relative entropy," Information Sciences, vol. 547, pp. 667–688, Feb. 2021.

[14] G. Casalino, M. Dominiak, F. Galetta, and K. Kaczmarek-Majer, "Incremental Semi-Supervised Fuzzy C-Means for Bipolar Disorder Episode Prediction," in 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). Bari, Italy: IEEE, May 2020, pp. 1–8.

[15] G. Casalino, G. Castellano, F. Galetta, and K. Kaczmarek-Majer, "Dynamic Incremental Semi-supervised Fuzzy Clustering for Bipolar Disorder Episode Prediction," in Discovery Science, A. Appice, G. Tsoumakas, Y. Manolopoulos, and S. Matwin, Eds. Cham: Springer International Publishing, 2020, vol. 12323, pp. 79–93.

[16] G. Casalino, G. Castellano, and C. Mencar, "Data Stream Classification by Dynamic Incremental Semi-Supervised Fuzzy Clustering," International Journal on Artificial Intelligence Tools, vol. 28, no. 08, p. 1960009, Dec. 2019.

[17] K. Kaczmarek-Majer, G. Casalino, G. Castellano, O. Hryniewicz, and M. Dominiak, "Explaining smartphone-based acoustic data in bipolar disorder: Semi-supervised fuzzy clustering and relative linguistic summaries," Information Sciences, vol. 588, pp. 174–195, Apr. 2022.

[18] H. Gan, Z. Yang, and R. Zhou, "Adaptive safety-aware semi-supervised clustering," Expert Systems with Applications, vol. 212, p. 118751, Feb. 2023.

[19] P. T. Huan, P. H. Thong, T. M. Tuan, D. T. Hop, V. D. Thai, N. H. Minh, N. L. Giang, and L. H. Son, "TS3FCM: Trusted safe semi-supervised fuzzy clustering method for data partition with high confidence," Multimedia Tools and Applications, vol. 81, no. 9, pp. 12 567–12 598, Apr. 2022.

[20] L. Guo, H. Gan, S. Xia, X. Xu, and T. Zhou, "Joint exploring of risky labeled and unlabeled samples for safe semi-supervised clustering," Expert Systems with Applications, vol. 176, p. 114796, Aug. 2021.

[21] H. Gan, Y. Fan, Z. Luo, R. Huang, and Z. Yang, "Confidence-weighted safe semi-supervised clustering," Engineering Applications of Artificial Intelligence, vol. 81, pp. 107–116, May 2019.

[22] H. Gan, Y. Fan, Z. Luo, and Q. Zhang, "Local homogeneous consistent safe semi-supervised clustering," Expert Systems with Applications, vol. 97, pp. 384–393, May 2018.

[23] K. Kmita, G. Casalino, G. Castellano, O. Hryniewicz, and K. Kaczmarek-Majer, "Confidence path regularization for handling label uncertainty in semi-supervised learning: use case in bipolar disorder monitoring," in 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2022, pp. 1–8.

[24] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pp. 98–110, May 1993.

[25] N. Pal, K. Pal, J. Keller, and J. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 517–530, Aug. 2005.

[26] H. Timm, C. Borgelt, C. Döring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis," Fuzzy Sets and Systems, vol. 147, no. 1, pp. 3–16, Oct. 2004.

[27] V. Antoine, J. A. Guerrero, T. Boone, and G. Romero, "Possibilistic clustering with seeds," in 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Rio de Janeiro: IEEE, Jul. 2018, pp. 1–7.

[28] V. Antoine, J. A. Guerrero, and G. Romero, "Possibilistic fuzzy c-means with partial supervision," Fuzzy Sets and Systems, vol. 449, pp. 162–186, Nov. 2022.

[29] D. A. Broniatowski, "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence," National Institute of Standards and Technology, Tech. Rep., Apr. 2021.

[30] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Turin, Italy: IEEE, Oct. 2018, pp. 80–89.

[31] A. Adadi and M. Berrada, "Explainable AI for Healthcare: From Black Box to Interpretable Models," in Embedded Systems and Artificial Intelligence, V. Bhateja, S. C. Satapathy, and H. Satori, Eds. Singapore: Springer Singapore, 2020, vol. 1076, pp. 327–337.

[32] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Boston, MA: Springer US, 1981.

[33] V. Antoine and N. Labroche, "Semi-supervised Fuzzy c-Means Variants: A Study on Noisy Label Supervision," in Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, J. Medina, M. Ojeda-Aciego, J. L. Verdegay, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, and R. R. Yager, Eds. Cham: Springer International Publishing, 2018, vol. 854, pp. 51–62.

[34] K. Kmita, "Source Code for Explainable Impact of Partial Supervision on Memberships in Semi-Supervised Fuzzy Clustering", 2023 [Source Code]. https://doi.org/10.24433/CO.3328061.v2.