Optimal Choices of Potential Functions in Fuzzy Clustering

Vladik Kreinovich[1], Hung T. Nguyen[2], and Yeung Yam[3]

[1]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
email vladik@cs.utep.edu

[2]Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA
email hunguyen@nmsu.edu

[3]Department of Mechanical & Automation Engineering
The Chinese University of Hong Kong
Shatin, NT
Hong Kong, China
email yyam@mae.cuhk.edu.hk

Abstract

Fuzzy logic-based clustering techniques are widely used in situations where statistical assumptions are not valid. Whether in estimating cluster centers for model identification purposes or in determining clusters the existing techniques are essentially based upon the choice of some potential functions. As in any design problems of this kind, the choice of such a function has to be justified on a theoretical basis. In this work, we set up a decision frame work and show that optimal potential functions are the ones which are used in current techniques.

1 Fuzzy Clustering: Existing Approaches and Formulation of the Problem

Clustering is important. Analysis of every new phenomenon starts with *classification*, when instead of *numerous* different *examples*, we have a *few classes*. Classification helped to analyze chemical elements, elementary particles, living organisms, astronomical objects, etc.

Statistical clustering is not always possible, so, alternative (in particular, fuzzy) methods are needed. In some situations, where assumptions about structure of data can be formulated in statistical terms, statistical techniques (see, e.g., [21]) are appropriate if we have sufficiently many data.

In other situations, we must use heuristic classification methods, in particular, methods that use fuzzy logic.

The main idea of fuzzy clustering is described in [3, 4, 5, 6, 7, 8, 11, 12, 13, 27, 49, 50].

The goal of fuzzy clustering: "typical" representatives and how to use them. We start with objects which we want to classify (i.e., to cluster). To classify, we use several (numerical) characteristics of these object. Let us denote the total number of these characteristics by s. The s real numbers that characterize each object can be naturally viewed as a point in s-dimensional space R^s. Thus, having n objects means that we have n points $x_1, \ldots, x_n$ in this space. These n points are the input for clustering.

As a result of clustering, we want to describe several clusters. Each cluster can be characterized by its "typical" element $t_j \in R^s$. After these typical elements $t_1, \ldots, t_q$ are found, we can then classify each object $x \in R^s$ according to which typical element it is closest to.

This "classification" is a fuzzy notion:

- if an element x is very close to, say, t_1, and not close to any other typical representative, then it is reasonable to conclude that x belongs to class 1;
- however, if an object $x \in R^s$ is almost equally close to two different representatives t_1 and t_2, then it is reasonable to conclude that this object belongs, to some extent, to *both* clusters 1 and 2.

To express this idea in precise terms, we select a function $f(x)$ (called *potential function*) such that for every two point x and y from R^s, the value $f(x-y)$ describes to what extent x and y are close. This function is usually non-negative, and the closer x and y, the larger the value of the potential function. Potentially, as a potential function, we can use a *membership function* which describes the relation "x and y are close"; however, from the mathematical viewpoint, the choice of membership function would mean that we only allow $f(x)$ to take values from the interval $[0,1]$, and sometimes, more general values are needed (in our main text, we will explain why we need such values).

When the potential function is selected, then we can say that an object x belongs to 1-st cluster with a degree $f(x-t_1)$, to the 2-nd cluster with the degree $f(x-t_2)$, ..., and to q-th cluster with the degree $f(x-t_q)$.

Since we do not require any normalization of the function $f(x)$, it is convenient to *normalize* these values so that they will add up to 1, in other words, to describe the degree to which x belongs to j-th cluster as

$$d_j(x) = \frac{f(x-t_j)}{f(x-t_1)+\ldots+f(x-t_q)}. \tag{1}$$

How to find "typical" representatives? The most widely used approach. We have described how to classify an object when the clusters (or, to be more precise, their typical representatives) have already been found. How can we find these representatives?

The most widely used fuzzy clustering method is the method of Fuzzy C-Means (Fuzzy ISODATA) [3, 4, 5, 6, 7, 8, 13, 27]. This method is based on the natural idea that each characteristic of a typical representative should be equal to an average over all elements of the corresponding cluster. If we have *crisp* clustering, then we would simply take the arithmetic average. However, since we have *fuzzy* clustering, it is natural to count, in this average, each element x_i with the weight $d_j(x_i)$ that is proportional to this element's degree of belonging to the cluster. In other words, it is natural to require that for each j,

$$t_j = \frac{d_j(x_1)\cdot x_1+\ldots+d_j(x_n)\cdot x_n}{d_j(x_1)+\ldots+d_j(x_n)}. \tag{2}$$

This method leads to *good quality* clustering. Its main *disadvantage* is that since the values $d_j(x_i)$, in their turn, depend on t_j, the equation (2) is, actually, a non-linear system of equations for determining the cluster "centers" $t_1,\ldots,t_q$, and solving this system of equations often requires lots of *computation time.*

How to find "typical" representatives? Recent approaches. To simplify computations, a new method has been recently proposed [49, 50] (see also [11, 12]). This method is based on the following idea: when we say that an element t_j is a typical representative of the cluster that consists of elements $x_{i_1},\ldots,x_{i_k}$, we mean that for each element $x\in R^s$, the degree $f(x-t_j)$ with which x is close to t_j is equal to the average of the degrees $f(x-x_{i_1}),\ldots,f(x-x_{i_k})$ with which x is close to all elements of this cluster:

$$f(x-x_{i_1})+\ldots+f(x-x_{i_k}) = k\cdot f(x-t_j). \tag{3}$$

If we have a *crisp* classification, then each of the original data points $x_1,\ldots,x_n$ belongs to one and only one cluster and therefore, by adding equalities (3) for all q clusters, we would conclude that

$$\sum_{i=1}^{n} f(x-x_i) = \sum_{j=1}^{q} k_j\cdot f(x-t_j), \tag{4}$$

where k_j is the total number of elements in j-th cluster (i.e., the *cardinality* of j-th cluster).

For a *fuzzy* clustering, it is reasonable to expect a similar formula, with k_j being the *fuzzy cardinality* of j-th cluster (see, e.g., [28]). So, to find t_j, we can do the following:

- compute, for all x, the function

$$M(x) = \sum_{i=1}^{n} f(x - x_i).$$

- represent this function $M(x)$ as a sum

$$M(x) = \sum_{j=1}^{q} k_j \cdot f(x - t_j)$$

  for the smallest possible number of clusters.

Theoretically, the smallest possible number of clusters is 1, in which case $M(x) = k_1 \cdot f(x - t_1)$. If one cluster is indeed sufficient, then, due to the properties of the "closeness" function $f(x)$, we can find t_1 easily: it is the value for which $M(x)$ is the largest possible. In this case, if $f(x)$ is normalized in such a way that $f(0) = 1$ (i.e., if $f(x)$ is a membership function, and x is close to x with degree of truth 1), we can take $k_1 = M(t_1)$.

In view of this observation, it is reasonable to select, as t_1, the value for which $M(x)$ is the largest possible. In this case, we cannot take $k_1 = M(t_1)$, because other clusters are also contributing to this value $M(t_1)$. Instead, we can take $k_1 = q \cdot M(t_1)$ for some number $q \in (0, 1)$. After that, we can subtract $k_1 \cdot f(x - t_1)$ from the original function $M(x)$, and use a similar method to represent the new function $M_1(x) = M(x) - k_1 \cdot f(x - t_1)$ as a sum

$$M_1(x) = \sum_{j=2}^{q} k_j \cdot f(x - t_j);$$

etc. We stop when the remainder becomes small enough.

This method is very similar to a very successful method of image reconstruction used in radio astronomy under the name of *CLEAN* (see [20, 23, 26, 29, 31, 47] and references therein). Due to the success of the CLEAN method, it is not surprising that this clustering method also turned out to be reasonably successful.

Main problem: how to choose a potential function? We have mentioned that the above fuzzy clustering methods turned out to be very successful, but we must clarify this statement: these methods are very successful provided we appropriately choose the potential function $f(x)$. For a different choice of $f(x)$, the resulting clustering may not be that good.

To the best of our knowledge, so far, the choice of the potential function was mainly done either empirically or heuristically. The following three families of potential functions are most widely used:

- in the original Fuzzy C-Means method, the function $f(x) = |x|^{-m}$ is used, where $|x|$ is the norm of a vector x, and $m > 0$ is a positive real number;
- in [49, 50], the potential function $f(x) = \exp(-\alpha \cdot |x|)$ is used; and
- in [11, 12], the Gaussian potential function $f(x) = \exp(-\alpha \cdot |x|^2)$ is used.

The first choice is used when we have no information about the typical cluster radius; the second and third choices presuppose that an approximate cluster radius is already known.

In this paper, we show that these three choices are indeed optimal in some reasonable sense. Thus, we provide a theoretical justification of these empirical and heuristic choices.

2 Optimal Potential Functions: General Idea

Optimal in what sense? The main idea. We are looking for the *best* (*optimal*) choice of a potential function.

Normally, the word "best" is understood in the sense of some *numerical* optimality criterion. However, in our case of *fuzzy* choice, it is often difficult to formulate the exact *numerical* criterion. Instead, we assume that there is an *ordinal* criterion, i.e., that we can compare arbitrary two choices, but that we cannot assign numerical values to these choices.

It turns out that in many cases, there are reasonable *symmetries*, and it is natural to assume that the (ordinal) optimality criterion is invariant with respect to these symmetries. Then, we are able to describe all choices that are optimal with respect to some invariant ordinal optimality criteria.

This general approach was described and used in [9, 33, 34, 38, 45], in particular, for fuzzy control. In this section, we will show that this approach is applicable to fuzzy clustering as well.

Let us borrow from the experience of modern physics and use symmetries. In modern physics, symmetry groups are a tool that enables to compress complicated differential equations into compact form (see, e.g., [24, 42, 46]). For example:

- Maxwell's equations of electrodynamics consist of *four* different differential equations for two vector fields: electric field $\vec{E}$ and magnetic filed $\vec{B}$.
- However, if we take into consideration that these equations are invariant with respect to Lorentz transformations (that form the basis of Special Relativity) then we can compress these equations into *two*: $F_{ab}^{,b} = j_a$, and $F_{ab,c} + F_{bc,a} + F_{ca,b} = 0$.

Moreover, the very differential equations themselves can be uniquely deduced from the corresponding symmetry requirements [17, 18, 30, 37] (see also [14, 15, 16]).

It is possible to use symmetry. As we have mentioned, in our previous papers, we have shown that the symmetry group approach can be used to find optimal membership functions, optimal t-norms and t-conorms, and optimal defuzzification procedures.

It is therefore reasonable to expect that the same approach can also be used to choose the best potential function for fuzzy clustering.

3 Optimal Potential Functions: Case When We Do Not Have a Prior Knowledge of the Cluster Radius

3.1 Motivations

We must choose a family of functions. We must select a potential function $f(x)$. The only way the potential function $f(x)$ is used in clustering is through the normalized formula (1). Because of the normalization, if we re-scale the values of the potential function, i.e., if we choose a constant $C > 0$ and consider a new potential function $\tilde{f}(x) = C \cdot f(x)$, this new potential function will lead to exactly the same values $d_j(x)$ as the old one. Therefore, from the viewpoint of fuzzy clustering, there is no way to distinguish between the functions $f(x)$ and $\tilde{f}(x) = C \cdot f(x)$. So, based on clustering behavior, we cannot choose a *single* function $f(x)$; we can only choose a 1-parametric *family* of functions $\{C \cdot f(x)\}$ that is characterized by a parameter C.

Comment about notations. In the following text, we will denote families of functions by capital letters, such as F, F', G, etc.

We must choose the best family of functions. We want to select the *best* family of functions.

What is a criterion for choosing a family of functions? What does it mean to choose a *best* family of functions? It means that we have some *criterion* that enables us to choose between the two families.

Traditionally, optimality criteria are *numerical*, i.e., to every family F, we assign some value $J(F)$ expressing its quality, and choose a family for which this value is maximal (i.e., when $J(F) \geq J(G)$ for every other alternative G). However, it is not necessary to restrict ourselves to such numeric criteria only.

For example, if we have several different families F that have the same classification ability $P(F)$, we can choose between them the one that has the minimal computational complexity $C(F)$. In this case, the actual criterion that we use to compare two families is not numeric, but more complicated:

> *A family F_1 is better than the family F_2 if and only if*
>
> – *either $P(F_1) > P(F_2)$,*
>
> – *or $P(F_1) = P(F_2)$ and $C(F_1) < C(F_2)$.*

A criterion can be even more complicated.

The only thing that a criterion *must* do is to allow us, for every pair of families (F_1, F_2), to make one of the following conclusions:

- the first family is better with respect to this criterion (we'll denote it by $F_1 \succ F_2$, or $F_2 \prec F_1$);
- with respect to the given criterion, the second family is better ($F_2 \succ F_1$);
- with respect to this criterion, the two families have the same quality (we'll denote it by $F_1 \sim F_2$);
- this criterion does not allow us to compare the two families.

Of course, it is necessary to demand that these choices be consistent.

> For example, if $F_1 \succ F_2$ and $F_2 \succ F_3$ then $F_1 \succ F_3$.

The criterion must be final, i.e., it must pick the unique family as the best one. A natural demand is that this criterion must choose a *unique* optimal family (i.e., a family that is better with respect to this criterion than any other family).

The reason for this demand is very simple:

- If a criterion *does not choose* any family at all, then it is of no use.
- If *several* different families are the best according to this criterion, then we still have the problem of choosing the best among them. Therefore we need some additional criterion for that choice, like in the above example:

  > If several families $F_1, F_2, \ldots$ turn out to have the same classification ability ($P(F_1) = P(F_2) = \ldots$), we can choose among them a family with minimal computational complexity ($C(F_i) \rightarrow \min$).

  So what we actually do in this case is abandon that criterion for which there were several "best" families, and consider a new "composite" criterion instead: F_1 is better than F_2 according to this new criterion if:

  - either it was better according to the old criterion,
  - or they had the same quality according to the old criterion and F_1 is better than F_2 according to the additional criterion.

In other words, if a criterion does not allow us to choose a unique best family, it means that this criterion is not final, we'll have to modify it until we come to a final criterion that will have that property.

The criterion must not change if we change the measuring unit for x. The exact mathematical form of a function $f(x)$ depends on the exact choice of units for measuring the s coordinates $x^1, \ldots, x^s$ of $x \in R^s$. If we replace each of these units by a new unit that is λ times larger, then the same physical value that was previously described by a numerical value x^k will now be described, in the new units, by a new numerical value $\tilde{x}^k = x^k/\lambda_j$. For example, if we replace centimeters by inches, with $\lambda = 2.54$, then $x^k = 5.08$ cm becomes $\tilde{x}^k = x^k/\lambda = 2$ in. After this transformation, x changes to $\tilde{x} = x/\lambda$.

How will the expression for closeness $f(x)$ change if we use the new units? In terms of $\tilde{x}$, we have $x = \lambda \cdot \tilde{x}$. Thus, if we change the measuring unit for x, the same dynamics that was originally represented by a function $f(x)$, will be described, in the new units, by a function $\tilde{f}(x) = f(\lambda \cdot x)$.

Since we assumed that we have no information about the cluster radii, there is no reason why one choice of unit should be preferable to the other. Therefore, it is reasonable to assume that the relative quality of different families should not change if we simply change the units, i.e., if the family F is better than a family G, then the transformed family $\tilde{F}$ should also be better than the family $\tilde{G}$.

The criterion must not be change is we apply a rotation. Similarly, it is reasonable to require that the relative quality of two different families of functions do not change if we apply an arbitrary *rotation* around 0 in s-dimensional space R^s.

We are now ready for the formal definitions.

3.2 Definitions

Definition 1.

- *By a family F, we mean a differentiable function $f(x)$ from R^s to R.*
- *We say that a function $e(x)$ **belongs** to the family $f(x)$ (or that $f(x)$ **contains** the function $e(x)$) if $e(x) = C \cdot f(x)$ for some $C > 0$.*
- *Two families F and G are considered **equal** if they contain the same functions.*

Denotation. *Let's denote the set of all possible families by Φ.*

- the set of all pairs (F_1, F_2) of elements $F_1 \in \Phi$, $F_2 \in \Phi$, is usually denoted by $\Phi \times \Phi$.
- An arbitrary subset R of a set of pairs $\Phi \times \Phi$ is called a *relation* on the set Φ. If $(F_1, F_2) \in R$, it is said that F_1 and F_2 are in relation R; this fact is denoted by $F_1 R F_2$.

Definition 2. *A pair of relations $(\prec, \sim)$ on a set Φ is called **consistent** if it satisfies the following conditions, for every $F, G, H \in \Phi$:*

(1) *if $F \prec G$ and $G \prec H$ then $F \prec H$;*

(2) *$F \sim F$;*

(3) *if $F \sim G$ then $G \sim F$;*

(4) *if $F \sim G$ and $G \sim H$ then $F \sim H$;*

(5) *if $F \prec G$ and $G \sim H$ then $F \prec H$;*

(6) *if $F \sim G$ and $G \prec H$ then $F \prec H$;*

(7) *if $F \prec G$ then it is not true that $G \prec F$, and it is not true that $F \sim G$.*

Comment. The intended meaning of these relations is as follows:

- $F \prec G$ means that with respect to a given criterion, G is better than F;
- $F \sim G$ means that with respect to a given criterion, F and G are of the same quality.

Under this interpretation, conditions (1)–(7) have simple intuitive meaning:

(1) if G is better than F, and H is better than G, then H is better than F;

(2) every F is of the same quality as itself;

(3) if G is of the same quality as F, then F is of the same quality as G;

(4) if F is of the same quality as G, and G is of the same quality as H, then F is of the same quality as H;

(5) if G is better than F, and H is of the same quality as G, then H is also better than F;

(6) if F is of the same quality as G, and H is better than G, then H is better than F;

(7) if G is better than F, then F cannot be better than G and F cannot be of the same quality as G.

Definition 3. *Assume a set Φ is given. Its elements will be called alternatives.*

- *By an optimality criterion, we mean a consistent pair $(\prec, \sim)$ of relations on the set Φ of all alternatives.*
  - *If $F \succ G$ we say that F is better than G;*
  - *if $F \sim G$ we say that the alternatives F and G are equivalent with respect to this criterion.*
- *We say that an alternative F is optimal (or best) with respect to a criterion $(\prec, \sim)$ if for every other alternative G either $F \succ G$ or $F \sim G$.*
- *We say that a criterion is final if there exists an optimal alternative, and this optimal alternative is unique.*

Comment. In this paper, we will consider optimality criteria on the set Φ of all families.

Definition 4. *Let $\lambda > 0$ be a positive real number.*

- *By a λ-rescaling of a function $f(x)$ we mean a function $\tilde{f}(x) = f(\lambda \cdot x)$.*
- *By a λ-rescaling of a family of functions F we mean the family consisting of λ-rescalings of all functions from F.*

Denotation. *λ-rescaling of a family F will be denoted by $R_\lambda(F)$.*

Definition 5. *We say that an optimality criterion on Φ is unit-invariant if for every two families F and G and for every number $\lambda > 0$, the following two conditions are true:*

i) if F is better than G in the sense of this criterion (i.e., $F \succ G$), then $R_\lambda(F) \succ R_\lambda(G)$;

ii) if F is equivalent to G in the sense of this criterion (i.e., $F \sim G$), then $R_\lambda(F) \sim R_\lambda(G)$.

Definition 6. *Let $T : R^s \to R^s$ be a rotation around 0 in s-dimensional space.*

- *By a T-rotation of a function $f(x)$ we mean a function $\tilde{f}(x) = f(Tx)$.*
- *By a T-rotation of a family of functions F we mean the family consisting of T-rotations of all functions from F.*

Denotation. *T-rotation of a family F around 0 will be denoted by $T(F)$.*

Definition 7. *We say that an optimality criterion on Φ is rotation-invariant if for every two families F and G and for every rotation T, the following two conditions are true:*

i) if F is better than G in the sense of this criterion (i.e., $F \succ G$), then $T(F) \succ T(G)$;

ii) if F is equivalent to G in the sense of this criterion (i.e., $F \sim G$), then $T(F) \sim T(G)$.

Comment. As we have already remarked, the demands that the optimality criterion is final, unit-invariant, and rotation invariant are quite reasonable. At first glance they may seem rather trivial and therefore weak, because these demands do not specify the exact optimality criterion. However, these demands are strong enough, as the following theorem shows:

3.3 Main result

Theorem 1. *If a family F is optimal in the sense of some optimality criterion that is final, unit-invariant, and rotation-invariant, then every function $f(x)$ from this family F has the form $C \cdot |x|^\alpha$ for some real numbers C and α.*

Comments.

- Thus, our general approach provides a precise mathematical justification for the (highly successful) potential functions used in Fuzzy C-Means approach.
- Since none of the optimal functions are from the interval $[0,1]$, our result explains why we cannot restrict ourselves to membership functions $f(x)$, and why we need to consider the potential functions which can attain values outside the interval $[0,1]$.
- For reader's convenience, all the proofs are placed in the special (last) Proofs section.

4 Particular Case: Potential Function Which is the Least Sensitive To Measurement Errors

4.1 Formulation of the problem in informal terms, and motivations for the following definitions

What we are planning to do in this section. In the previous section, we analyzed the problem of choosing the optimal potential function under an arbitrary (reasonable) optimality criterion, and we ended up with the choice of $f(x) = C \cdot |x|^\alpha$ for arbitrary C and α. In this section, we will consider one of these criteria: the least possible sensitivity to measurement errors, and we will show which values α are optimal for this particular criterion.

Comment. A similar sensitivity analysis was undertaken for *fuzzy control* in [39, 41].

Measurement errors and how they affect the values of the potential function $f(x)$: the general idea. In the previous text, we assumed that the values $x_i^1, \ldots, x_i^s$ that characterize each object are known precisely. In reality, these values come from measurements, and these measurements are never 100% accurate.

As a result of the inevitable measurement errors, for each of these quantities, its measured value $\tilde{x}_i^k$ is, in general, different from the actual value x_i^k. Thus, in our clustering, instead of the desired values $f(x)$, we will use different values $f(\tilde{x})$.

It is natural to require that the potential function be chosen in such a way that it is the least sensitive to these errors.

How to describe measurement errors? The measurement error Δx_i^k is usually assumed to be a Gaussian distributed random variable with 0 average, and different measurement errors are independent random variables (see, e.g., [22, 43]).

A Gaussian random variable with 0 average is uniquely characterized by its standard deviation σ. Since we do not have any reason to believe that some measurements are more accurate than others, it is reasonable to assume that all measurement errors have the same standard deviation σ.

It is also reasonable to assume that the measurement errors are relatively small (so that we can neglect the terms that are quadratic or of higher order in terms of these errors). The reason for this restriction is that if these measurement errors $\tilde{x} - x$ can be big, then the numerical values $\tilde{x}$ which we used for classification are drastically different from the actual values x, and the resulting classification becomes based on wrong data and thus, stops making sense.

How to describe the effect of measurement errors on the values of the potential function. For small measurement errors, we can use the standard methodology of sensitivity analysis: expand $f(x)$ into the Taylor series, and neglect the terms that are quadratic (or of higher order) in terms of $\Delta x = \tilde{x} - x$. As a result, we conclude that

$$\Delta f(x) = f(\tilde{x}) - f(x) = \frac{\partial f}{\partial x^1}\Delta x^1 + \ldots + \frac{\partial f}{\partial x^s}\Delta x^s + o(\Delta x), \tag{5}$$

where $o(\Delta x)$ includes terms that are quadratic (or of higher order) in Δx. We assumed that $E[\Delta x^1] = \ldots = E[\Delta x^s] = 0$ (here, E stands for mathematical expectation), and that $E[(\Delta x^1)^2] = \ldots = E[(\Delta x^s)^2] = \sigma^2$. Since $\Delta x^1, \ldots, \Delta x^s$ are independent random variables, we have $E[\Delta x^i \cdot \Delta x^j] = E[\Delta x^i] \cdot E[\Delta x^j] = 0$ for $i \neq j$. Using these equalities, we can conclude from (5) that $E[\Delta f] = 0 + o(\sigma)$ and

$$E[\Delta f^2] = \left[\left(\frac{\partial f}{\partial x^1}\right)^2 + \ldots + \left(\frac{\partial f}{\partial x^s}\right)^2\right]\sigma^2 + o(\sigma^2). \tag{6}$$

We cannot use this value $E[\Delta f^2]$ as a measure of sensitivity of a potential function f at a point x, because it depends not only on this function f, but also on the standard deviation σ. So, as an estimate for sensitivity of a potential function f in x, it is reasonable to take the value $E[\Delta f^2]/\sigma^2$ that describes to what extent the error Δf is larger than the measurement error:

$$\frac{E[\Delta f^2]}{\sigma^2} = \left(\frac{\partial f}{\partial x^1}\right)^2 + \ldots + \left(\frac{\partial f}{\partial x^s}\right)^2 + o(1). \tag{6}$$

This ratio, generally speaking, also depends on σ: namely, this $o(1)$ part (i.e., a part that tends to 0 as $\sigma \to 0$) can depend on σ. However, since we are interested in the small values of σ, we can neglect this part (or, in mathematical terms, use the limit value of this ratio when $\sigma \to 0$). Therefore, we arrive at the following expression for estimating the average sensitivity of a potential function f at a point x:

$$S(f, x) = \left(\frac{\partial f}{\partial x^1}\right)^2 + \ldots + \left(\frac{\partial f}{\partial x^s}\right)^2. \tag{7}$$

This number describes the average sensitivity of f in a point $x \in R^s$. Therefore, as an overall average sensitivity $S(f)$, it is reasonable to take an average of $S(f, x)$ over all possible x, i.e.,

$$S(f) = \int S(f, x)\, dx. \tag{8}$$

Substituting (7) into (8), we arrive at the following formula:

$$S(f) = \int \left[\left(\frac{\partial f}{\partial x^1} \right)^2 + \ldots + \left(\frac{\partial f}{\partial x^s} \right)^2 \right] dx. \tag{9}$$

How to find the least sensitive potential function. To find a function f for which the integral $S(f)$ takes the smallest possible value, we can apply the standard techniques of variational calculus (see, e.g., [19, 25, 44, 48]). Variational equations for this integral lead to the so called Laplace equation

$$\frac{\partial^2 f}{(\partial x^1)^2} + \ldots + \frac{\partial^2 f}{(\partial x^s)^2} = 0. \tag{10}$$

So, the desired function f with the smallest sensitivity must satisfy this equation. We are now ready for the formal definitions.

4.2 Definitions and result

Definition 8. *We say that a differentiable function $f(x)$ from R^s to R^s is:*

- *rotation invariant if $f(Tx) = f(x)$ for all rotations T around 0;*
- *asymptotically decreasing if $f(x) \to 0$ as $x \to \infty$, and*
- *least sensitive with respect to measurement errors if it satisfies the Laplace equation (10).*

Theorem 2. $(s \geq 3)$ *If a potential function $f(x)$ is rotation invariant, asymptotically decreasing, and least sensitive with respect to measurement errors, then $f(x) = C \cdot |x|^{-(s-2)}$ for some real number C.*

Comments.

- Theorem 2 does not require the family to be final and unit invariant as in Theorem 1. The form $f(x) = C \cdot |x|^{\alpha}$ arises from the least sensitivity criterion (i.e., $f(x)$ satisfies the Laplace equation) and in this particular case, we have $\alpha = -(s-2)$.
- If we do not require that the function $f(x)$ is asymptotically decreasing, then we get a more general expression $f(x) = C \cdot |x|^{-(s-2)} + C_1$, where C_1 is an additional real number.

5 Optimal Potential Functions: General Case

5.1 Motivations

What we are planning to do in this section. In the previous two sections, we considered the case when we have no prior information about the cluster radii. Let us now consider the general case, when we have some information about these radii.

Main idea. The main objective of the potential function is to formalize statements of the type "x is close to y". These statements form an important part of expert knowledge about the quantity x.

One expert may say that the unknown value of x is close to, say, an object $(1.0, 2.5, 3.0)$; another expert may say that x is close to $(0.9, 2.2, 2.8)$. We would like to be able to take both expert estimates into consideration. In other words, we must take into consideration the possibility that one expert says "x is close to y_1", another says that "x is close to y_2", and the resulting expert statement about x is "x is close to y_1 and x is close to y_2".

The statement "x is close to y_1" is described by a function $f(x-y_1)$; the statement "x is close to y_2" is described by the function $f(x-y_2)$; thus, the conjunction of these two statements can be described as $f_\&(f(x-y_1), f(x-y_2))$ for an appropriate t-norm ("and"-operation) $f_\&(a,b)$.

Thus, if we are looking for a family of functions of the type $f_i(x) = f(x-y)$, then we must require that this family be closed under some t-norm, i.e., if two functions $f_1(x)$ and $f_2(x)$ belong to this family, then the combination $f_\&(f_1(x), f_2(x))$ describes the situation when two experts independently express statements that correspond to $f_1(x)$ and $f_2(x)$.

Let us consider the simplest possible strictly Archimedean t-norm. It is reasonable to restrict ourselves to strictly Archimedean t-norms because it is known that every t-norm can be approximated, with an arbitrary accuracy, by strictly Archimedean t-norms [40].

It is known that strictly Archimedean t-norms are isomorphic to the algebraic product $f_\&(a,b) = a \cdot b$ (see, e.g., [41, 28]). Therefore, without losing generality, we can assume that the t-norm is actually the product. Thus, we will assume that our family of functions F is closed under multiplication: if $f_1(x) \in F$ and $f_2(x) \in F$, then $f_1(x) \cdot f_2(x) \in F$.

It is also reasonable to require that the family F contain the function $f(x) = 1$ that corresponds to the absence of expert information about x.

Terminological comment. In mathematics, families of functions that are closed under multiplication are called *multiplicative semigroups*. Semigroups that contain 1 are called *semigroups with unity*. Thus, in mathematical terms, the above requirements can be reformulated as the requirement that the family of functions F be a multiplicative semigroup with unity.

We want a finite-parametric family. We are interested in having a family of functions that would enable us to describe any possible knowledge. Since our final goal is to use that information for automated classification, these functions must be representable in a computer, i.e., we must have a function with several adjustable parameters that would enable us to represent any function from F. Since in finite time we can estimate the values of only finitely many parameters, we need a function with finitely many parameters. So the family F must be *finite-dimensional* in the sense that fixing the values of finitely many (m) parameters would be sufficient to pick any function from this family.

The optimality criterion must be shift-invariant and unit-invariant. If we know the cluster radius, then the corresponding *function* $f(x)$ is not unit-invariant. However, since this radius can be an arbitrary number, it is reasonable to require that the optimality criterion that selects the best *family* be unit-invariant.

The desired family must contain the functions of the type $f(x-y)$ for all possible y; simply what we care about is the *difference* between the values x and y, the optimality criterion should not change if we simply use different starting points for measuring x^k, i.e., replace x by $\tilde{x} = x + a$, because this shift does not change the difference: $\tilde{x} - \tilde{y} = (x+a) - (y+a) = x - y$.

Now, we are ready for the formal definitions.

5.2 Definitions and results

We will from here on adopt the following definition of a family of potential functions.

Definition 9.

- *By a family we will understand a set of smooth everywhere positive membership functions defined on R^s.*
- *A family F is called a multiplicative semigroup with unity if $1 \in F$, and if for every two functions $f, g \in F$, their product $f \cdot g$ also belongs to F.*
- *We say that a family is m-dimensional, where m is an integer, if there exists a connected open region U in an m-dimensional space R^m and a differentiable mapping $f : U \times R^s \to [0,1]$ such that F coincides with the set of all functions $x \to f(\vec{c}, x)$ for different $\vec{c} \in U$.*

Comment. This definition describes in mathematical terms that we need to fix m parameters to describe a function from F.

Denotation. *The set of all m-dimensional multiplicative semigroups with unity will be denoted by Φ_m.*

Comment. In the present section, we consider optimality criteria on the set Φ_m of all m-dimensional multiplicative semigroups with unity.

Definition 10. *Let $a \in R^s$ be an arbitrary vector.*

- *By a* a*-shift of a function $f(x)$ we mean a function $\tilde{f}(x) = f(x+a)$.*
- *By a* a*-shift of a family of functions F we mean the family consisting of a-shifts of all functions from F.*

Denotation. *a-shift of a family F will be denoted by $S_a(F)$.*

Definition 11. *We say that an optimality criterion on Φ_m is* shift-invariant *if for every two families F and G and for every vector $a \in R^s$, the following two conditions are true:*

i) if F is better than G in the sense of this criterion (i.e., $F \succ G$), then $S_a(F) \succ S_a(G)$;

ii) if F is equivalent to G in the sense of this criterion (i.e., $F \sim G$), then $S_a(F) \sim S_a(G)$.

Theorem 3. *If a multiplicative semigroup with unity F is optimal in the sense of some optimality criterion that is final, unit-invariant, and shift-invariant, then every function $f(x)$ from this family F has the form $\exp(-P(x))$ for some polynomial $P(x) = P(x^1, \ldots, x^s)$.*

Comment. To justify the use of Gaussian functions, we can now use rotation invariance:

Corollary. *If a multiplicative semigroup with unity F is optimal in the sense of some optimality criterion that is final, unit-invariant, and shift-invariant, then every rotation invariant function $f(x)$ from this family F has the form $f(x) = C\exp(-\alpha_1 \cdot |x|^2 - \alpha_2 \cdot |x|^4 - \ldots - \alpha_k \cdot |x|^{2k})$ for some real numbers $\alpha_1, \ldots, \alpha_k$.*

Comments.

- In practice, the coefficients α_i will be chosen according to available cluster radius information. In some sense, the cluster radius is related to the resolution of the data. As such, since the resolution is more fundamental, the α_i will be determined from the knowledge of the resolution.
- The simplest possible potential function of this type is when only one of the non-trivial coefficients α_i is different from 0. In this case, we get a Gaussian potential function. Thus, our general approach provides a precise mathematical justification for (successful) Gaussian potential functions used in [11, 12].
- In this general case, we can use only potential functions whose values are from the interval $[0, 1]$, so, there is no need for values outside this interval.
- For *fuzzy control*, a similar justification of Gaussian membership functions was proposed in [32, 34, 35, 36]

6 Optimal Consistent Potential Functions

What we are planning to do in this section. In the previous section, we considered arbitrary potential functions, and chose a potential function that is optimal with respect to some reasonable criterion.

It may be reasonable, however, instead of considering *arbitrary* possible potential functions, to impose some *conditions* on potential functions, and select the best only among the potential functions that satisfy these additional conditions. In this section, we will describe what happens if we impose one of these conditions: consistency.

Main idea. Let us assume that the potential function is rotation invariant, i.e., that $f(x-y)=f_0(|x-y|)$ for some function $f_0(r)$ of one variable. The value $f_0(r)$ describes, crudely speaking, the expert's degree of belief that objects x and y are close if $|x-y|\le r$.

The notion of *closeness* is a fuzzification of the crisp notion of *equality*. Equality is transitive: if x is equal to y, and y is equal to z, then x is equal to z. For closeness, it is natural to impose a similar requirement: if x is close to y, and y is close to z, then x is close to z. In other words, if we know the degree of belief d_1 that x is close to y, and the degree of belief d_2 that y is close to z, then x is close to z with at least the degree of belief $f_{\&}(d_1,d_2)=d_1\cdot d_2$.

Each degree of belief d in closeness can be reformulated in terms of the corresponding distance, i.e., distance r for which $f_0(r)=d$. So, if the degree of belief d_1 corresponds to distance r_1 (in the sense that $f_0(r_1)=d_1$), and degree of belief d_2 corresponds to the distance r_2 (in the sense that $f_0(r_2)=d_2$), then the original knowledge about the closeness between x, y, and z can be expressed as follows: we know that $|x-y|\le r_1$, and we know that $|y-z|\le r_2$. In this case, geometrically, the only thing we can conclude about x and z is that $|x-z|\le r_1+r_2$. This conclusion correspond to the degree of belief $f_0(r_1+r_2)$.

Thus, for the statement "x is close to z", we have two different different degrees of belief:

- the degree $d_1\cdot d_2=f_0(r_1)\cdot f_0(r_2)$ that comes from fuzzy logic, and
- the degree $f_0(r_1+r_2)$ that comes from geometry.

It is reasonable to require that these two approaches be *consistent* and thus, that the corresponding degrees coincide: $f_0(r_1+r_2)=f_0(r_1)\cdot f_0(r_2)$. Such *consistent* potential functions are easy to describe:

Definition 12. *A continuous rotation invariant potential function $f(x)=f_0(|x|)$ is called consistent if for every $r_1,r_2>0$, we have*

$$f_0(r_1+r_2)=f_0(r_1)\cdot f_0(r_2).$$

Theorem 4. *Every consistent potential function has the form*

$$f(x) = \exp(-\alpha \cdot |x|)$$

for some real value α*.*

Comment. Thus, we have justified the use of these functions in [49, 50]. Again, the constant α will in practice be chosen according to the cluster radius information or, more fundamentally, the prescribed resolution of the data.

7 Proofs

7.1 Proof of Theorem 1

This proof is based on the following auxiliary result of independent interest:

Proposition 1. *If an optimality criterion is final and unit-invariant, then the optimal family* F_{opt} *is also unit-invariant, i.e.,* $R_\lambda(F_{opt}) = F_{opt}$ *for every number* λ*.*

Proof of Proposition 1. Since the optimality criterion is final, there exists a unique family F_{opt} that is optimal with respect to this criterion, i.e., for every other F:

- either $F_{opt} \succ F$
- or $F_{opt} \sim F$.

To prove that $F_{opt} = R_\lambda(F_{opt})$, we will first show that the re-scaled family $R_\lambda(F_{opt})$ is also optimal, i.e., that for every family F:

- either $R_\lambda(F_{opt}) \succ F$
- or $R_\lambda(F_{opt}) \sim F$.

If we prove this optimality, then the desired equality will follow from the fact that our optimality criterion is final and therefore, there is only one optimal family (so, since the families F_{opt} and $R_\lambda(F_{opt})$ are both optimal, they must be the same family).

Let us show that $R_\lambda(F_{opt})$ is indeed optimal. How can we, e.g., prove that $R_\lambda(F_{opt}) \succ F$? Since the optimality criterion is unit-invariant, the desired relation is equivalent to $F_{opt} \succ R_{\lambda^{-1}}(F)$. Similarly, the relation $R_\lambda(F_{opt}) \sim F$ is equivalent to $F_{opt} \sim R_{\lambda^{-1}}(F)$.

These two equivalences allow us to complete the proof of the proposition. Indeed, since F_{opt} is optimal, we have one of the two possibilities:

- either $F_{opt} \succ R_{\lambda^{-1}}(F)$,
- or $F_{opt} \sim R_{\lambda^{-1}}(F)$.

In the first case, we have $R_\lambda(F_{opt}) \succ F$; in the second case, we have $R_\lambda(F_{opt}) \sim F$.

Thus, whatever family F we take, we always have either $R_\lambda(F_{opt}) \succ F$, or $R_\lambda(F_{opt}) \sim F$. Hence, $R_\lambda(F_{opt})$ is indeed optimal and thence, $R_\lambda(F_{opt}) = F_{opt}$. The proposition is proven.

A similar statement is true for rotation invariance. Similarly, we can prove that if an optimality criterion is final and rotation-invariant, then the optimal family F_{opt} is also rotation-invariant, i.e., $T(F_{opt}) = F_{opt}$ for every rotation T.

Conclusions: the optimal family is unit-invariant and rotation invariant. Let us now prove the theorem. Since the criterion is final, there exists an optimal family $F_{opt} = \{C \cdot f(x)\}$. Due to the Proposition, the optimal family is unit-invariant and rotation-invariant.

Using rotation invariance. In particular, since $f(x) \in F_{\rm opt}$, rotation invariance means that for every rotation T, the function $f(Tx)$ also belongs to the optimal family, i.e., that for every T, there exists a real number $C(T)$ such that

$$f(Tx) = C(T) \cdot f(x)$$

for all $x \in R^s$.

Since the function $f(x)$ is assumed to be differentiable (hence, continuous), we can conclude that the ratio $C(T) = f(Tx)/f(x)$ is a continuous function of T.

Let us consider a rotation T that is a composition of two other rotations $T = T_1 \circ T_2$. Then, the above equality takes the form

$$f(T_1 \circ T_2(x)) = C(T_1 \circ T_2) \cdot f(x).$$

On the other hand, we can apply a similar equality first for T_2, and then for T_1, thus getting

$$f(T_2x) = C(T_2) \cdot f(x)$$

and

$$f(T_1 \circ T_2(x)) = f(T_1(T_2x)) = C(T_1) \cdot f(T_2x) = C(T_1) \cdot C(T_2) \cdot f(x).$$

Comparing the two formulas for $f(T_1 \circ T_2(x))$, we conclude that $C(T_1 \circ T_2) = c(T_1) \cdot C(T_2)$.

Similarly, we can conclude that for every sequence of n rotations, we have:

$$C(T_1 \circ \ldots \circ T_n) = C(T_1) \cdot \ldots \cdot a(T_n).$$

In particular, if we take $T_1 = \ldots = T_{2n+1} =$ rotation by an angle $2\pi/(2n+1)$ around the same axis, we have an identity transformation id as $T_1 \circ \ldots \circ T_{2n+1}$ (for which $C(id) = 1$), and therefore, $C^{2n+1}(T_i) = 1$. Hence, $C(T_i) = 1$.

An arbitrary rotation by an angle $2\pi \cdot p/(2n+1)$, with integer p and n, can be represented as a composition of p rotations by an angle $2\pi/(2n+1)$. For each of these angles, as we have already shown, $C = 1$. Therefore, for their composition, we also have $C(T) = 1$.

Let us now show that $C(T) = 1$ for an arbitrary rotation T. Indeed, let α be this rotation's angle. The real number $\alpha/(2\pi)$ can be represented as a limit of rational numbers $p/(2n+1)$; therefore, the angle α is equal to the limit of angles $2\pi \cdot p/(2n+1)$, and hence, the rotation T can be represented as a limit of rotations T_k by angles $2\pi \cdot p/(2n+1)$. We already know that for all these rotations, $C(T_k) = 1$, and since the function $C(T)$ is continuous, we conclude that $C(T) = 1$ for an arbitrary rotation T, i.e., $f(Tx) = f(x)$.

Every two points x, x' from R^s for which $|x| = |x'|$ can be transformed to each other by an appropriate rotation T around 0, i.e., $Tx = x'$. Hence, $f(x') = f(Tx) = f(x)$. Thus, the value $f(x)$ can depend only on the length $|x|$ of the vector x, i.e., $f(x) = f_0(|x|)$ for some function $f_0(r)$ of one real variable.

Using unit invariance. To determine the exact type of this dependence, let us use the *unit-invariance*. From unit-invariance, it follows that for every λ, there exists a real number $A(\lambda)$ for which $f(\lambda \cdot x) = A(\lambda) \cdot f(x)$. Substituting the expression of $f(x)$ in terms of $f_0(r)$, we conclude that

$$f_0(\lambda \cdot r) = A(\lambda) \cdot f_0(r)$$

for every two positive real numbers $r, \lambda > 0$.

Since the function $f(x)$ is differentiable, we can conclude that the ratio $A(\lambda) = f(\lambda \cdot x)/f(x)$ is differentiable as well, and that the function $f_0(r)$ is also differentiable for $r > 0$. Thus, we can differentiate both sides of the above equation with respect to λ, and substitute $\lambda = 1$. As a result, we get the following differential equation for the unknown function $f_0(r)$:

$$r \cdot \frac{df_0}{dr} = \alpha \cdot f_0,$$

where by α, we denoted the value of the derivative $dA/d\lambda$ taken at $\lambda = 1$. Moving terms dr and r to the right-hand side and all the term containing f_0 to the left-hand side, we conclude that

$$\frac{df_0}{f_0} = \alpha \cdot \frac{dr}{r}.$$

Integrating both sides of this equation, we conclude that $\ln(f_0) = \alpha \cdot \ln(r) + C$ for some constant C, and therefore, that $f_0(r) = \text{const} \cdot r^\alpha$. Thus, $f(x) = f_0(|x|) = \text{const} \cdot |x|^\alpha$. The theorem is proven.

7.2 Proof of Theorem 2

Similarly to the proof of Theorem 1, we conclude that since the function $f(x)$ is rotation invariant (i.e., $f(Tx) = f(x)$ for all rotations T around 0), then

$f(x) = f_0(|x|)$ for some function $f_0(r)$ of one real variable r. Such rotation-invariant solutions of Laplace equation are well known (see, e.g., [10, 48]) and lead to the desired expression for the potential function $f(x)$.

7.3 Proof of Theorem 3

The optimal family exists and is invariant. Similarly to Proposition 1, we can conclude that the optimal family F_{opt} exists, and that this optimal family is unit-invariant, and that this family is *shift-invariant* in the sense that $F_{opt} = S_a(F_{opt})$ for all vectors $x \in R^s$.

From multiplicative to additive semigroup. Let's simplify the problem a little bit. A family F is closed under multiplication, i.e., it contains a product of every two of its elements. There is a well-known way to go from multiplications to a simpler operation (addition): namely, if we consider logarithms instead of the original membership functions, then the set of these logarithms will be closed with respect to addition (i.e., contains a sum of any two of its elements), because $\log(ab) = \log(a) + \log(b)$. We can always consider logarithms, because we consider only positive membership functions.

So let us consider the set L of all functions of the type $\log(f(x))$, where $f(x) \in F$. Let's first prove that the set L is closed under addition, i.e., if $l_1, l_2 \in L$, then $l_1 + l_2 \in L$. In mathematical terms we can express it by saying that L is an *additive semigroup*.

Indeed, if $l_1 \in L$, then by the definition of L, we have $l_1 = \log(f_1)$ for some $f_1 \in F$; here $f_1 = \exp(l_1)$, so this is equivalent to saying that $\exp(l_1) \in F$. Likewise, from $l_2 \in L$ we conclude that $\exp(l_2) \in F$. Since F is closed under multiplication, we conclude that the product $\exp(l_1) \cdot \exp(l_2)$ also belongs to F. Therefore, $\log(\exp(l_1) \cdot \exp(l_2)) \in L$, but this logarithm is precisely $l_1 + l_2$.

L is an invariant m-parametric family. From the fact that a family F is invariant we conclude that if a function $l(x)$ belongs to L, then for every $\lambda > 0$ and a the functions $l(\lambda \cdot x)$ and $l(x + a)$ also belong to L. So, L is invariant.

F is an m-dimensional family, i.e., it could be obtained by choosing values of parameters $\vec{u} \in R^m$ of some function $f(\vec{u}, x)$ in a connected open region. All functions from L are just logarithms of functions from F, so they are obtained from $\log f(\vec{u}, x))$ for different values of $\vec{u}$. Therefore, L is also an m-dimensional family.

From an additive semigroup to an additive group. Let us now consider the set D of all the functions $p(x)$ that can be represented as differences between the two functions from L, i.e., the set of all the functions of the type $p(x) = l_1(x) - l_2(x)$ for some $l_i \in L$.

Let us prove that this set D is a group under addition. We must prove that $0 \in D$, that if $p \in D$, then $-p \in D$, and that if p and q belong to D, then $p + q \in D$.

Since $1 \in F_{\rm opt}$, we have $0 = \log(1) \in L$ and hence, $0 = 0 - 0 \in D$.

If p belongs to D, this means that $p = l_1 - l_2$, where $l_i \in L$, but in this case $-p = l_2 - l_1$ and, therefore, $-p$ also belongs to D.

Finally, if $p \in D$ and $q \in D$, this means that $p = l_1 - l_2$ and $q = m_1 - m_2$, where $l_i \in L$ and $m_i \in L$. In this case, $p + q = (l_1 + m_1) - (l_2 + m_2)$, where both sums $l_1 + m_1$ and $l_2 + m_2$ belong to L, because L is an additive semigroup. Thus, $p + q \in D$.

D is an invariant finite-parametric family. It is easy to show that if $p(x) \in D$, then $p(x + a) \in D$ and $p(\lambda \cdot x) \in D$ for any a and $\lambda > 0$.

Indeed, if $p(x) \in D$, then $p(x) = l_1(x) - l_2(x)$ for some $l_i(x) \in L$. Since $l_i(x) \in L$, we have $l_1(x + a) \in L$ and $l_2(x + a) \in L$, and, therefore, their difference $l_1(x + a) - l_2(x + a)$ also belongs to D, but this difference is equal to $p(x + a)$. For $p(\lambda \cdot x)$, the proof is similar.

Let us now show that the family D is finite-parametric. Indeed, since we need m parameters to describe l_1 and m to describe l_2, we need to have at most $2m$ parameters to describe any function from D, so D is a $2m$-dimensional family.

D is a linear space. So D is a continuous finite-dimensional additive subgroup of the group of all functions. All such subgroups are known: they are linear subspaces. So we come to a conclusion that D is a finite-dimensional linear space.

This means that there exists a finite set of functions $p_1(x), p_2(x), \ldots, p_r(x)$ from D (they are called a *basis*) that are linearly independent, and such that any other function from D can be represented as a linear combinations of the functions from this base, i.e., as $\sum C_i \cdot p_i(x)$ for some coefficients C_i.

Let us use shift-invariance. Shift-invariance means that if $e(x) \in D$, then $e(x + a) \in D$ (i.e., the shifted function $e(x + a)$ can be represented as a linear combination of the basis functions $p_j(x)$). In particular, since each function $p_i(x)$ belongs to the family D, we conclude that for every $a \in R^s$, there exist values $C_i(a)$ for which

$$p_i(x + a) = C_{i1}(a) \cdot p_1(x) + \ldots + C_{ir}(a) \cdot p_r(x).$$

If we take r different values of x, then the corresponding equations form a system of r linear equations to determine r coefficients $C_{i1}(a), \ldots, C_{ir}(a)$. The well-known Cramer's rule describes the solution of a system of linear equation as a ratio of two determinants. Since $p_j(x)$ are differentiable functions, we can thus conclude that the functions $C_{ij}(a)$ are differentiable too.

Since both sides of the equation in concern are differentiable, let us differentiate both sides with respect to a^l, and then substitute $a^1 = \ldots = a^s = 0$. As a result, we get the following system of differential equations:

$$\frac{\partial p_j(x)}{\partial x^l} = \sum_{k=1}^{r} c_{jkl} \cdot p_k(x),$$

where we denoted

$$c_{jkl} = \frac{\partial C_{jk}(a^1, \ldots, a^s)}{\partial a^l}_{|a^1 = \ldots = a^s = 0}.$$

If we fix all the variables but one (e.g., except for x^1), we conclude that the functions $p_1(x^1), \ldots, p_r(x^1)$ satisfy a system of linear differential equations with constant coefficients. A general solution of such a system is well known (see, e.g., [2]): it has a form

$$p_j(x^1) = \sum_p A_{jp} \cdot \exp(\alpha_p \cdot x^1) \cdot (x^1)^{k_p},$$

where α_p are complex numbers (eigenvalues of the coefficient matrix), A_{jp} are complex numbers, and k_p are non-negative integers.

If we take into consideration the dependence on x^2, then all the coefficients should depend on x^2, i.e.,

$$p_j(x^1, x^2) = \sum_p A_{jp}(x^2) \cdot \exp(\alpha_p(x^2) \cdot x^1) \cdot (x^1)^{k_p(x^2)}.$$

Since the dependence on x^2 is smooth (hence, continuous), and k_p is an integer, we conclude that k_p is a constant: $k_p(x^2) = k_p$. The dependence on all other coefficients on x^2 can be determined from the fact that, similarly, for a fixed x^1, we must have a similar expression in terms of x^2:

$$p_j(x^2) = \sum_p A'_{jp} \cdot \exp(\alpha'_p \cdot x^2) \cdot (x^2)^{k'_p}.$$

Thus, the only possible dependence of A_{jp} on x^2 is a dependence of the type $\exp(\alpha'_p \cdot x^2) \cdot (x^2)^{k'_p}$, and the only possible dependence of α_p on x^2 is linear, i.e., we get

$$p_j(x^1, x^2) = \sum_p A_{jp} \cdot \exp(\alpha_{p1} \cdot x^1 + \alpha_{p2} \cdot x^2 + \alpha'_p \cdot x^1 \cdot x^2) \cdot (x^1)^{k_{p1}} \cdot (x^2)^{k_{p2}}.$$

Similarly, for s variables, we get

$$p_j(x^1, \ldots, x^s) = \sum_p A_{jp} \cdot \exp(B_{jp}(x^1, \ldots, x^s)) \cdot M_{jp}(x^1, \ldots, x^s),$$

where each B_{jp} is a multi-linear function, and M_{jp} is a monomial.

Let us now use unit invariance. Due to unit invariance, if $p_j(x)$ belongs to the family D, then, for every $\lambda > 0$, the re-scaled function $p_j(\lambda \cdot x)$ also belongs to D.

If the term B_{jp} is different from constant, e.g., if we have $B_{jp}(x) = c \cdot x^1 + \ldots$, then for different λ, we have infinitely many linearly independent functions

$p_j(\lambda \cdot x)$, all within the same finite-dimensional linear space, which is impossible. Thus, each terms B_{jp} is a constant, and therefore, each function $p_j(x)$ is a polynomial.

So, every function $e(x)$ from the set D is a linear combination of polynomials and hence, a polynomial itself. Since $L \subseteq D$, we can conclude that all elements of the family L are also polynomials.

By definition of the family L, every function $f(x) \in F_{\rm opt}$ has the form $\exp(l(x))$ for some $l \in L$ and thus, this function has the form $\exp(-P(x))$ for some polynomial $P(x) = P(x^1, \ldots, x^s)$.

Since we started with a representation in which the coefficients $A_{jp}, \ldots$ were complex numbers, we can only conclude that $P(x) = c + c_1 \cdot x^1 + \ldots + c_s \cdot x^s + c_{11} \cdot (x^1)^2 + c_{12} \cdot x^1 \cdot x^2 + \ldots$ is a polynomial with complex coefficients. Let us show that these coefficients are real numbers. Indeed, the function $f(x)$ is positive, therefore, $P(x) = -ln(f(x))$ is a real number for all x. So, for all x, the values of the polynomial $P(x)$ are real numbers. It is known that the coefficients of a polynomial can be obtained from its values, namely:

- The free term c can be computed as $c = P(0, \ldots, 0)$; since the polynomial takes real values, the coefficient c is a real number.
- the coefficients c_i at x^i can be defined as first partial derivatives of the polynomial P with respect to x^i, taken at the point $x^1 = \ldots = x^s = 0$. Since the polynomial P takes real values only, its first derivatives are also real numbers, and therefore, the coefficients c_i are real numbers.
- Similarly, the coefficients c_{ij} at $x^i \cdot x_j$ can be computed from the second derivatives of the polynomial P and are, therefore, real values; the coefficients c_{ijk} can be computed from third derivatives, etc.

Thus, all the coefficients of the polynomial $P(x)$ are real numbers. So, $f(x) = \exp(-P(x))$ for a polynomial with real coefficients. The theorem is proven.

Proof of the Corollary. According to Theorem 3, every function $f(x)$ from the family F has the form $f(x) = \exp(-P(x))$ for some polynomial $P(x) = P(x^1, \ldots, x^s)$. This function is rotation-invariant if and only if the corresponding polynomial $P(x)$ is rotation-invariant.

Rotation invariance means that $P(x) = P(Tx)$ for every rotation T around 0. Every two points x and x' for which $|x| = |x'|$ can be transformed into each other by an appropriate rotation T around 0: $x' = Tx$. Therefore, if $|x| = |x'|$, we have $P(x) = P(x')$. Thus, the value of the polynomial $P(x) = P(x^1, \ldots, x^s)$ only depends on $|x|$: $P(x) = P_0(|x|)$ for some function $P_0(r)$ of one real variable.

For every real number r, we can construct a vector $x_r = (r, 0, \ldots, 0)$ for which $|x_r| = r$. For this vector, $P_0(r) = P(x_r) = P(r, 0, \ldots, 0)$. Since P is a polynomial of all its variables, we can thus conclude that $P_o(r) = P(r, 0, \ldots, 0)$ is also a polynomial, i.e., that $P_0(r) = b_0 + b_1 \cdot r + b_2 \cdot r^2 + \ldots + b_p \cdot r^p$ for some integer p. Thus, the function $P(x) = P_0(|x|)$ takes the form $P(x) = b_0 + b_1 \cdot |x| + b_2 \cdot |x|^2 + \ldots + b_p \cdot |x|^p$.

This expression $P(x)$ must be a polynomial of all its variables. Even terms in the above expression are polynomials: indeed, $|x|^2 = (x^1)^2 + \ldots + (x^s)^2$ is a polynomial, and thus, so is $|x|^{2i} = (|x|^2)^i$. However, odd terms like $|x| = \sqrt{(x^1)^2 + \ldots + (x^s)^2}$ and $|x|^{2i+1}$ are *not* polynomials. Thus, the only possibility for a function $P(x)$ to be a polynomial is when all these odd terms are equal to 0. In this case, $P(x)$ is a linear combination of the even powers of $|x|$: $P(x) = \alpha_0 + \alpha_1 \cdot |x|^2 + \alpha_2 \cdot |x|^4 + \ldots + \alpha_k \cdot |x|^{2k}$ for some k. Thus, if we denote $C = \exp(-\alpha_0)$, we get the desired expression for $f(x) = \exp(-P(x))$.

7.4 Proof of Theorem 4

It is known (see, e.g., [1]), that every continuous solution to the functional equation $f_0(r_1 + r_2) = f_0(r_1) \cdot f_0(r_2)$ has the form $f_0(r) = \exp(-\alpha \cdot r)$ for some real value α. Hence, $f(x) = f_0(|x|) = \exp(-\alpha \cdot |x|)$. The theorem is proven.

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCCW-0089, by NSF grants No. DUE-9750858 and EEC-9322370, and by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518.

One of the authors (V.K.) is greatly thankful to Andrei Finkelstein, Olga Kosheleva, and Ronald R. Yager for helpful discussions.

Part of this work was conducted while one of the author (H.T.N.) was visiting the Department of Mechanical and Automation Engineering at the Chinese University of Hong Kong under the support of RGC Earmarked grant RGC519/95E.

References

[1] J. Aczel, *Lectures on Functional Equations and Their Applications*, Academic Press, NY-London, 1966.

[2] R. E. Bellman. *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1970.

[3] J. C. Bezdek, "Numerical taxonomy with fuzzy sets", *Journal of Mathematical Biology*, 1974, Vol. 1, pp. 57–71.

[4] J. C. Bezdek, "Cluster validity with fuzzy sets", *Journal of Cybernetics*, 1973, Vol. 3, No. 3, pp. 58–71.

[5] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, NY, 1981.

[6] J. C. Bezdek, R. Hathaway, M. Sabin, and W. Tucker, "Convergence theory for fuzzy C-Means: counterexample and repairs", *IEEE Trans. Systems, Man, and Cybernetics*, 1987, Vol. SMC-17, pp. 873–877.

[7] J. C. Bezdek, R. Hathaway, M. Sabin, and W. Tucker, "Convergence theory for fuzzy C-Means: counterexample and repairs", In: J. Bezdek (ed.), *The Analysis of Fuzzy Information*, CRC Press, 1987, Vol. 3, Chapter 8.

[8] J. C. Bezdek and S. K. Pal (eds.) *Fuzzy models for pattern recognition*, IEEE Press, N.Y., 1992.

[9] B. Bouchon-Meunier, V. Kreinovich, A. Lokshin, and H. T. Nguyen, "On the formulation of optimization under elastic constraints (with control in mind)", *Fuzzy Sets and Systems,* 1996, Vol. 81, No. 1, pp. 5–29.

[10] A. Broman, *Introduction to Partial Differential Equations*, Dover, N.Y., 1989.

[11] S. Chiu, "Fuzzy model identification based on cluster estimation", *J. of Intelligent and Fuzzy Systems*, 1994, Vol. 2, No. 3, pp. 267–278.

[12] S. Chiu, "Selecting input variables for fuzzy models", *J. of Intelligent and Fuzzy Systems*, 1996, Vol. 4, pp. 243–256.

[13] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters", *Journal of Cybernetics*, 1973, Vol. 3, No. 3, pp. 32–57.

[14] A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry, error estimation, and other applications of set-valued analysis", *ACM SIGNUM Newsletter*, 1996, Vol. 31, No. 4, pp. 3–25.

[15] A. Finkelstein, O. Kosheleva, and V. Kreinovich, "Astrogeometry: towards mathematical foundations", *International Journal of Theoretical Physics*, 1997, Vol. 36, No. 4, pp. 1009–1020.

[16] A. Finkelstein, O. Kosheleva, and V. Kreinovich, Andrei Finkelstein, Olga Kosheleva, and Vladik Kreinovich, "Astrogeometry: geometry explains shapes of celestial bodies", *Geombinatorics*, 1997, Vol. VI, No. 4, pp. 125–139.

[17] A. M. Finkelstein and V. Kreinovich, "Derivation of Einstein's, Brans-Dicke and other equations from group considerations," *On Relativity Theory. Proceedings of the Sir Arthur Eddington Centenary Symposium, Nagpur India 1984*, Vol. 2, Y. Choque-Bruhat and T. M. Karade (eds), World Scientific, Singapore, 1985, pp. 138–146.

[18] A. M. Finkelstein, V. Kreinovich, and R. R. Zapatrin, "Fundamental physical equations uniquely determined by their symmetry groups," *Lecture Notes in Mathematics*, Springer-Verlag, Berlin-Heidelberg-N.Y., Vol. 1214, 1986, pp. 159–170.

[19] C. Fox, it An Introduction to the Calculus of Variations, Dover, N.Y., 1987.

[20] A. Freedman, R. Bose, and B. D. Steinberg, "Techniques to improve the CLEAN deconvolution algorithm", *J. Franklin Inst.*, 1995, Vol. 332B, No. 5, pp. 535–553.

[21] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, San Diego, CA, 1990.

[22] W. A. Fuller, *Measurement error models*, J. Wiley & Sons, New York, 1987.

[23] *Galactic and extra-galactic radio astronomy*, Springer-Verlag, NY, 1974.

[24] *Group theory in physics: proceedings of the international symposium held in honor of Prof. Marcos Moshinsky, Cocoyoc, Morelos, Mexico, 1991*, American Institute of Physics, N.Y., 1992.

[25] F. H. Hildebrand, *Methods of Applied Mathematics* (Dover, N.Y., 1992.

[26] *Instrumentation and techniques in radio astronomy*, IEEE Press, NY, 1988.

[27] A. Kandel, *Fuzzy techniques in pattern recognition*, Wiley-Interscience, NY, 1982.

[28] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications,* Prentice Hall, Upper Saddle River, NJ, 1995.

[29] O. M. Kosheleva, V. Ya. Kreinovich, and B. S. Minchenko, "A modification of the CLEAN algorithm of radioimage reconstruction." *Proceedings of the XYII National Radioastronomical Conference*, Erevan, 1985, p. 123 (in Russian).

[30] V. Kreinovich. "Derivation of the Schroedinger equations from scale invariance," *Theoretical and Mathematical Physics*, 1976, Vol. 8, No. 3, pp. 282–285.

[31] V. Ya. Kreinovich, "A modification of the generalized CLEAN method", *Notices of the American Mathematical Society*, 1979, Vol. 26, No. 5, p. A-438, Publ. No. 79T–C58.

[32] V. Kreinovich, Ching-Chuang Chang, L. Reznik, and G. N. Solopchenko. "Inverse problems: fuzzy representation of uncertainty generates a regularization", *Proceedings of NAFIPS'92: North American Fuzzy Information Processing Society Conference, Puerto Vallarta, Mexico, December 15–17, 1992*, NASA Johnson Space Center, Houston, TX, 1992, pp. 418–426.

[33] V. Kreinovich, C. Quintana, and R. Lea, "What procedure to choose while designing a fuzzy control? Towards mathematical foundations of fuzzy control", *Working Notes of the 1st International Workshop on Industrial Applications of Fuzzy Control and Intelligent Systems*, College Station, TX, 1991, pp. 123–130.

[34] V. Kreinovich, C. Quintana, R. Lea, O. Fuentes, A. Lokshin, S. Kumar, I. Boricheva, and L. Reznik. "What non-linearity to choose? Mathematical foundations of fuzzy control", *Proceedings of the 1992 International Conference on Fuzzy Systems and Intelligent Control*, Louisville, KY, 1992, pp. 349–412.

[35] V. Kreinovich, C. Quintana, and L. Reznik, "Gaussian membership functions are most adequate in representing uncertainty in measurements". *Proceedings of NAFIPS'92: North American Fuzzy Information Processing Society Conference, Puerto Vallarta, Mexico, December 15–17, 1992*, NASA Johnson Space Center, Houston, TX, 1992, pp. 618–625.

[36] V. Kreinovich and L. K. Reznik. "Methods and models of formalizing a priori information (on the example of processing measurements results)," *Analysis and Formalization of Computer Experiments, Proceedings of the Mendeleev Metrology Institute*, 1986, pp.37–41 (in Russian).

[37] V. Kreinovich and L. Longpré, "Unreasonable effectiveness of symmetry in physics", *International Journal of Theoretical Physics*, 1996, Vol. 35, No. 7, pp. 1549–1555.

[38] H. T. Nguyen and V. Kreinovich, *Applications of continuous mathematics to computer science*, Kluwer, Dordrecht, 1997.

[39] H. T. Nguyen, V. Kreinovich, and D. Tolbert, "A measure of average sensitivity for fuzzy logics", *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 1994, Vol. 2, No. 4, pp. 361–375.

[40] H. T. Nguyen, V. Kreinovich, and P. Wojciechowski, "Strict Archimedean t-Norms and t-Conorms as Universal Approximators", *International Journal of Approximate Reasoning* (to appear).

[41] H. T. Nguyen and E. A. Walker, *A first course in fuzzy logic,* CRC Press, Boca Raton, Florida, 1997.

[42] P. J. Olver, *Equivalence, invariants, and symmetry*, Cambridge University Press, Cambridge, N.Y., 1995.

[43] S. Rabinovich, *Measurement errors: theory and practice*, American Institute of Physics, N.Y., 1993.

[44] K. Rectorys, *Variational Methods in Mathematics, Science and Engineering*, D. Reidel, Dordrecht–Holland, Prague, Czechoslovakia, 1980.

[45] M. H. Smith and V. Kreinovich, "Optimal strategy of switching reasoning methods in fuzzy control", Chapter 6 in H. T. Nguyen, M. Sugeno, R. Tong, and R. Yager (eds.), *Theoretical aspects of fuzzy control*, J. Wiley, N.Y., 1995, pp. 117–146.

[46] *Symmetries in physics: proceedings of the international symposium held in honor of Prof. Marcos Moshinsky, Cocoyoc, Morelos, Mexico, 1991*, Springer-Verlag, Berlin, N.Y., 1992.

[47] Sze M. Tan, "An analysis of the properties of CLEAN and smoothness stabilized CLEAN - some warnings", *Mon. Not. R. Astron. Soc.*, 1986, Vol. 220, pp. 971–1001.

[48] V. S. Vladimirov, *Equations of Mathematical Physics*, Marcel Dekker, N.Y., 1971.

[49] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method", *IEEE Trans. Systems, Man and Cybernetics*, 1994, Vol. 24, No. 8, pp. 1279–1284.

[50] R. R. Yager and D. P. Filev, "Generation of fuzzy rules by mountain clustering", *Journal of Intelligent and Fuzzy Systems*, 1994, Vol. 2, No. 3, pp. 209–219.