Mapping Points Back from the Concept Space with Minimum Mean Squared Error

Wladyslaw Homenda, Tomasz Penza

Mapping points back from the concept space with minimum mean squared error

Wladyslaw Homenda[1,2] and Tomasz Penza[2]

[1] Faculty of Economics and Informatics in Vilnius, University of Bialystok
Kalvariju g. 135, LT-08221 Vilnius, Lithuania
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract. In this article we present a method to map points from the concept space, associated with the fuzzy c–means algorithm, back to the feature space. We assume that we have a probability density function f defined on the feature space (e.g. a normalized density of a data set). For a given point $\boldsymbol{w}$ of concept space, we give explicitly a set of points in feature space that are mapped onto $\boldsymbol{w}$ and we give a formula for a reverse mapping to the feature space which results in minimum mean squared error, with respect to density f, of the operation of mapping a point of feature space into the concept space and back. We characterize the circumstances under which points can be mapped back into the feature space unambiguously and provide a formula for the inverse mapping.

1 Introduction

There is an effort among researchers today, to capture the mechanisms of human cognition that allow the mind to process abstract information. Their goal is to enable the machine to discern patterns, spot relations and make connections between objects and ideas, to reason and predict on the basis of knowledge and observation. To make this possible, computers must be able to form and process concepts. A cluster in a data set can be thought of as representing an abstract concept. We can thus use data clustering methods to extract concepts. With each clustering method, comes a way to measure to what degree a given data point belongs to a given cluster. If we regard these degrees of membership in different clusters as coordinates, we obtain a concept space of dimension equal to the number of clusters. We can then map numeric data from the feature space into the concept space and process it at the level of concepts. Many points of the feature space may end up being mapped onto the same point of concept space. Consequently, we run into a problem when we try to put the result of concept-level computations back into the feature space. In certain circumstances, if we want to have a reverse mapping from the concept space into the feature space, we have to accept that it will make errors, i.e. when we map an observation into the concept space, this reverse mapping will not map it back onto itself.

In this text we work with a transformation from the feature space to the concept space associated with the fuzzy c–means algorithm which we denote

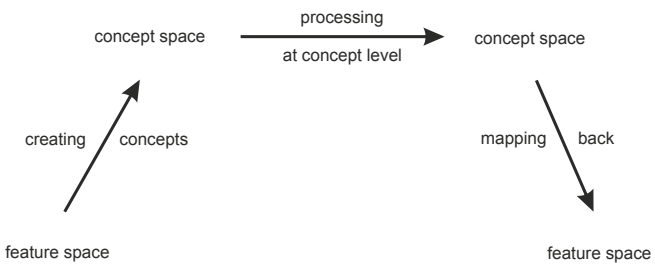

Fig. 1. Processing of data

with symbol $\mathcal{M}$. We will describe sets of points that $\mathcal{M}$ maps onto the same point of concept space. We will also determine under what conditions $\mathcal{M}$ is invertible and give a formula for its inverse. We will assume that there is a continuous probability density function defined on the feature space and we will present a formula for a reverse mapping G that allows minimum mean squared error (MSE), with respect to that probability density, of an operation of mapping an observation to the concept space via $\mathcal{M}$ and back via G. In practice such probability density might arise as a density of data points of some data set in the feature space. We propose an application of this reverse mapping as a defuzzification–like transformation of the results of concept–level processing, which is the last stage of the process outlined at figure 1.1. There is a long record of defuzzification methods and we refer to selected results in References without explicit discussion on them. Our approach is different in that we study a general reverse mapping and defuzzification discussed in this paper is a special case of such general transformation.

2 Preliminaries

2.1 Fuzzy c–means algorithm

An analogue of k–means algorithm within fuzzy clustering is the fuzzy c-means algorithm (FCM) which lets us find c fuzzy clusters in a data set $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ – a subset of feature space $\mathbb{R}^h$. We denote the i–th cluster by C_i and its centroid as $\boldsymbol{\mu}_i$. For a given observation $\boldsymbol{x}_i$, its degree of membership in the cluster C_j is denoted by w_{ij} and given by the formula

$$w_{ij} = \frac{||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^{-\frac{2}{m-1}}}{\sum_{k=1}^{c} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^{-\frac{2}{m-1}}} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||\boldsymbol{x}_i - \boldsymbol{\mu}_j||}{||\boldsymbol{x}_i - \boldsymbol{\mu}_k||} \right)^{\frac{2}{m-1}}}$$

where $m > 1$ is a parameter called the fuzzifier. For every data point, its membership in different clusters sums up to 1. Each fuzzy cluster is represented by its weighted mean $\boldsymbol{\mu}_i = \sum_{j=1}^{n} w_{ji}\boldsymbol{x}_j / \sum_{j=1}^{n} w_{ji}$. These two formulas are mutually

referential. At each step of the algorithm only one of them will be satisfied, but the clustering that the algorithm will output will approximately satisfy them both.

The fuzzifier m controls how fuzzy the resulting clustering will be. The higher the fuzzifier, the less distinct the values of membership and thus the fuzzier the clustering. When m converges to 1, for each observation one of its cluster memberships converges to 1 and the rest converge to 0, so in the limit the clustering becomes hard.

Before running the algorithm, we must choose the number c of clusters that we want to form, the fuzzifier m and the threshold ε. The threshold will let us terminate the algorithm when upon iteration all the memberships w_{ij} changed by less than ε. The algorithm is as follows.

1. Randomly initialize all the degrees of membership w_{ij} with numbers from the unit interval.
2. Compute all the weighted means $\boldsymbol{\mu}_i$ using current values w_{ij}.
3. Compute all the membership degrees w_{ij} using current values $\boldsymbol{\mu}_i$.
4. If any of values w_{ij} changed by more than ε, go to step 2.

FCM endeavors to minimize the loss function $\sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2$ which never increases after an iteration of the loop. Thanks to the threshold ε the algorithm always converges. The final clustering depends on the initial values w_{ij}, which are assigned randomly, thus the clusters will vary each time we run FCM. It is advised to run the algorithm several times and choose the best result. We can evaluate clusterings using cluster validity indices. They can also help us pick an optimal number of clusters c.

2.2 Mapping $\mathcal{M}$

Suppose that in the feature space $\mathbb{R}^h$, we used FCM with fuzzifier set to m to find c fuzzy clusters and we obtained the set of centroids $\mathcal{C} = \{\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2,\ \ldots,\ \boldsymbol{\mu}_c\}$. Mapping $\mathcal{M}_{\mathcal{C}} : \mathbb{R}^h \longrightarrow [0,1]^c$, assigns to $\boldsymbol{x} \in \mathbb{R}^h$ a vector of its degrees of membership in the respective clusters, which is a point of the concept space $[0,1]^c$. For $\boldsymbol{x} \in \mathbb{R}^h \backslash \mathcal{C}$, we have

$$\mathcal{M}_{\mathcal{C}}(\boldsymbol{x}) = \left(\frac{1}{\sum_{k=1}^{c} \left(\frac{||\boldsymbol{x} - \boldsymbol{\mu}_i||}{||\boldsymbol{x} - \boldsymbol{\mu}_k||} \right)^{\frac{2}{m-1}}} \right)_{i=1,\ 2,\ \ldots,\ c}$$

We extend $\mathcal{M}_{\mathcal{C}}$ continuously onto $\mathbb{R}^h$ by defining $\mathcal{M}_{\mathcal{C}}(\boldsymbol{\mu}_i)$ to be a vector whose i–th coordinate is 1 and all its other coordinates are 0. For now we will write simply $\mathcal{M}$, but we will invoke the subscript in section 3.3, when we will need to refer to mapping $\mathcal{M}_{\widehat{\mathcal{C}}}$ for a subset $\widehat{\mathcal{C}}$ of the set of centroids $\mathcal{C}$. $\mathcal{M}$ is continuously differentiable on $\mathbb{R}^h \backslash \mathcal{C}$. We will invert $\mathcal{M}$ in circumstances when it is possible and otherwise we will provide a formula for the most accurate reverse mapping to the feature space.

If we exclude the trivial case of just one centroid, the values of $\mathcal{M}$ on set $\mathbb{R}^h \backslash \mathcal{C}$ are all contained in the subset $W = \left\{(w_1,\ w_2,\ ...,\ w_c) \in (0,1)^c \ \middle|\ \sum_{i=1}^{c} w_i = 1\right\}$ of concept space. For $c = 3$ this subset would be an interior of a triangle inside the cube $[0,1]^3$ with vertices $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, i.e. points of the concept space corresponding to the centroids. Even though elements of W have c coordinates, they have only $c-1$ degrees of freedom, because the last coordinate is given by the first $c-1$ coordinates: $w_c = 1 - \sum_{i=1}^{c-1} w_i$. Let P denote the projection onto the first $c-1$ coordinates, i.e. $P(\boldsymbol{w}) = P(w_1,\ w_2,\ ...,\ w_c) = (w_1,\ w_2,\ ...,\ w_{c-1})$. For any $\boldsymbol{w} \in W$ and $X \subset W$, we will denote their images $P(\boldsymbol{w})$ and $P(X)$ under P by $\underline{\boldsymbol{w}}$ and $\underline{X}$ respectively. Set $\underline{W} = \left\{(w_1,\ w_2,\ ...,\ w_{c-1}) \in (0,1)^{c-1} \ \middle|\ \sum_{i=1}^{c-1} w_i < 1\right\}$ is an open subset of $\mathbb{R}^{c-1}$. For any point $\underline{\boldsymbol{w}}$ in $\underline{W}$, w_c can be reconstructed using the equation above, so P is a bijection. Identity $\boldsymbol{w} = (\underline{\boldsymbol{w}}, w_c)$ relates the elements of these two sets and whenever symbols $\boldsymbol{w}$ and $\underline{\boldsymbol{w}}$ appear in the same context, they are always bound by this identity. P is actually a homeomorphism between $\underline{W}$ and W which makes W a $(c-1)$–dimensional manifold.

3 Sets mapped onto the same point of concept space

3.1 Preimage of a point of concept space

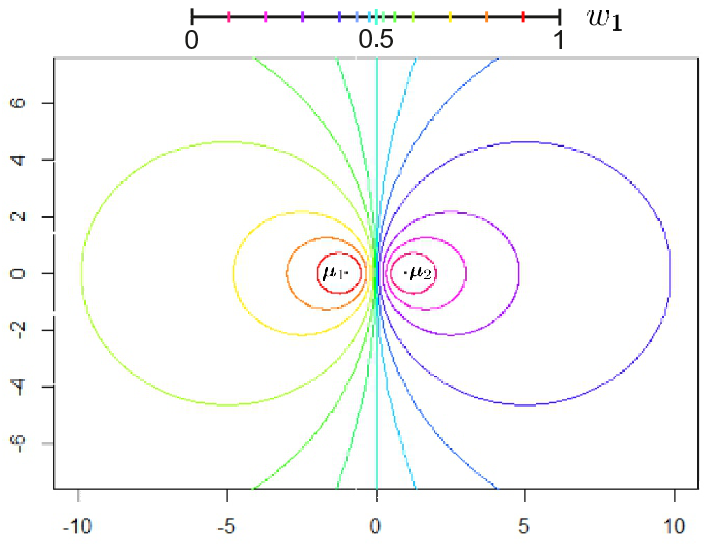

Fig. 2. Preimages of points $(w_1, 1-w_1)$ for $h = 2$ and $c = 2$.

Preimages of points of W under $\mathcal{M}$, when they are nonempty, are spheres of some dimension embedded in $\mathbb{R}^h$, except for a special case when the preimage is

a linear manifold. We define $\mathbb{R}^0$ to be a space containing only a zero 1 x 1 vector and $\mathcal{B}^0(0,R)$ and $\mathcal{S}^{-1}(0,0)$ to be equal to $\mathbb{R}^0$.

For a set containing at least two elements, $\mathcal{C} = \{\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2,\ \ldots\ ,\ \boldsymbol{\mu}_c\} \subset \mathbb{R}^h$, a point in W, $\boldsymbol{w} = (w_1,\ w_2,\ \ldots,\ w_c)$ and $m > 1$, we define the following values

1. M is a $(c-1) \times h$ matrix whose i–th row is $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^T$
2. $\boldsymbol{u}$ is a column vector of length $c-1$ such that $u_i = \frac{1}{2}||\boldsymbol{\mu}_i - \boldsymbol{\mu}_c||^2$
3. $\boldsymbol{v}$ is a column vector of length $c-1$ such that $v_i = -\frac{1}{2}(w_i^{1-m} - w_c^{1-m})$
4. $\boldsymbol{M}^+$ is the Moore–Penrose pseudoinverse of $\boldsymbol{M}$
5. $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{M}\boldsymbol{M}^+$
6. $\boldsymbol{a} = \boldsymbol{M}^+\boldsymbol{u}, \quad \boldsymbol{b} = \boldsymbol{M}^+\boldsymbol{v}$
7. $r = \text{rank}\boldsymbol{M}$, $d = h - r$
8. If $d > 0$, $\boldsymbol{U}$ is a $h \times d$ matrix whose columns form an orthonormal basis of the linear subspace $\{(\boldsymbol{I} - \boldsymbol{M}^+\boldsymbol{M})\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^h\}$ of $\mathbb{R}^h$ of dimension d. If $d = 0$, $\boldsymbol{U}$ is a zero vector of $\mathbb{R}^h$
9. If $\boldsymbol{Q}\boldsymbol{u} = 0$, matrix $\boldsymbol{H}$ is $\left[\ \boldsymbol{U}\ \middle|\ \frac{\boldsymbol{b}}{||\boldsymbol{b}||}\right]$ for $d > 0$ and $\frac{\boldsymbol{b}}{||\boldsymbol{b}||}$ for $d = 0$. If $\boldsymbol{Q}\boldsymbol{u} \neq 0$, it is $\boldsymbol{U}$.

Matrix $\boldsymbol{U}$ can be obtained by selecting a maximal set of linearly independent columns of matrix $\boldsymbol{I} - \boldsymbol{M}^+\boldsymbol{M}$, orthogonalizing it, normalizing each vector and taking the resulting set of vectors as columns of $\boldsymbol{U}$. Values of $\boldsymbol{M}$, $\boldsymbol{u}$, $\boldsymbol{a}$, $\boldsymbol{Q}$, d and $\boldsymbol{U}$ depend only on $\mathcal{C}$, while values of $\boldsymbol{v}$, $\boldsymbol{b}$ and $\boldsymbol{H}$ depend on both $\mathcal{C}$ and $\boldsymbol{w}$. We will treat $\boldsymbol{v}$, $\boldsymbol{b}$ and $\boldsymbol{H}$ as functions of variable $\underline{\boldsymbol{w}} = (w_1,\ w_2,\ \ldots,\ w_{c-1}) \in \underline{W}$. These functions are continuous.

We need above definitions to provide a description of the preimage $\mathcal{M}^{-1}(\boldsymbol{w})$ of $\boldsymbol{w} \in W$, which is split into two cases depending on if the set of centroids lies on a sphere or not. In each case if the conditions are not met, then the preimage is empty (no point of $\mathbb{R}^h$ is mapped onto $\boldsymbol{w}$). We define $\boldsymbol{w}_{eq} = (\frac{1}{c},\ \frac{1}{c},\ \ldots,\ \frac{1}{c})$. Symbol $\mathcal{S}^n$ stands for the unit n–sphere $\mathcal{S}^n = \left\{\boldsymbol{y} \in \mathbb{R}^{n+1} \ \middle|\ ||\boldsymbol{y}|| = 1\right\}$ which is a subset of $\mathbb{R}^{n+1}$.

Set of centroids lies on a sphere (equivalent to $\boldsymbol{Q}\boldsymbol{u} = \boldsymbol{0}$)

If $\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{0}$ and $w_c^{1-m} \geq 2(\boldsymbol{a} \cdot \boldsymbol{b} + ||\boldsymbol{a}||||\boldsymbol{b}||)$, then the preimage is nonempty. For $\boldsymbol{w} \neq \boldsymbol{w}_{eq}$, it is a d–sphere of radius R embedded in $\mathbb{R}^h$ centered at point $\boldsymbol{p}$

$$\mathcal{M}^{-1}(\boldsymbol{w}) = \left\{\boldsymbol{p} + R\boldsymbol{H}\boldsymbol{y} \ \middle|\ \boldsymbol{y} \in \mathcal{S}^d\right\}$$

with

$$R = \frac{\sqrt{(2\boldsymbol{a} \cdot \boldsymbol{b} - w_c^{1-m})^2 - 4||\boldsymbol{a}||^2||\boldsymbol{b}||^2}}{2||\boldsymbol{b}||}$$

$$\boldsymbol{p} = \boldsymbol{\mu}_c + \boldsymbol{a} + \frac{w_c^{1-m} - 2\boldsymbol{a} \cdot \boldsymbol{b}}{2||\boldsymbol{b}||^2}\boldsymbol{b}$$

For $\boldsymbol{w} = \boldsymbol{w}_{eq}$ the preimage is a linear manifold of dimension d

$$\mathcal{M}^{-1}(\boldsymbol{w}_{eq}) = \left\{ \boldsymbol{\mu}_c + \boldsymbol{a} + \boldsymbol{U}\boldsymbol{y} \;\middle|\; \boldsymbol{y} \in \mathbb{R}^d \right\}$$

Set of centroids does not lie on a sphere (equivalent to $\boldsymbol{Q}\boldsymbol{u} \neq \boldsymbol{0}$)

If $\boldsymbol{Q}\boldsymbol{v} \neq \boldsymbol{0}$, $\boldsymbol{Q}\boldsymbol{u} \cdot \boldsymbol{Q}\boldsymbol{v} = -||\boldsymbol{Q}\boldsymbol{u}||||\boldsymbol{Q}\boldsymbol{v}||$ and

$$||\boldsymbol{b}||^2||\boldsymbol{Q}\boldsymbol{u}||^2 + (2\boldsymbol{a}\cdot\boldsymbol{b} - w_c^{1-m})||\boldsymbol{Q}\boldsymbol{u}||||\boldsymbol{Q}\boldsymbol{v}|| + ||\boldsymbol{a}||^2||\boldsymbol{Q}\boldsymbol{v}||^2 \leq 0$$

$$\left(||\boldsymbol{b}||^2||\boldsymbol{Q}\boldsymbol{u}||^2 + (2\boldsymbol{a}\cdot\boldsymbol{b} - w_c^{1-m})||\boldsymbol{Q}\boldsymbol{u}||||\boldsymbol{Q}\boldsymbol{v}|| + ||\boldsymbol{a}||^2||\boldsymbol{Q}\boldsymbol{v}||^2 = 0, \text{ when } d = 0\right)$$

then the preimage is nonempty. It is a $(d-1)$–sphere of radius R embedded in $\mathbb{R}^h$ centered at point $\boldsymbol{p}$. If $d > 0$

$$\mathcal{M}^{-1}(\boldsymbol{w}) = \left\{ \boldsymbol{p} + R\boldsymbol{U}\boldsymbol{y} \;\middle|\; \boldsymbol{y} \in \mathcal{S}^{d-1} \right\}$$

and if $d = 0$

$$\mathcal{M}^{-1}(\boldsymbol{w}) = \{\boldsymbol{p}\}$$

with

$$R = \sqrt{-||\boldsymbol{b}||^2\left(\frac{||\boldsymbol{Q}\boldsymbol{u}||}{||\boldsymbol{Q}\boldsymbol{v}||}\right)^2 - (2\boldsymbol{a}\cdot\boldsymbol{b} - w_c^{1-m})\frac{||\boldsymbol{Q}\boldsymbol{u}||}{||\boldsymbol{Q}\boldsymbol{v}||} - ||\boldsymbol{a}||^2}$$

$$\boldsymbol{p} = \boldsymbol{\mu}_c + \boldsymbol{a} + \frac{||\boldsymbol{Q}\boldsymbol{u}||}{||\boldsymbol{Q}\boldsymbol{v}||}\boldsymbol{b}$$

Let n represent a dimension of the spherical preimages for a given $\mathcal{C}$. If preimage of $\boldsymbol{w} \neq \boldsymbol{w}_{eq}$ is nonempty, then it is given by $\mathcal{M}^{-1}(\boldsymbol{w}) = \left\{ \boldsymbol{p} + R\boldsymbol{H}\boldsymbol{y} \;\middle|\; \boldsymbol{y} \in \mathcal{S}^n \right\}$. Columns of $h \times (n+1)$ matrix $\boldsymbol{H}$ form an orthonormal set of vectors which entails that a linear transformation defined by this matrix is distance–preserving. It maps objects from $\mathbb{R}^{n+1}$ into $\mathbb{R}^h$ without distortion, so the formula for a preimage describes embedding of $\mathcal{S}^n$ into $\mathbb{R}^h$, scaling it to have radius R and the translating it so that it is centered at poin $\boldsymbol{p}$.

Functions $\boldsymbol{p}$ and R are defined on set $W\backslash\{\boldsymbol{w}_{eq}\}$. They are continuous which together with continuity of $\boldsymbol{H}$ implies that a slight change in $\underline{\boldsymbol{w}}$ results in only a slight change of position, radius and spatial orientation of sphere $\mathcal{M}^{-1}(\boldsymbol{w})$. If $d > 0$, then the set of centroids lies on some linear manifold of dimension smaller than h. Mapping $\mathcal{M}$ is not invertible if and only if $\boldsymbol{Q}\boldsymbol{u} = \boldsymbol{0}$ or $d > 0$, in other words if and only if the set of centroids lies on a sphere or on a linear manifold of dimension smaller than h. Thus, if there are at least $h+1$ centroids, then $\mathcal{M}$ is typically invertible (unless they are unfortunately positioned) and we can map points from the concept space back into the feature space precisely. On the other hand, if there are at most h centroids, then they always lie on some sphere, so $\mathcal{M}$ is not invertible. If $\mathcal{M}$ is invertible, then $\mathcal{M}^{-1}(\boldsymbol{w}) = \boldsymbol{p}(\underline{\boldsymbol{w}})$.

3.2 Derivatives of $\boldsymbol{b}$, $\boldsymbol{p}$ and $\boldsymbol{R}$.

Functions $\boldsymbol{b}$ and $\boldsymbol{p}$ are continuously differentiable everywhere on $\overline{W \backslash \boldsymbol{w}_{eq}}$ and R on its subset where it is nonzero. We introduce symbols for three $(c-1) \times 1$ vectors $\mathbf{1} = \big[1,\ 1,\ \ldots,\ 1\big]^T$, $\boldsymbol{v}_I = (1-m)\big[w_1^{-m},\ w_2^{-m},\ \ldots,\ w_{c-1}^{-m}\big]^T$ and $\boldsymbol{v}_c = (1-m)\big[w_c^{-m},\ w_c^{-m},\ \ldots,\ w_c^{-m}\big]^T$. We let $\mathrm{diag}(\boldsymbol{x})$ denote a diagonal matrix with coordinates of vector $\boldsymbol{x}$ on its diagonal and we put $s = \dfrac{R^2 + ||\boldsymbol{a}||^2}{||\boldsymbol{b}||^2}$. We can now write the differentials

$$D\boldsymbol{v} = -\tfrac{1}{2}\big[\boldsymbol{v}_c \mathbf{1}^T + \mathrm{diag}(\boldsymbol{v}_I)\big]$$

$$D\boldsymbol{b} = \boldsymbol{M}^+ D\boldsymbol{v}$$

In case of $\boldsymbol{Qu} = \mathbf{0}$ we have

$$\nabla R = \frac{1}{R}\Big[D\boldsymbol{b}^T\big(\sqrt{s}\boldsymbol{a} - s\boldsymbol{b}\big) + \frac{\sqrt{s}}{2}\boldsymbol{v}_c\Big]$$

$$D\boldsymbol{p} = \frac{1}{||\boldsymbol{b}||^2}\boldsymbol{b}\Big[D\boldsymbol{b}^T\big(2\sqrt{s}\boldsymbol{b} - \boldsymbol{a}\big) - \frac{1}{2}\boldsymbol{v}_c\Big]^T - \sqrt{s}D\boldsymbol{b}$$

In case of $\boldsymbol{Qu} \neq \mathbf{0}$ we define $t_0 = \dfrac{||\boldsymbol{Qu}||}{||\boldsymbol{Qv}||}$ and we have

$$\nabla R = \frac{t_0}{2R}\left[\frac{1}{||\boldsymbol{Qv}||^2}\big(2t_0||\boldsymbol{b}||^2 + 2\boldsymbol{a}\cdot\boldsymbol{b} - w_c^{1-m}\big)D\boldsymbol{v}^T\boldsymbol{Q}^T\boldsymbol{Qv} - 2D\boldsymbol{b}^T\big(\boldsymbol{a} + t_0\boldsymbol{b}\big) - \boldsymbol{v}_c\right]$$

$$D\boldsymbol{p} = t_0\boldsymbol{M}^+\Big[\boldsymbol{I} - \frac{1}{||\boldsymbol{Qv}||^2}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{Q}^T\boldsymbol{Q}\Big]D\boldsymbol{v}$$

3.3 Reduced set of centroids

In this section, we will specify a $\widehat{c}$ element subset $\widehat{\mathcal{C}}$ of the set of centroids $\mathcal{C}$, such that for every $\boldsymbol{w} \in W$, preimage $\mathcal{M}^{-1}(\boldsymbol{w})$ under $\mathcal{M}$ is the same as preimage of some $\widehat{\boldsymbol{w}} \in \widehat{W}$ under mapping $\mathcal{M}_{\widehat{\mathcal{C}}} : \mathbb{R}^h \longrightarrow [0,1]^{\widehat{c}}$, which corresponds to the set of centroids $\widehat{\mathcal{C}}$. We will use the hat symbol to indicate that a given object corresponds to this reduced set of centroids, e.g. $\widehat{\boldsymbol{b}}$ is $\boldsymbol{b}$ computed for the set of centroids $\widehat{\mathcal{C}}$.

The i–th row of $\boldsymbol{M}$ is of the form $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^T$, so let's say that i–th row of $\boldsymbol{M}$ and centroid $\boldsymbol{\mu}_i$ correspond to each other. Let S be a set containing $\boldsymbol{\mu}_c$ and all the centroids that correspond to some set of r linearly independent rows of $\boldsymbol{M}$. If $\boldsymbol{Qu} = \mathbf{0}$, then we set $\widehat{\mathcal{C}} = S$. This set lies on a sphere. If $\boldsymbol{Qu} \neq \mathbf{0}$, we take another centroid $\boldsymbol{\mu}_k$ such that $S \cup \{\boldsymbol{\mu}_k\}$ does not lie on a sphere and we set $\widehat{\mathcal{C}} = S \cup \{\boldsymbol{\mu}_k\}$. We can find $\boldsymbol{\mu}_k$ by checking for each centroid if amending S with it results in $\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{u}} \neq \mathbf{0}$.

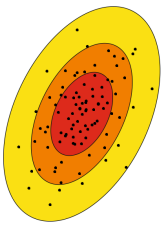

Fig. 3. Contour plot of a simple probability density derived from a data set in a two–dimensional feature space.

Let $k_1,\ k_2,\ \ldots,\ k_{\widehat{c}}$ be the indices of centroids that we put into $\widehat{\mathcal{C}}$. We define transformation $T : W \longrightarrow \widehat{W}$ with formula

$$T(\boldsymbol{w}) = \frac{1}{\sum_{i=1}^{\widehat{c}} w_{k_i}} \big(w_{k_1},\ w_{k_2},\ \ldots,\ w_{k_{\widehat{c}}}\big)$$

For any $\boldsymbol{w} \in W$, $\mathcal{M}^{-1}(\boldsymbol{w}) = \mathcal{M}_{\widehat{\mathcal{C}}}^{-1}\big(T(\boldsymbol{w})\big)$. T is invertible and continuous on W. We let $\widehat{\boldsymbol{w}}$ denote $T(\boldsymbol{w})$ and we let $\underline{\widehat{\boldsymbol{w}}}$ denote $\widehat{P}(T(\boldsymbol{w}))$. Elements of sets $\widehat{W}$ and $\underline{\widehat{W}}$ are in one-to-one correspondence with elements of set W (and with each other), so we will use symbols $\widehat{\boldsymbol{w}}$ and $\underline{\widehat{\boldsymbol{w}}}$ respectively to denote elements of these sets. Thus we have for $\boldsymbol{w} \in W$, $\mathcal{M}^{-1}(\boldsymbol{w}) = \mathcal{M}_{\widehat{\mathcal{C}}}^{-1}\big(\widehat{\boldsymbol{w}}\big)$. For every $\boldsymbol{w} \in W$, we have $\widehat{R}(\underline{\widehat{\boldsymbol{w}}}) = R(\underline{\boldsymbol{w}})$, $\widehat{\boldsymbol{p}}(\underline{\widehat{\boldsymbol{w}}}) = \boldsymbol{p}(\underline{\boldsymbol{w}})$, $\widehat{\boldsymbol{U}} = \boldsymbol{U}$, $\widehat{\boldsymbol{b}}(\underline{\widehat{\boldsymbol{w}}}) = \boldsymbol{b}(\underline{\boldsymbol{w}})$ and $\widehat{\boldsymbol{H}}(\underline{\widehat{\boldsymbol{w}}}) = \boldsymbol{H}(\underline{\boldsymbol{w}})$. Though unless $\widehat{\mathcal{C}} = \mathcal{C}$, $D\widehat{\boldsymbol{b}}(\underline{\widehat{\boldsymbol{w}}})$ is not equal to $D\boldsymbol{b}(\underline{\boldsymbol{w}})$, because these matrices have different number of columns. The same is true for $D\widehat{\boldsymbol{p}}$ and $\nabla\widehat{R}$.

4 Reverse mapping with minimal MSE

Let's assume that there is a continuous probability distribution, with a continuous probability density function f, defined on the feature space $\mathbb{R}^h$. Function f can be interpreted as density of points of some data set $\mathcal{D} \subset \mathbb{R}^h$. For a real data set, such density can be approximated using statistical density estimation methods. Figure 2 shows an example of a simple probability density function derived from a data set. In this chapter, we will look for a reverse mapping G from the concept space to the feature space which allows minimum mean squared error (MSE), with respect to f, of an operation of mapping points of feature space to concept space via $\mathcal{M}$ and then back to feature space via G, cf. figure 3. Let $\boldsymbol{X}$ be a random vector distributed according to density f. MSE of this operation is given by expression $\mathbb{E}\big|\big|\boldsymbol{X} - G(\mathcal{M}(\boldsymbol{X}))\big|\big|^2$, so the function that minimizes it is equal to conditional expectation $G(\boldsymbol{w}) = \mathbb{E}\big(\boldsymbol{X}\big|\mathcal{M}(\boldsymbol{X}) = \boldsymbol{w}\big)$ almost everywhere.

4.1 Parameterization of feature space

For $n > 0$, we parametrize almost all of the unit sphere $\mathcal{S}^n \subset \mathbb{R}^{n+1}$ by n angular coordinates $\boldsymbol{\theta} = (\theta_1,\ \theta_2,\ ...,\ \theta_n)$ that range over set $\Theta = (0, \pi)^{n-1} \times (0, 2\pi) \subset \mathbb{R}^n$.

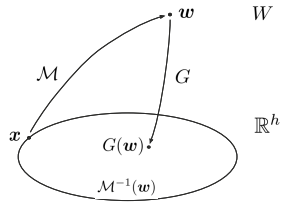

Fig. 4. Point $\boldsymbol{x}$ of $\mathbb{R}^h$ is mapped by $\mathcal{M}$ onto a point $\boldsymbol{w}$ of W which in turn is mapped back into $\mathbb{R}^h$ by G.

$$\boldsymbol{y}(\boldsymbol{\theta}) = \begin{bmatrix} \cos(\theta_1) \\ \sin(\theta_1)\cos(\theta_2) \\ \sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ \vdots \\ \sin(\theta_1)\cdots\sin(\theta_{n-1})\cos(\theta_n) \\ \sin(\theta_1)\cdots\sin(\theta_{n-1})\sin(\theta_n) \end{bmatrix} \quad \boldsymbol{y}^{-1}(\boldsymbol{x}) = \begin{bmatrix} \arccos \frac{x_1}{\sqrt{x_{n+1}^2+x_n^2+\cdots+x_1^2}} \\ \arccos \frac{x_2}{\sqrt{x_{n+1}^2+x_n^2+\cdots+x_2^2}} \\ \vdots \\ \arccos \frac{x_{n-1}}{\sqrt{x_{n+1}^2+x_n^2+x_{n-1}^2}} \\ 2\mathrm{arccot}\frac{x_n+\sqrt{x_{n+1}^2+x_n^2}}{x_{n+1}} \end{bmatrix}$$

Vector $\boldsymbol{y}$ has length $n+1$ and $\boldsymbol{y}^{-1}$ length n. The (i,j)–th element of $(n+1)\times n$ matrix $D\boldsymbol{y}$ is

$$\frac{\partial y_i}{\partial \theta_j} = \begin{cases} 0 & \text{if } \ i < j \\ -\sin(\theta_1)\ldots\sin(\theta_{i-1})\sin(\theta_i) & \text{if } \ i = j \\ \sin(\theta_1)\ldots\sin(\theta_{j-1})\cos(\theta_j)\sin(\theta_{j+1})\ldots\sin(\theta_{i-1})\cos(\theta_i) & \text{if } \ j < i \le n \\ \sin(\theta_1)\ldots\sin(\theta_{j-1})\cos(\theta_j)\sin(\theta_{j+1})\ldots\sin(\theta_n) & \text{if } \ i = n+1 \end{cases}$$

For $n = 0$, we set $\Theta = \{-1, 1\}$ and we parametrize $\mathcal{S}^0$ with an identity $y : \Theta \longrightarrow \{-1, 1\}$. Let $\mathcal{I}_s$ be a subset of W such that preimages of its points are nondegenerate spheres. We define transformation $\psi : \underline{\widehat{\mathcal{I}_s}} \times \Theta \longrightarrow \mathbb{R}^h$ as follows

$$\psi(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) = \psi_{\boldsymbol{w}}(\boldsymbol{\theta}) = \boldsymbol{p}(\underline{\boldsymbol{w}}) + R(\underline{\boldsymbol{w}})\boldsymbol{H}(\underline{\boldsymbol{w}})\boldsymbol{y}(\boldsymbol{\theta})$$

For a fixed $\underline{\widehat{\boldsymbol{w}}}$, it is a bijection between Θ and $\mathcal{M}^{-1}(\boldsymbol{w})$. Mapping ψ is a bijection onto $\mathcal{M}^{-1}(\mathcal{I}_s)$ whose complement in $\mathbb{R}^h$ is of measure zero if $\mathcal{M}$ is not invertible. ψ is continuously differentiable. (if $n = 0$, it is continuously differentiable for a fixed $\boldsymbol{\theta}$). If $\boldsymbol{Q}\boldsymbol{u} = 0$, then

$$D_{\underline{\widehat{\boldsymbol{w}}}}\psi = D\widehat{\boldsymbol{p}} + \boldsymbol{H}\boldsymbol{y}\nabla\widehat{R}^T + \frac{Ry_{n+1}}{||\boldsymbol{b}||}\big(\boldsymbol{I} - \frac{1}{||\boldsymbol{b}||^2}\boldsymbol{b}\boldsymbol{b}^T\big)D\widehat{\boldsymbol{b}}$$

and if $\boldsymbol{Q}\boldsymbol{u} \neq 0$, then

$$D_{\underline{\widehat{\boldsymbol{w}}}}\psi = D\widehat{\boldsymbol{p}} + \boldsymbol{U}\boldsymbol{y}\nabla\widehat{R}^T$$

If $n > 0$, then also

$$D_{\boldsymbol{\theta}}\psi = R\boldsymbol{H}D\boldsymbol{y}$$

For $n > 0$, we define $\mathcal{J}(\underline{\boldsymbol{w}}, \boldsymbol{\theta})$ as the Jacobian of ψ, $\mathcal{J}(\underline{\boldsymbol{w}}, \boldsymbol{\theta}) = \big|\det[D_{\underline{\boldsymbol{w}}}\psi \mid D_{\boldsymbol{\theta}}\psi]\big|$ and for $n = 0$, as $\mathcal{J}(\underline{\boldsymbol{w}}, \boldsymbol{\theta}) = \big|\det D_{\underline{\boldsymbol{w}}}\psi\big|$.

4.2 Reverse mapping

We define a subset of W, $\mathcal{I}_f = \{\boldsymbol{w} \in W \mid \exists \boldsymbol{x} \in \mathcal{M}^{-1}(\boldsymbol{w}) \quad f(\boldsymbol{x}) > 0\}$. The following mapping $G : \mathcal{I}_f \longrightarrow \mathbb{R}^h$, allows the minimum MSE of the operation we discussed above. For $\boldsymbol{w} \in \mathcal{I}_f \cap \mathcal{I}_s$,

$$G(\boldsymbol{w}) = \frac{\int\limits_{\Theta} \psi_{\boldsymbol{w}}(\boldsymbol{\theta}) f(\psi_{\boldsymbol{w}}(\boldsymbol{\theta})) \mathcal{J}(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) \mathrm{d}\lambda^n}{\int\limits_{\Theta} f(\psi_{\boldsymbol{w}}(\boldsymbol{\theta})) \mathcal{J}(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) \mathrm{d}\lambda^n}$$

and for all other $\boldsymbol{w} \neq \boldsymbol{w}_{eq}$ in its domain $G(\boldsymbol{w}) = \boldsymbol{p}(\underline{\boldsymbol{w}})$. Symbol λ^n stands for Lebesgue measure on $\mathbb{R}^n$. Value of an integral of a vector–valued function is a vector of integrals of its component functions. For $\boldsymbol{w} \in \mathcal{I}_f \cap \mathcal{I}_s$, we can rewrite the above formula as

$$G(\boldsymbol{w}) = \boldsymbol{p}(\underline{\boldsymbol{w}}) + \Big(\int\limits_{\Theta} f(\psi_{\boldsymbol{w}}(\boldsymbol{\theta})) \mathcal{J}(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) \mathrm{d}\lambda^n \Big)^{-1} R(\underline{\boldsymbol{w}}) \boldsymbol{H}(\underline{\boldsymbol{w}}) \int\limits_{\Theta} \boldsymbol{y}(\boldsymbol{\theta}) f(\psi_{\boldsymbol{w}}(\boldsymbol{\theta})) \mathcal{J}(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) \mathrm{d}\lambda^n$$

If $\boldsymbol{w}_{eq} \in \mathcal{I}_f$, then $G(\boldsymbol{w}_{eq})$ can be arbitrary as this single value has no bearing on the value of MSE. G is continuous on $\mathcal{I}_f \backslash \{\boldsymbol{w}_{eq}\}$.

For $\boldsymbol{w} \in \mathcal{I}_f \cap \mathcal{I}_s$, $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{x} \in \mathcal{M}^{-1}(\mathcal{I}_s)$, we define

$$\phi(\boldsymbol{w}, \boldsymbol{\theta}) = \sqrt{\det \left(D_{\underline{\widehat{\boldsymbol{w}}}} \psi(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta})^T D_{\underline{\widehat{\boldsymbol{w}}}} \psi(\underline{\widehat{\boldsymbol{w}}}, \boldsymbol{\theta}) \right)}$$

$$\boldsymbol{\Phi}_{\boldsymbol{w}}(\boldsymbol{x}) = \phi\big(\boldsymbol{w}, \psi_{\boldsymbol{w}}^{-1}(\boldsymbol{x})\big) = \phi\Big(\boldsymbol{w}, \boldsymbol{y}^{-1}\Big(\tfrac{1}{R}\boldsymbol{U}^T\big(\boldsymbol{x} - \boldsymbol{p}\big)\Big)\Big)$$

Integrals below are over the embedded n–sphere that is the preimage of $\boldsymbol{w}$ and is an n–dimensional manifold. If $\boldsymbol{Q}\boldsymbol{u} \neq \boldsymbol{0}$ and $n > 0$, then for $\boldsymbol{w} \in \mathcal{I}_f \cap \mathcal{I}_s$,

$$G(\boldsymbol{w}) = \frac{\int\limits_{\mathcal{M}^{-1}(\boldsymbol{w})} \boldsymbol{x} f(\boldsymbol{x}) \boldsymbol{\Phi}_{\boldsymbol{w}}(\boldsymbol{x}) \mathrm{d}S^n}{\int\limits_{\mathcal{M}^{-1}(\boldsymbol{w})} f(\boldsymbol{x}) \boldsymbol{\Phi}_{\boldsymbol{w}}(\boldsymbol{x}) \mathrm{d}S^n}$$

This alternative form enables approximation of this value by summing over sets of points spaced regularly and finely on the n–sphere $\mathcal{M}^{-1}(\boldsymbol{w})$ embedded in $\mathbb{R}^h$.

4.3 Algorithm

Based on this work we may consider the following algorithm for mapping a point from the concept space to the feature space. We are given set $\mathcal{C} \subset \mathbb{R}^h$ of centroids of c clusters and a probability density function f defined on $\mathbb{R}^h$ or a data set $\mathcal{D} \subset \mathbb{R}^h$. If we are given a data set, we will find values of f with

density estimation methods. We compute $\boldsymbol{M}$, $\boldsymbol{M}^+$, $\boldsymbol{Q}$, $\boldsymbol{u}$, $\boldsymbol{a}$, r, d and $\boldsymbol{U}$. Based on $\boldsymbol{Qu}$ we ascertain which case we are dealing with and we compute $\boldsymbol{H}$ and n. We determine a reduced set of clusters $\widehat{\mathcal{C}}$.

The input of the algorithm is $\boldsymbol{w} \in [0,1]^c$ and the output is $\boldsymbol{x} \in \mathbb{R}^h$. When value of $\boldsymbol{x}$ is set, the algorithm stops. Let $\boldsymbol{\mu}$ be a centroid of the cluster in which $\boldsymbol{w}$ has the highest membership.

1. Project $\boldsymbol{w}$ onto W orthogonally/in the direction of the origin.
2. Compute $\boldsymbol{v}$, $\boldsymbol{b}$ and $\boldsymbol{Qv}$.
3. If conditions for $\mathcal{M}^{-1}(\boldsymbol{w})$ to be nonempty are not met, set $\boldsymbol{x} = \boldsymbol{\mu}$.
4. If $\boldsymbol{w} \neq \boldsymbol{w}_{eq}$, compute $\boldsymbol{p}$ and R.
5. If $d = 0$ and $\boldsymbol{Qu} \neq \boldsymbol{0}$ or if $R = 0$, set $\boldsymbol{x} = \boldsymbol{p}$.
6. Check if f is positive on any point of $\mathcal{M}^{-1}(\boldsymbol{w})$. If not, set $\boldsymbol{x} = \boldsymbol{p}$.
7. Compute an approximation of $G(\boldsymbol{w})$ and set $\boldsymbol{x}$ to its value.

We tested above algorithm against assigning to $\boldsymbol{w} \in W$ a mean of $\mathcal{M}^{-1}(\boldsymbol{w})$ (which in case of FCM is $\boldsymbol{p}$). Each data set consisted of c clouds of points generated using multivariate normal probability distributions of different means and covariance matrices. Results are in table 1.

Table 1. Comparison of MSE

h	c	$\boldsymbol{p}$	$G(\boldsymbol{w})$
3	3	22.58	15.35
4	4	25.37	19.19
7	7	8.21	6.75
4	3	48.65	25.95
5	4	23.72	20.22
6	5	25.71	23.23

4.4 Application

The reverse mapping from the concept space to the feature space that we derived, is best fitted – in terms of MSE – to a given probability density function defined on the feature space which can be a density of some data set in the feature space. We will discuss an application of this mapping to the problem of defuzzification of results of computation on fuzzy data. Suppose we have a training set T whose each element consists of k input points $\boldsymbol{x}_1^i,\ \boldsymbol{x}_2^i,\ \ldots,\ \boldsymbol{x}_k^i$ of feature space $\mathbb{R}^h$ and one output point $\boldsymbol{x}_0^i$ of feature space. All the feature space points from T put together, constitute a data set $\mathcal{D}$ on which we use FCM to find c fuzzy clusters and a set of their centroids $\mathcal{C}$. We transform every $\boldsymbol{x}_j^i$ into $\mathcal{M}(\boldsymbol{x}_j^i)$, to obtain a concept-level training set T_c. Using some machine learning method, we train on T_c an operation F_c, working on the level of concept space, that returns a point of concept space based on k input points of concept space. Let G be a minimum MSE reverse mapping associated with $\mathcal{C}$, found for f – a density of

data set $\mathcal{D}_{\text{out}} \subset \mathcal{D}$ consisting only of output points. G has minimum MSE when defuzzifying output points of the training set T_{c}, so we may expect it to defuzzify other outputs well.

Furthermore, for all sets of input points from T_{c}, F_{c} is fitted to return a point close to a correct output point $\mathcal{M}(\boldsymbol{x}_0^i)$ and G is fitted to map to a point of feature space close to $\boldsymbol{x}_0^i$. Thus, the composition $F = G \circ F_{\text{c}} \circ \mathcal{M}$ may be a good fit for training set T as an operation that returns a point of feature space based on k input points of feature space. This approach may be useful in case of data which is easier modeled at the concept level than at the numeric level.

5 Conclusion

We presented a reverse mapping from concept space to feature space, best fitted – in terms of MSE – to a given probability density function defined on feature space or a given data set in feature space. Similar reverse mappings could be obtained for different transformations into the concept space. Maybe transformation that allows more precise reverse mapping could be constructed. Many aspects of this work leave room for further research. In the described application the concept-level operation may return points that do not lie in set W. Effectiveness of various approaches to amend this situation can be investigated, such as orthogonal projection onto W or projection onto W in the direction of the origin. Also there may be better ways to treat the points of W that drop out at steps 3 and 6 of the algorithm. For a point of W that has a nonempty preimage, but density f is zero on all its points, it might be a good treatment to map it onto the point of its preimage that is closest to the mean of $\mathcal{D}$ weighted by f. The second important direction of research is approximation of f in a situation when we do not have access to any data set in the feature space.

References

1. J. C. Bezdek, R. Ehrlich, W. Full, FCM: the Fuzzy c-Means Clustering Algorithm, Computer and Geosciences, 10 (2-3), 191-203, 1984
2. E.V. Broekhoven, B.D. Baets, Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trapezoidal fuzzy partitions, Fuzzy Sets and Systems, 157(3), 904-918, 2006
3. R. N. Dave, K. Bhaswan, Adaptive Fuzzy c-shells Clustering and Detection of Ellipses, IEEE Transactions on Neural Networks 3(5), 643-662, 1992
4. W. van Leekwijck, E. E. Kerre, "Defuzzification: criteria and classification". Fuzzy Sets and Systems 108 (2), 159-178, 1999
5. X. Liu, Parameterized defuzzification with maximum entropy weighting function - Another view of the weighting function expectation method, Mathematical and Computer Modelling, 45(1-2), 177-188, 2007
6. D. P. Madau, L. A. Feldkamp, "Influence value defuzzification method". Fuzzy Systems 3, 1819-1824, 1996
7. E. Roventa, T. Spircu, Averaging procedures in defuzzification processes, Fuzzy Sets and Systems, 136(3), 375-385, 2003