AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 4, 375–390

The Problem of Classification when the Data are Non-precise

Mayer Alvo and François Théberge
University of Ottawa, Canada

Abstract: Non-precise data arise in a natural way in several contexts. For example, the water level of a river does not usually consist of a single number as can be seen from the intensity of the wetness as a function of depth of a survey rod. The temperature of a room varies as a function of distance from a reference point. The color intensities associated with a pixel which describe observations from remote sensing are non-precise numbers because they vary as a function of the reflection from the sun. In these examples, it is the imprecision of the observation itself that is of interest rather than the uncertainty due to statistical variation. Even in the absence of stochastic error, there would still be an imprecision in the measurement. Viertl (1997) developed the subject of statistical inference for such non-precise data and associated it very closely to fuzzy set theory. Precise data can be described by an indicator function whereas non-precise data is described by characterizing functions. In this article, we first review the notation and then consider the problems of classification for non-precise data.

Zusammenfassung: Unscharfe Daten entstehen auf natürliche Art in diversen Situationen. Beispielsweise ist der Pegelstand eines Flusses gewöhnlich keine einzelne Zahl. Die Stärke der Feuchtigkeit des Messstabes kann als Funktion der Tiefe gesehen werden. Die Raumtemperatur variiert als Funktion des Abstands zum Referenzpunkt. Die Farbstärken eines Pixels, die Beobachtungen bei der Fernerkundung beschreiben, sind unpräzise Zahlen, da diese in Abhängigkeit von der Sonnenspiegelung variieren. In all diesen Beispielen ist es vielmehr die Ungenauigkeit der Beobachtung selbst die interessiert, als die Unbestimmtheit wegen statistischer Streuung. Sogar bei Fehlen eines stochastischen Fehlers ist noch immer Ungenauigkeit in der Messung. Viertl (1997) entwickelte das Fach der statistischen Inferenz für derartige unscharfe Daten und verknüpfte es stark mit der Theorie unscharfer Mengen. Präzise Daten können durch eine Indikatorfunktion beschrieben werden während unpräzise Daten durch charakterisierende Funktionen dargestellt werden. In diesem Artikel besprechen wir zuerst die Notation und dann betrachten wir die Probleme der Klassifikation unscharfer Daten.

Keywords: Classification, Non-precise Data, Fuzzy Data.

1 Characterizing Functions

In the presentation below, we shall draw heavily on the analogy with inference for precise data. We shall put aside the presence of statistical error and assume that a precise measurement of a given quantity yields a single value. A precise measurement can be

uniquely represented by an indicator function, $I_{[x_0]}(x)$ which takes value 1 if $x = x_0$ and 0 otherwise. A non-precise observation will be mathematically modelled in terms of characterizing function which are a generalization of an indicator function. Quoting from Viertl (1997),

Definition 1 *A characterizing function $\xi(\cdot)$ of a non-precise number is a real function of a real variable such that*

(i) $\xi : \mathbb{R} \to [0, 1]$

(ii) $\exists x_0 \in \mathbb{R} : \xi(x_0) = 1$

(iii) $\forall \alpha \in (0, 1]$, *the set* $B_\alpha = \{x \in \mathbb{R} : \xi(x) \geq \alpha\} = [a_\alpha, b_\alpha]$ *is a finite closed interval called an α-cut of ξ.*

Viertl (1997) has shown that a characterizing function can be uniquely determined by the family of α-cuts $\{B_\alpha : \alpha \in (0, 1]\}$ and moreover

$$\xi(x) = \max_{\alpha \in (0,1]} \alpha I_{B_\alpha}(x)\,, \quad \forall x \in \mathbb{R}\,. \tag{1}$$

It should be noted that continuous functions fulfilling only conditions (i) and (ii) above can, through the notion of the convex hull, be made to also obey condition (iii). The characterizing function is the unique representation of a non-precise measurement. All inference is drawn on the basis of this representation.

Viertl (1997) points out that characterizing functions can be viewed as representing the rate of change of values and he provides a prescription for its construction. Referring to the example on the water level of a river, let $w(h)$ represent the intensity of the wetness of a survey rod as a function of the depth $h_1 \leq h \leq h_2$, where h_1, h_2 provide the range of values. Then the characterizing function can be given as

$$\xi(h) = \frac{w'(h)}{\max\limits_{h_1 \leq h \leq h_2} w'(h)}\,. \tag{2}$$

The derivative measures the rate of change of the wetness. For values of h close to h_1 or h_2, $\xi(h)$ should be near 0 since the rod would be either always wet or always dry and one expects very little change. Precise data is described by an indicator function $I_{[x_0]}(x)$, showing that the rate of change is 0 for values on either side of x_0.

Example 2 *Consider the characterizing function given by:*

$$\xi(x) = \exp\left\{-\frac{(x-\omega)^2}{2\tau^2}\right\}\,.$$

The α-cut boundaries are given by

$$B_\alpha = \{x \in \mathbb{R} : \xi(x) \geq \alpha\} = \left\{x \in \mathbb{R} : |x - \omega| \leq \sqrt{-2\tau^2 \log(\alpha)}\right\}\,.$$

Characterizing functions can also be defined for a non-precise $n-$dimensional vector x^*.

Definition 3 *A characterizing function $\xi_{x^*}(\cdot)$ of a non-precise vector x^* is a real function of n variables such that*

(i) $\xi_{x^*}(\cdot) : \mathbb{R}^n \to [0,1]$

(ii) $\exists \mathbf{x}_0 \in \mathbb{R}^n : \xi(\mathbf{x}_0) = 1$

(iii) $\forall \alpha \in (0,1]$, *the set* $B_\alpha(x^*) = \{\mathbf{x} \in \mathbb{R}^n : \xi_{x^*}(\mathbf{x}) \geq \alpha\}$ *is a star shaped compact subset of* $\mathbb{R}^n$, *by which we mean that the line segment joining any two points in the set lies entirely in the set.*

An example of a non-precise vector is the location of an object on a radar screen. The object appears as a cloud in two-dimensional space. The characterizing function may be constructed in terms of the light intensity function. Given n non-precise observations, $x_1^*, x_2^*, \ldots, x_n^*$, each taking values in a space M with corresponding characterizing functions $\xi_1, \ldots, \xi_n$, it is possible to define a characterizing function $\xi : M^n \to [0,1]$ for the combined sample via the product or the minimum rule respectively as

$$\xi(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \xi_i(x_i) \tag{3}$$

or

$$\xi(x_1, x_2, \ldots, x_n) = \min_{1 \leq i \leq n} \xi_i(x_i)\,. \tag{4}$$

The α-cuts for the combined sample (and hence the characterizing function) based on the minimum combination rule are easy to obtain from the α-cuts of the individual non-precise observations. Referring to Example 2, if every observation in a sample of size n has the same characterizing function, then the characterizing function for the combined sample using the product rule is:

$$\xi(x_1, \ldots, x_n) = \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \omega)^2}{2\tau^2}\right\}.$$

It can be shown that both the product and the minimum combination rules lead to functions which satisfy the conditions of Definition 3 above. In practice, the minimum rule appears to be the more useful of the two. We now turn attention to functions of non-precise observations, such as the usual sample mean and sample variance.

Definition 4 *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a real valued continuous function whose arguments are non-precise vectors x^* with characterizing function ξ. The characterizing function of the non-precise value $y^* = g(x^*)$ is defined $\forall y \in \mathbb{R}$ as*

$$\psi(y) = \begin{cases} \sup\{\xi(x) : x \in \mathbb{R}^n,\ g(x) = y\}\,, & \textit{for } g^{-1}([y]) \neq \varnothing \\ 0, & \textit{for } g^{-1}([y]) = \varnothing \end{cases}. \tag{5}$$

To demonstrate that the definition is reasonable, consider the sample sum. If the characterizing function of the individual measurements is represented by a rate of change, then the range of change in the sum is dictated by the greatest rate of change among the individual components. It can be shown that once again, ψ defined above is a characterizing function.

Assuming that $g : M \to \mathbb{R}$ is a continuous function with $M \subseteqq \mathbb{R}^n$ and $\sup(\xi(\cdot)) \subseteqq M$, Viertl (1997) showed that the general form for the α-cuts $(B_\alpha(y^*); \alpha \in (0,1])$ that define the characterizing function above consists of intervals of the form

$$B_\alpha(y^*) = \left[\min_{x \in B_\alpha(x^*)} g(x), \max_{x \in B_\alpha(x^*)} g(x) \right] . \tag{6}$$

In practice, this result coupled with (1) leads to the construction of the characterizing function. In order to deal with more general problems of inference in point and interval estimation as well as with Bayesian analysis, Viertl (1997) introduced the following generalization:

Definition 5 *A function $g^*(\cdot)$ with non-precise values is a mapping which assigns to every element $x \in M$ a non-precise number $g^*(x)$.*

Example 6 *Consider a sample $x_1^*, \ldots, x_n^*$ of size n, each with characterizing function given in Example 2. Let $g(x) = \sum_{i=1}^n x_i/n$. Then the characterizing function $\psi(\cdot)$ of $g(x_1^*, \ldots, x_n^*)$ is given $\forall y \in \mathbb{R}$ by its values:*

$$\psi(y) = \begin{cases} \sup \left\{ \exp\left[\frac{-\sum_{i=1}^n (x_i - \omega)^2}{2\tau^2} \right] : x \in \mathbb{R}^n , \sum_{i=1}^n x_i = ny \right\} , & g^{-1}([y]) \neq \varnothing \\ 0 , & g^{-1}([y]) = \varnothing \end{cases} ,$$

$$= \exp\left[\frac{-n(y-\omega)^2}{2\tau^2} \right] , \quad \textit{in the first case.}$$

This can be seen from the fact that

$$\sum_{i=1}^n (x_i - \omega)^2 = \sum_{i=1}^n (x_i - y)^2 + n(y - \omega^2) \tag{7}$$

and we can choose the x_i's so that $x_1 = \cdots = x_n = y$.

In the next example, we consider the sample variance.

Example 7 *Consider a sample $x_1^*, \ldots, x_n^*$ of size n, each with characterizing function given as in Example 2. Let $g(x) = \sum_{i=1}^n (x_i - \overline{x})^2/(n-1) = S_x^2$. Then the characterizing function of $g(x_1^*, \ldots, x_n^*)$ is $\forall y \in \mathbb{R}$ given by*

$$\psi(y) = \left\{ \begin{array}{ll} \sup \left\{ \exp\left[\frac{-\sum_{i=1}^n (x_i - \omega)^2}{2\tau^2} \right] : x \in \mathbb{R}^n , \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1} = y^2 \right\} , & g^{-1}([y]) \neq \varnothing \\ 0 , & g^{-1}([y]) = \varnothing \end{array} \right\} ,$$

$$= \exp\left[-\frac{(n-1)y^2}{2\tau^2} \right] , \quad \textit{in the first case.}$$

To demonstrate this, using (7), we may write

$$\psi(y) = \sup_{S_x^2 = y} \exp\left(\frac{-(n-1)y^2}{2\tau^2} - \frac{-n(\overline{x} - \omega)^2}{2\tau^2} \right) .$$

The first term is constant, and choosing $\overline{x} = \omega$, the second term is 0, so that

$$\psi(y) = \exp\left(\frac{-(n-1)y^2}{2\tau^2}\right) .$$

Setting $x_1 = \cdots = x_{n-1} = \omega \pm \sqrt{y^2/n}$ and $x_n = n\omega - (n-1)\left(\omega \pm \sqrt{y^2/n}\right)$ yields this optimum.

In both examples, the characterizing functions decrease exponentially fast as the sample size increases. The basis for inference involving non-precise data is the construction of the characterizing function $\xi(\cdot)$ of the n-dimensional non-precise vector describing the combined sample x^*. The statistical function $S(x_1, x_2, \ldots, x_n)$ which is the basis of inference for precise data $x = (x_1, x_2, \ldots, x_n)$ is then adapted for non-precise data by computing its characterizing function in accordance with the rule $\forall y \in \mathbb{R}$

$$\psi(y) = \left\{ \begin{array}{ll} \sup\left\{\xi(x) : x \in \mathbb{R}^n, S(x_1, x_2, \ldots, x_n) = y\right\} & \text{for } S^{-1}([y]) \neq \varnothing \\ 0, & \text{for } S^{-1}([y]) = \varnothing \end{array} \right\} . \quad (8)$$

We make use of this procedure in the problem of classification.

2 The Problem of Classification

As an application of a classification problem involving non-precise data, we may wish to identify the species of fish on the basis of echo sounder measurements. The depth at which fish travel does not consist of a single number but rather is a non-precise number. Moreover, different species may travel at depths described by different characterizing functions. In the simplest situation, for precise data, samples are observed from two populations π_1, π_2 described respectively by density functions $f_1(\cdot)$ and $f_2(\cdot)$. Let $c(i|j)$ be the cost of misclassification of class j as class i, $i \neq j$, and let $p(i)$ be the prior probability for class i. We would like to classify a new observation which may be vector valued, into either π_1, π_2 so as to minimize the expected cost of misclassification (*ECM*). It can be shown (Johnson and Wichern, 1999, chapter 11) that the optimal region consists of classifying the new observation x into population π_1 provided

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)p(2)}{c(2|1)p(1)} .$$

In what follows we will assume equal priors and equal costs, so that

$$\frac{c(1|2)p(2)}{c(2|1)p(1)} = 1 .$$

As an example, suppose that π_1, π_2 are described by normal populations with known means and covariances given respectively by (μ_1, Σ_1) and (μ_2, Σ_2). Assume $\Sigma_1 = \Sigma_2 = \Sigma$ and let

$$T(x; \mu_1, \mu_2, \Sigma) \equiv (\mu_1 - \mu_2)'\Sigma^{-1}\left(x - (\mu_1 + \mu_2)/2\right) . \quad (9)$$

The optimal classification rule for precise data consists of classifying a new observation x into population π_1 if and only if

$$T(x; \mu_1, \mu_2, \Sigma) \geq 0\,. \tag{10}$$

When the parameters (μ_1, μ_2, Σ) are unknown, samples of sizes n_1, n_2 respectively

$$x^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}\right)\,, \quad x^{(2)} = \left(x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)}\right)$$

are taken and used to calculate the corresponding sample means $\hat{\mu}_1, \hat{\mu}_2$ and pooled sample covariance

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2}{(n_1 - 1) + (n_2 - 1)}\,,$$

where $\hat{\Sigma}_1, \hat{\Sigma}_2$ are the respective sample covariances. The rule then consists of classifying a new observation x into population π_1 if and only if

$$T(x; \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}) \geq 0\,. \tag{11}$$

The statistic T can be considered a score function. Viewed in terms of $T(x; \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma})$, the region for classification of points into π_1 is always a subset of $\mathbb{R}$. For non-precise data $x^{(1)*}, x^{(2)*}, x^*$, the characterizing function of the non-precise value t^* of the statistic $T(x; \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma})$ is given by its values $(\forall t \in \mathbb{R})$

$$\psi(t) = \left\{ \begin{array}{ll} \sup\left\{\xi(x^{(1)}, x^{(2)}, x) : x^{(i)} \in M^{(i)}, x \in M, T = t\right\}\,, & T^{-1}(t) \neq \varnothing \\ 0\,, & T^{-1}(t) = \varnothing \end{array} \right\}\,, \tag{12}$$

where $M, M^{(i)}$ are the respective spaces for the non-precise observations. If the support of t^* is contained in either the interval $[0, \infty)$ or its complement $(-\infty, 0)$, the observation is classified into either π_1 or π_2, respectively. On the other hand, if the support has a non-empty intersection with the intervals $[0, \infty), (-\infty, 0)$, then the classification is ambiguous. We now consider some examples.

Example 8 *Consider the case for two univariate normal populations with known means μ_1, μ_2 and common known variance σ^2. Assume that the characterizing function of a new measurement x is given as in Example 2. Then the characterizing function of the non-precise value t^* is given by*

$$\psi(t) = \exp\left\{ -\frac{\sigma^2}{2\tau^2(\mu_1 - \mu_2)^2}\left[t - \frac{(\mu_1 - \mu_2)}{\sigma}\left(\omega - \frac{(\mu_1 + \mu_2)}{2}\right)\right]^2 \right\}\,.$$

The decision of where to place the measurement clearly depends on the support of t^*. Assume that $\mu_1 > \mu_2$. For values of $\omega \gg (\mu_1 + \mu_2)/2$, the characterizing function will be centered around a large positive number. Consequently, the measurement is likely to be classified into π_1. Conversely for values of $\omega \ll (\mu_1 + \mu_2)/2$, the measurement is likely to be classified into π_2. For values of $\omega \approx (\mu_1 + \mu_2)/2$, the characterizing function will be centered around 0 and then the classification will be ambiguous.

The example above can be generalized to the case where the parameters are unknown and are estimated on the basis of non-precise data. The latter will then serve to modify the characterizing function of t^*.

We now consider the general classification problem involving several populations. For precise data, Fisher recommended the use of sample linear discriminants. These are defined as follows. Let $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)})$ be a sample of observations from the ith population and define the mean vectors $\overline{x}_i = \sum_j x_j^{(i)}/n_i$, $\overline{x} = \sum_i \sum_j x_j^{(i)} / \sum_i n_i$. Define as well the between groups and within groups variation matrices respectively

$$B = \sum_i n_i(\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})' , \tag{13}$$
$$W = \sum_i \sum_j (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)' .$$

Let (λ_s) denote the non-zero eigenvalues of $W^{-1}B$ arranged in decreasing order and let (e_s) denote the corresponding eigenvectors. Then, the vector of coefficients l which maximizes the ratio $l'Bl/l'Wl$ is given by $l_1 = e_1$. The first sample linear discriminant is given by $d_1 = e_1'x$. In general, the sth sample discriminant is given by $d_s = e_s'x$ and these are used to classify a future observation x as follows. Compute the discriminants $y = (d_1, d_2, \ldots)'$ along with their vector of means $\mu_{Yi} = (e_1'\mu_i, e_2'\mu_i, \ldots)'$ under population π_i. We assign x to that population for which the distance $\|y - \mu_Y\|^2$ is smallest.

Let $D_i(x^{(1)}, x^{(2)}, \ldots) = \|y - \mu_{Yi}\|^2$ represent the distance of the discriminants to their mean under the ith population. Then the characterizing function of D_i is given by the following, $\forall t \in \mathbb{R}$

$$\psi_i(t) = \begin{cases} \sup\left\{\xi(x^{(1)}, x^{(2)}, \ldots, x) : x^{(j)} \in M^{(j)}, x \in M, D_i = t\right\}, & D_i^{-1}(t) \neq \varnothing \\ 0, & D_i^{-1}(t) = \varnothing . \end{cases} \tag{14}$$

In the case where a single characterizing function, say from population π_k, emerges clearly to the left of all the others, then the decision consists of classifying the measurement into π_k. In instances where the regions of support of the characterizing functions overlap, there will be ambiguity in the classification. The calculations involved in (14) are illustrated in the next section for normal populations using the notion of α-cuts.

Example 9 *Suppose that it is desired to classify a non-precise measurement x^* into one of several multivariate populations having means μ_i and covariances Σ_i. The usual classification rule in the case where the data are precise consists of allocating x to that population π_k for which the quadratic score $d_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + \log p_k$ is largest. Here, $\{p_k t\}$ represent the prior probabilities of selection of the populations. If the parameters are unknown, they are replaced by standard estimates $\hat{\mu}_k, \hat{\Sigma}_k, \hat{p}_k$ and the quadratic score becomes*

$$\hat{d}_k(x) = -\frac{1}{2}\log|\hat{\Sigma}_k| - \frac{1}{2}(x - \hat{\mu}_k)'\hat{\Sigma}_k^{-1}(x - \hat{\mu}_k) + \log \hat{p}_k . \tag{15}$$

For non-precise data, the characterizing function corresponding to the ith score becomes

$$\psi_i(t) = \begin{cases} \sup\left\{\xi(x^{(1)}, x^{2)}, \ldots, x) : x^{(j)} \in M^{(j)}, x \in M, \hat{d}_i = t\right\}, & \textit{for } \hat{d}_i^{-1}(t) \neq \varnothing \\ 0, & \textit{for } \hat{d}_i^{-1}(t) = \varnothing, \end{cases} \tag{16}$$

$\forall t \in \mathbb{R}$.

We now consider some numerical examples.

3 Classification - Example

In order to illustrate our classification rule with non-precise observations, we consider three population classes with truncated Gaussian characterizing functions. (For all x such that $\xi(x) < \alpha$ for some small α, we set $\xi(x) = 0$. This yields a finite support.) We consider samples of size $n_i = 25$ for each class $i = 1, 2, 3$, and

- We set $c_1 = 0, c_2 = 2$, and $c_3 = 3$, the "centers" of each class.
- For each non-precise observation j from class i, we set $\mu_{i,j} = c_i + U_{i,j}$, where the $U_{i,j}$ are iid uniform random variables on $[-1, 1]$. We also set $\sigma_{i,j} = V_{i,j}$, iid uniform random variables on $[0.1, 0.5]$.
- We generated a "new observation", x^*, also with a truncated Gaussian characterizing function with $\mu = 0.65$ and $\sigma = 0.05$.

We then build the characterizing functions for the D_i, the distances of the discriminants to their means under each class, as given in formula (14). This is done as follows. We set a fixed value α, and we consider the α-cuts for each observation to evaluate the minimum and maximum values taken by D_i for this value of α. We do this for several values of α to obtain a sketch of the characterizing functions for the D_i. This is illustrated in the top graphic of Figure 1 where, from left to right, we see the characterizing functions for classes 1, 2, and 3, respectively. From this plot, we see that for $\alpha > 0.78$ (roughly), x^* belongs to class 1. For $0.1 < \alpha < 0.78$, there is ambiguity between populations 1 and 2, and for $\alpha < 0.1$, there is ambiguity between all three populations. We define the values $\alpha_c(1, 2) = 0.78$ and $\alpha_c(1, 3) = 0.1$ as *critical points*. Another way to describe the classification of x^* is to say that it belongs to class 1 with confidence $1 - \alpha_c(1, 2) = 0.22$, and to class 1 or 2 with confidence $1 - \alpha_c(1, 3) = 0.9$.

Formally, looking at classes $i \neq j$, we define

$$\alpha_c(i, j) = \begin{cases} \max\{\alpha; B_\alpha(D_i) \cap B_\alpha(D_j) \neq \emptyset\}, & \text{if such } \alpha \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

the point at which the α-cuts intersect, if any. In the bottom plot of Figure 1, we compute $\alpha_c(1, 2)$ and $\alpha_c(1, 3)$ for a range of *precise* observations $x \in [0, 2]$. When we overlay the characterizing function of x^*, we clearly see the critical values 0.78 and 0.1. This approach can be generalized to any number of classes and dimensions.

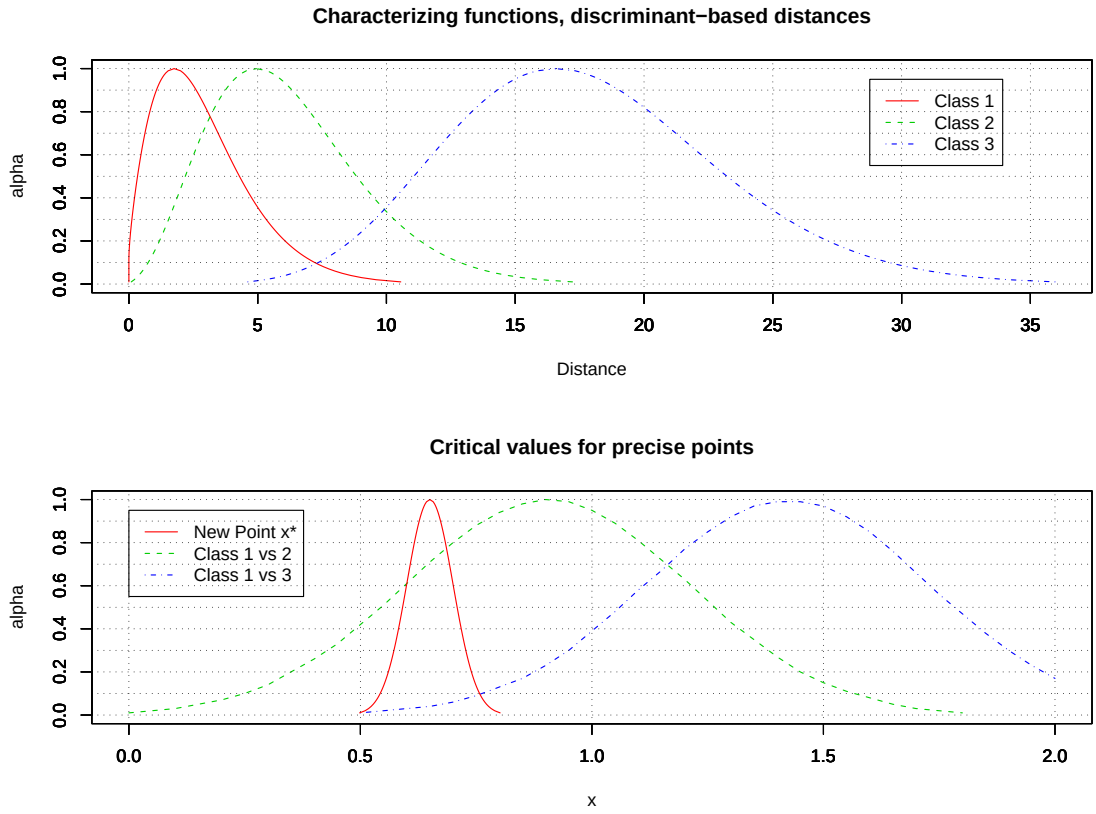

Figure 1: Nonprecise classification with 3 populations.

4 Classification - Discussion

The main difference in classification between precise and non-precise numbers lies in the interpretation of $f_i(x)$, class i probability density function (pdf) for precise quantities versus $\xi_i(x)$, class i characterizing function for non-precise quantities. For non-precise numbers, class i numbers take all values $\xi_i(x) > 0$ simultaneously, and $\xi_i(x)$ represents the *intensity* at the (precise) value x. Properties such as the area under the curve being 1 (in the continuous case) no longer holds here. This interpretation is crucial in our definitions of classification functions.

As an illustration, consider some two-dimensional objects, such as weather patterns, groups of animals, etc., in $\mathbb{R}^2$. In the simple example shown in Figure 2, we have two square-shaped objects, with respective area of 1 and 25. We assume that the density is uniform for both objects, with characterizing functions $\xi_i(x, y) = 1$, $i = 1, 2$, respectively inside the squares, and 0 elsewhere. In this case, a point (x, y) that belongs to both objects is such that $\xi_1(x, y)/\xi_2(x, y) = 1$, since both objects have the same intensity at (x, y). However, if we consider the objects via pdf's, we get $f_1(x, y)/f_2(x, y) = 25$, which gives much more weight to the smaller object.

In this section, we look at likelihood scores $S_i(x)$ for $i \in \{1, \ldots, n\}$ and x a precise value, given n characterizing functions $\xi_i(\cdot)$, $i = 1, \ldots, n$.

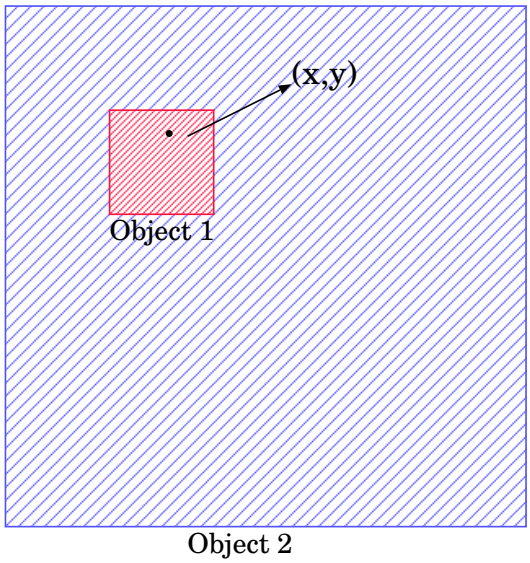

Figure 2: Three classes in two dimensions.

4.1 Two Likelihood Scores

If we deal with precise numbers, the pdf $f_i(x)$ is a measure of the likelihood of class i for observation x, and the straightforward classification strategy is to choose class i that maximizes $f_i(x)$. We can assign the following scores for each class

$$S_{f_i}(x) = \frac{f_i(x)}{\sum_j f_j(x)},$$

which are interpreted as a membership function, or fuzzy classification values. We use this definition first and consider the following scores for each class i

$$S^I_{\xi_i}(x) = \frac{\xi_i(x)}{\sum_j \xi_j(x)}. \tag{18}$$

In this case however, since the $\xi_i(\cdot)$ take values all in the same range $[0, 1]$, we can define another score based on the α-cuts.

For a given observation x, let $1 \geq \xi_1(x) \geq \xi_2(x) \geq \cdots \geq \xi_N(x) \geq 0$ without loss of generality (this is simply a re-definition of the class labels). We see that

- x belongs to no α-cut for $1 \geq \alpha > \xi_1(x)$.
- x belongs to class 1 α-cut for $\xi_1(x) \geq \alpha > \xi_2(x)$.
- x belongs to classes 1 and 2 α-cuts for $\xi_2(x) \geq \alpha > \xi_3(x)$.
- x belongs to classes 1, 2, and 3 α-cuts for $\xi_3(x) \geq \alpha > \xi_4(x)$.
- $\cdots$
- x belongs to all N α-cuts for $\xi_N(x) \geq \alpha \geq 0$.

At a given α value, the interpretation is that x could belong to all classes i such that x is in the α-cut for this class. We assign scores according to the following table.

α-range	class 1	class 2	$\cdots$	class N
$(\xi_1(x), 1]$	0	0	$\cdots$	0
$(\xi_2(x), \xi_1(x)]$	$\xi_1(x) - \xi_2(x)$	0	$\cdots$	0
$(\xi_3(x), \xi_2(x)]$	$(\xi_2(x) - \xi_3(x))/2$	$(\xi_2(x) - \xi_3(x))/2$	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[0, \xi_N(x)]$	$\xi_N(x)/N$	$\xi_N(x)/N$	$\cdots$	$\xi_N(x)/N$

Summing all entries from this table we get $\xi_1(x)$, so the score assigned to class i is given by the sum of all values in column i, divided by $\xi_1(x)$

$$S_{\xi_i}^{II}(x) = \frac{\sum_{j=i}^{N-1}(\xi_j(x) - \xi_{j+1}(x))/j + \xi_N(x)/N}{\xi_1(x)}. \tag{19}$$

Example

We consider a simple 2-class example where $\xi_1(x)$ is a triangle defined in the range $[0, 2]$ with $\xi_1(1) = 1$, while $\xi_2(x) = 1$ over the range $[0, 2]$. This is illustrated in Figure 3, along with the corresponding probability density functions.

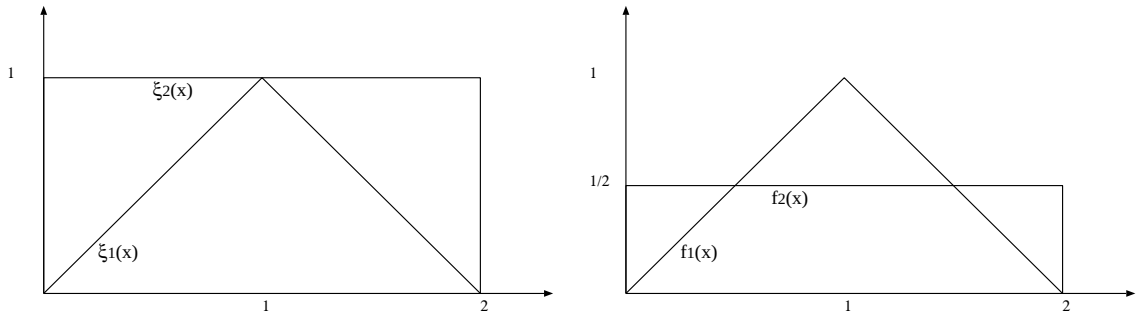

Figure 3: Two-class example.

Here are some values for the classification scores of class 1.

x	$S_{f_1}(x)$	$S_{\xi_1}^{I}(x)$	$S_{\xi_1}^{II}(x)$
0 or 2	0	0	0
1/2 or 3/2	1/2	1/3	1/4
1	2/3	1/2	1/2

We see that $S_{f_1}(1)$ differs from the other ones at $x = 1$, since it must be the case that $\int f_i(x)dx = 1$, so the triangular-shaped pdf gets more weight at $x = 1$. In the non-precise case however, observing $x = 1$ is as likely to come from either class. We also see that S_{ξ}^{I} and S_{ξ}^{II} differ, and in the next subsection, we show that $S_{\xi_1}^{II}(x)$ is always a better choice with respect to some global error criterion.

4.2 An Error Criterion

In the N-class classification problem for non-precise numbers with characterizing functions $\xi_i(x)$, $i = 1, \ldots, N$, let S_ξ be a scoring function such that $S_{\xi_i}(x) \geq 0$ is the score for class i at x, and $\sum_{i=1}^{N} S_{\xi_i}(x) = 1$, $\forall x$. We define the following error function in the continuous case

$$\Upsilon(S_\xi) = \sum_{i=1}^{N} \int_{-\Delta}^{\Delta} \xi_i(x)(1 - S_{\xi_i}(x))dx \tag{20}$$

for some large Δ. (In practice, we usually assume finite support so Δ is finite.) So for each class i, we sum all the score given to the other classes, weighted by $\xi_i(\cdot)$. In the discrete case, the integral is replaced by a summation over all cases for which $\xi_i(x) > 0$.

In the example seen previously where $\xi_1(x)$ is triangular-shaped and $\xi_2(x)$ is uniform, we get $\Upsilon(S_\xi^I) \approx 0.301$, $\Upsilon(S_\xi^{II}) \approx 0.292$, and $\Upsilon(S_f) \approx 0.338$ when using the pdf's.

For our next example, we consider two Gaussian distributions (respectively, characterizing functions) with mean and variance $(0, 1)$ for class 1 and $(1, \gamma)$ for class 2, and we look at several values for γ. In Figure 4, we plot the quantities $\Upsilon(S_\xi^I)/\Upsilon(S_f)$ (ratio I) and $\Upsilon(S_\xi^{II})/\Upsilon(S_f)$ (ratio II). When $\gamma = 1$, we see that $\Upsilon(S_\xi^I) = \Upsilon(S_f)$ as expected. Moreover, we see that $\Upsilon(S_\xi^{II}) < \Upsilon(S_\xi^I)$ for all cases considered. This is in fact a general result that we prove next.

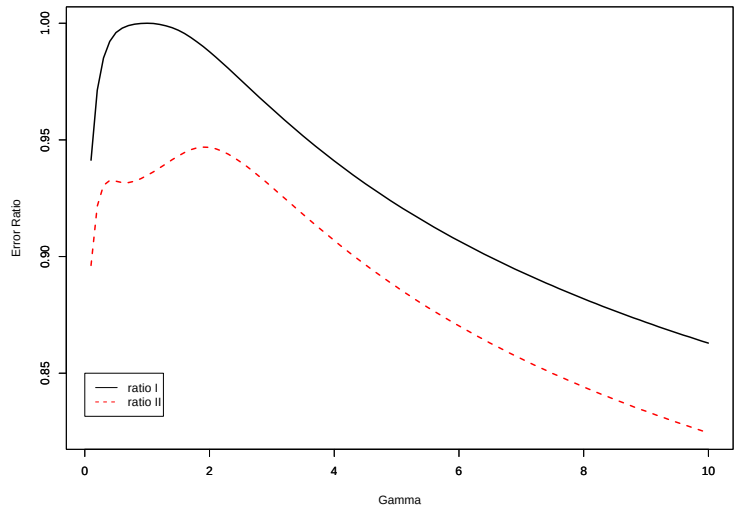

Figure 4: Gaussian example.

Theorem 10 *Given a classification problem with N non-precise classes having respective characterizing functions $\xi_i(x)$, $i = 1, \ldots, N$, then $\Upsilon(S_\xi^{II}) \leq \Upsilon(S_\xi^I)$.*

In order to prove this result, we need to show a few results in Lemmas 11 to 13, where we introduce the following (lighter) notation, given N classes

$$\Upsilon_N^I = \Upsilon(S_\xi^I)\,, \quad \Upsilon_N^{II} = \Upsilon(S_\xi^{II})\,.$$

All proofs are given in the Appendix.

Lemma 11 $\Upsilon_N^I = \Upsilon_{N-1}^I + \dfrac{\xi_N}{\sum_{i=1}^N \xi_i}\left(\sum_{i=1}^{N-1}\xi_i + \dfrac{\sum_{i=1}^{N-1}\xi_i^2}{\sum_{i=1}^{N-1}\xi_i}\right).$

Lemma 12 $\Upsilon_N^{II} = \Upsilon_{N-1}^{II} + \dfrac{1}{\xi_1}\left(\dfrac{\xi_N}{N(N-1)}\sum_{i=1}^{N-1}\xi_i + \xi_1\xi_N - \dfrac{\xi_N^2}{N}\right).$

Lemma 13 *For* $1 \geq \xi_1 \geq \xi_2 \geq \cdots \geq \xi_N \geq 0$ *with* $N > 1$,

$$\sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{k=1}^{N-1}\xi_1\xi_k(\xi_k - \xi_N) \geq \sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{k=1}^{N-1}\xi_i\xi_j(\xi_k - \xi_N)\,.$$

We illustrate the likelihood scores given in (18) and (19) with an example based on the one presented in Section 3. Here, we assume that each population is represented by a truncated Gaussian characterizing function with respective centers $\mu_1 = 0$, $\mu_2 = 2$, and $\mu_3 = 3$. We assume that all $\sigma = 1$. In Figure 5, we compute the likelihood functions for the three classes for precise values of $x \in [0, 5]$. In particular, we look at $x = 0.65$ and notice that with both measures, this point is more likely to belong to class 1, with respective scores of 0.64 and 0.75.

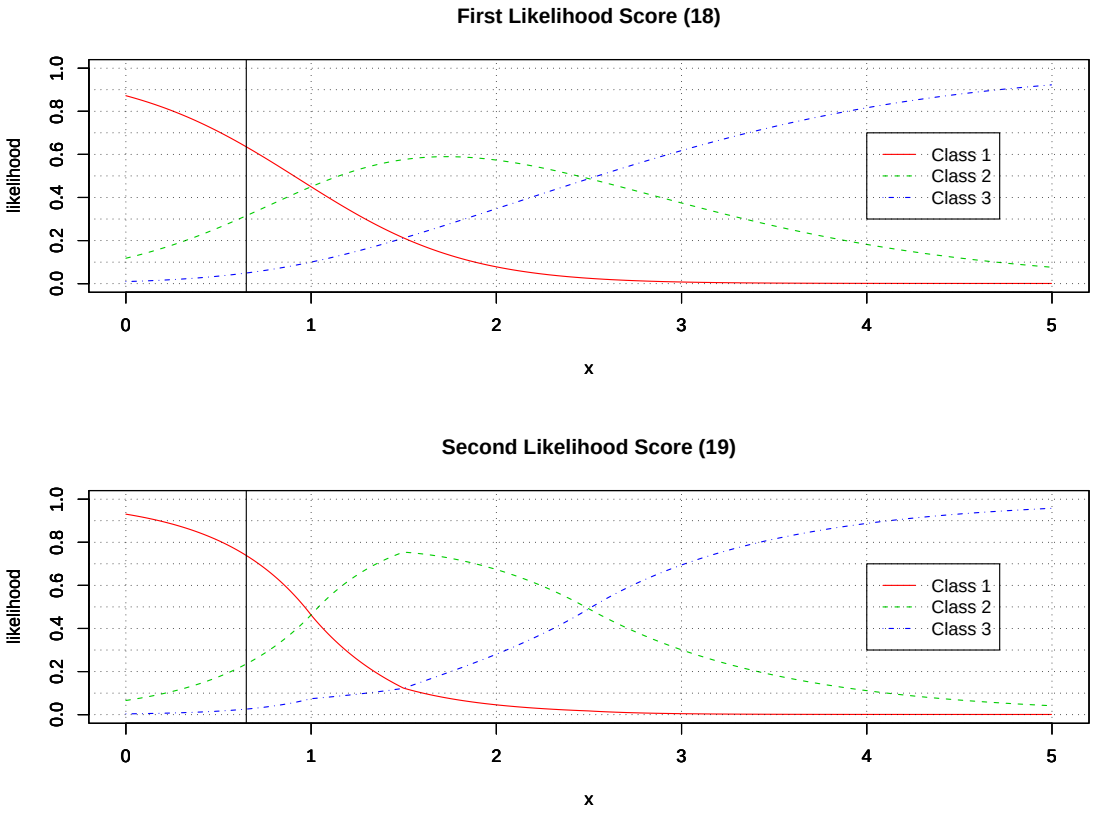

Figure 5: Likelihood with three populations.

5 Conclusion

In this paper, we presented a framework to address the problem of classification for non-precise quantities. This was achieved by writing the characterizing functions of the distances to discriminants with respect to each class of observations. The notion of critical value $\alpha_c(i,j)$ between classes i and j was also introduced. We also compared a new likelihood score to a straightforward extension from probability density functions. We showed that our likelihood score is always a better choice under some global error criterion.

Acknowledgement

The authors would like to thank Professor Reinhard Viertl for his very useful comments on a previous version of the manuscript.

References

Johnson, R., and Wichern, D. (1999). *Applied Multivariate Statistical Analysis* (4th ed.). Prentice-Hall.

Viertl, R. (1997). On statistical inference for non-precise data. *Environmetrics*, *8*, 541-568.

A Proofs

Proof of Lemma 11:

$$\begin{aligned}
\Upsilon_N^I - \Upsilon_{N-1}^I &= \xi_N - \frac{\sum_{i=1}^N \xi_i^2}{\sum_{i=1}^N \xi_i} + \frac{\sum_{i=1}^{N-1} \xi_i^2}{\sum_{i=1}^{N-1} \xi_i} \\
&= \xi_N + \frac{\sum_{i=1}^{N-1} \xi_i^2 \sum_{i=1}^N \xi_i - \sum_{i=1}^N \xi_i^2 \sum_{i=1}^{N-1} \xi_i}{\sum_{i=1}^N \xi_i \sum_{i=1}^{N-1} \xi_i} \\
&= \xi_N + \frac{\xi_N \sum_{i=1}^{N-1} \xi_i^2 - \xi_N^2 \sum_{i=1}^{N-1} \xi_i}{\sum_{i=1}^N \xi_i \sum_{i=1}^{N-1} \xi_i} \\
&= \frac{\xi_N}{\sum_{i=1}^N \xi_i} \left(\sum_{i=1}^N \xi_i + \frac{\sum_{i=1}^{N-1} \xi_i^2}{\sum_{i=1}^{N-1} \xi_i} - \xi_N \right) \\
&= \frac{\xi_N}{\sum_{i=1}^N \xi_i} \left(\sum_{i=1}^{N-1} \xi_i + \frac{\sum_{i=1}^{N-1} \xi_i^2}{\sum_{i=1}^{N-1} \xi_i} \right) . \square
\end{aligned}$$

Proof of Lemma 12: By definition, for $i < N$, we have

$$1 - S_{\xi_i}^{II}(N) = 1 - S_{\xi_i}^{II}(N-1) + \frac{\xi_N}{N(N-1)\xi_1} .$$

Therefore,

$$\begin{aligned}\Upsilon_N^{II} &= \Upsilon_{N-1}^{II} + \sum_{i=1}^{N-1} \xi_i \frac{\xi_N}{N(N-1)\xi_1} + \xi_N \left(1 - \frac{\xi_N}{N\xi_1}\right) \\ &= \Upsilon_{N-1}^{II} + \frac{1}{\xi_1}\left(\sum_{i=1}^{N-1} \frac{\xi_i \xi_N}{N(N-1)\xi_1} + \xi_1\xi_N - \frac{\xi_N^2}{N}\right) . \square\end{aligned}$$

Proof of Lemma 13: We show this result by induction on N. We let LHS and RHS represent respectively the left and right hand sides of the inequality in the lemma.

For $N = 2$, $LHS = 2\xi_1^2(\xi_1 - \xi_2)$ and $RHS = \xi_1^2(\xi_1 - \xi_2) + \xi_1\xi_2(\xi_1 - \xi_2) \leq LHS$, since $\xi_2 \leq \xi_1$.

We assume the result holds up to $N-1$, and we decompose the LHS and RHS into four terms, respectively L_i and R_i for $i = 1, \ldots, 4$.

$$\begin{aligned}LHS &= \sum_{i=2}^{N}\sum_{j=2}^{N-1}\sum_{k=2}^{N-1} \xi_1\xi_k(\xi_k - \xi_N) + \sum_{j=1}^{N-1}\sum_{k=1}^{N-1} \xi_1\xi_k(\xi_k - \xi_N) \\ &\quad + \sum_{i=2}^{N}\sum_{k=1}^{N-1} \xi_1\xi_k(\xi_k - \xi_N) + \sum_{i=2}^{N}\sum_{j=2}^{N-1} \xi_1^2(\xi_1 - \xi_N) \\ &\triangleq L_1 + L_2 + L_3 + L_4 .\end{aligned}$$

$$\begin{aligned}RHS &= \sum_{i=2}^{N}\sum_{j=2}^{N-1}\sum_{k=2}^{N-1} \xi_i\xi_j(\xi_k - \xi_N) + \xi_1 \sum_{j=1}^{N-1}\sum_{k=1}^{N-1} \xi_j(\xi_k - \xi_N) \\ &\quad + \xi_1 \sum_{i=2}^{N}\sum_{k=1}^{N-1} \xi_i(\xi_k - \xi_N) + (\xi_1 - \xi_N)\sum_{i=2}^{N}\sum_{j=2}^{N-1} \xi_i\xi_j \\ &\triangleq R_1 + R_2 + R_3 + R_4 .\end{aligned}$$

Next, we compare the terms pairwise.

1. $L_1 \geq \sum_{i=2}^{N}\sum_{j=2}^{N-1}\sum_{k=2}^{N-1} \xi_2\xi_k(\xi_k - \xi_N) \geq R_1$ from the induction hypothesis, and since $\xi_1 \geq \xi_2$.

2. $L_2 = (N-1)\xi_1 \sum_{k=1}^{N-1} \xi_k(\xi_k - \xi_N)$ and $R_2 = \xi_1 \sum_{j=1}^{N-1}\sum_{k=1}^{N-1} \xi_j(\xi_k - \xi_N)$, so $(L_2 - R_2)/\xi_1 = (N-1)\sum_{k=1}^{N-1} \xi_k^2 - \sum_{j=1}^{N-1}\sum_{k=1}^{N-1} \xi_j\xi_k \geq 0$ from Cauchy-Schwartz inequality.

3. $L_2 = L_3$ and $R_2 \geq R_3$ so $L_3 \geq R_3$.

4. $L_4 = (N-2)\xi_1^2(\xi_1 - \xi_N) \geq R_4$ since all $\xi_i \leq \xi_1$.

Thus, $LHS \geq RHS$. □

Proof of Theorem 10: For $N = 1$, we get $\Upsilon_N^I = \Upsilon_N^{II} = 0$. For $N > 1$, we write $\Upsilon_N^I = \Upsilon_{N-1}^I + \Delta_N^I$ and $\Upsilon_N^{II} = \Upsilon_{N-1}^{II} + \Delta_N^{II}$ and show that $\Delta_N^{II} \leq \Delta_N^I$. From Lemmas 11 and 12, we get

$$\begin{aligned}\frac{1}{\xi_N}(\Delta_N^I - \Delta_N^{II}) &= \frac{\sum_{i=1}^{N-1}\xi_i}{\sum_{i=1}^{N}\xi_i} + \frac{\sum_{i=1}^{N-1}\xi_i^2}{\sum_{i=1}^{N}\xi_i \sum_{i=1}^{N-1}\xi_i} - 1 - \frac{\sum_{i=1}^{N-1}\xi_i}{\xi_1 N(N-1)} + \frac{\xi_N}{N\xi_1} \\ &= \frac{-\xi_N}{\sum_{i=1}^{N}\xi_i} + \frac{\sum_{i=1}^{N-1}\xi_i^2}{\sum_{i=1}^{N}\xi_i \sum_{i=1}^{N-1}\xi_i} - \frac{\sum_{i=1}^{N-1}\xi_i}{\xi_1 N(N-1)} + \frac{\xi_N}{N\xi_1}\,.\end{aligned}$$

To show this is non-negative is equivalent to showing that

$$\frac{\sum_{i=1}^{N-1}\xi_i}{\xi_1 N(N-1)} - \frac{\xi_N}{N\xi_1} \leq \frac{\sum_{i=1}^{N-1}\xi_i^2 - \xi_N \sum_{i=1}^{N}\xi_i}{\sum_{i=1}^{N}\xi_i \sum_{i=1}^{N-1}\xi_i}\,,$$

which we can write as

$$\left(\sum_{i=1}^{N}\xi_i\right)\left(\sum_{i=1}^{N-1}\xi_i\right)\left(\sum_{i=1}^{N-1}(\xi_i - \xi_N)\right) \leq N(N-1)\xi_1 \sum_{i=1}^{N-1}\xi_i(\xi_i - \xi_N)\,,$$

which was shown in Lemma 13. □

Author's address:

Prof. Mayer Alvo
Prof. François Théberge

Department of Mathematics and Statistics
University of Ottawa
585 King Edward
Ottawa, ON, K1N 6N5
Canada

E-mail: theberge@ieee.org