European Journal of Operational Research 251 (2016) 930–937

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Computational Intelligence and Information Management

Measures of dispersion for multidimensional data

Adam Kołacz[a], Przemysław Grzegorzewski[a,b,*]

[a] Faculty of Mathematics and Computer Science, Warsaw University of Technology, Koszykowa 75, Warsaw 00–662, Poland
[b] Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw 01–447, Poland

ARTICLE INFO

Article history:
Received 22 February 2015
Accepted 4 January 2016
Available online 11 January 2016

Keywords:
Descriptive statistics
Dispersion
Interquartile range
Multidistance
Spread

ABSTRACT

We propose an axiomatic definition of a dispersion measure that could be applied for any finite sample of k-dimensional real observations. Next we introduce a taxonomy of the dispersion measures based on the possible behavior of these measures with respect to new upcoming observations. This way we get two classes of unstable and absorptive dispersion measures. We examine their properties and illustrate them by examples. We also consider a relationship between multidimensional dispersion measures and multidistances. Moreover, we examine new interesting properties of some well-known dispersion measures for one-dimensional data like the interquartile range and a sample variance.

1. Introduction

Various summary statistics are always applied wherever decisions are based on sample data. The main goal of those characteristics is to deliver a synthetic information on basic features of a data set under study. It seems that the most commonly used summary statistics are central tendency measures (like the mean, median, mode, etc.) indicating a typical behavior of the examined variable. However, no measure of central tendency can reveal the whole picture of a variable. Indeed, two or more samples may have the same mean (or other central tendency) although they differ significantly. Therefore, besides central tendency a dispersion of observations in a sample is also of interest. Moreover, in many cases we have to monitor variability as carefully as the location parameters. As a typical example let us consider the Statistical Process Control where no alarm signal found on the $\bar{X}$-chart cannot be automatically interpreted as the process is under control until the S-chart (or R-chart) confirms no alarm caused by the increase of variability.

Many tools have been proposed to characterize dispersion, like the range, interquartile range, sample variance, standard deviation an so on. They differ in construction, properties and situations they are intended for use. It is also worth mentioning that several terms are used in the literature as regards dispersion measures like measures of variability, scatter, spread or scale. Some authors reserve the notion of the dispersion measure only to those cases when variability is considered relative to a given fixed point (like a sample variance which averages squared deviation of the data points from their mean) and then use the term spread as a more general one (see Bickel & Lehmann (1976, 1979); Wilcox (2005)). However, such distinction in terminology is neither consistent nor commonly accepted. Thus in our paper we do not attach importance to such distinctions.

Some of the considered tools measure the absolute spread (like those mentioned before), while the other indicate the relative scatter (e.g. the coefficient of variation or Gini coefficient). Most of them are dedicated to quantitative data (ratio scale) but one can found also a few that might be used to characterize qualitative observations (nominal scale).

What is interesting is that almost all well-known measures of dispersion could be used only for one-dimensional data. It is rather inconvenient especially that most of the contemporary data sets available and processed in practice is multidimensional. Of course, having such multidimensional data set one may apply univariate dispersion measures to each variable separately, but this way we loose information on possible relations between variables. Then, as a possible remedium, one may consider e.g. a covariance matrix which delivers both variances of all single variables and covariances for all pairs of variables. Hence, having a data set of k-dimensional observations we get a matrix of k^2 numbers instead of a single real value of a desired measure of

* Corresponding author at: Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01–447 Warsaw, Poland. Tel.: +48 223810207; fax: +48 223810105.
E-mail addresses: A.Kolacz@mini.pw.edu.pl (A. Kołacz), pgrzeg@ibspan.waw.pl (P. Grzegorzewski).

http://dx.doi.org/10.1016/j.ejor.2016.01.011
0377-2217/

dispersion characterizing somehow the whole multidimensional sample.

Keeping in mind all the remarks mentioned above we propose a general definition of a dispersion measure that could be applied for any finite sample of k-dimensional real observations, i.e. $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$. Next we examine basic properties of so defined measures and illustrate them by examples. We also consider the relationship between multidimensional dispersion measures and multidistances introduced by Martín and Mayor (2009, 2011).

Recently, Gagolewski (2015) considered the dispersion measures from the aggregation theory point of view. He showed that although aggregation theory mainly focuses on central tendency measures (see Beliakov, Pradera, & Calvo (2007); Calvo, Mayor, and Mesiar (2002); Grabisch, Marichal, Mesiar, and Pap (2009)), it may deliver an interesting insight to measures of spread of one-dimensional quantitative data. In our case we show that some considerations on general multidimensional dispersion measures may also lead to some interesting conclusions for one-dimensional data sets.

The paper is organized as follows: In Section 2 we present the desired requirements each measure of dispersion should satisfy. Next, we distinguish two basic types of dispersion measures: unstable and absorptive dispersion measures (Section 3 and Section 4, respectively). Section 5 is devoted to some interesting properties of the interquartile range that appear in practice when we try to estimate it from data. In Section 6 we prove a theorem showing a relation between unstable and absorptive dispersion measures. Finally, in Section 7 we examine the relationship between dispersion measures and multidistances.

2. Dispersion measures

Consider a sample of n observations from the k-dimensional real space, i.e. $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$. Descriptive statistics, also called summary statistics, provide various measured describing different aspects of the underlying data. Besides central tendency measures, the next group of the most useful summary statistics is formed by measures of dispersion. Although each person has some intuition about measures of dispersion, it seems that a formal definition would be desirable.

Definition 2.1. A function $\Delta : \bigcup_{n=1}^{\infty} \left(\mathbb{R}^k\right)^n \to [0, \infty)$ is called a **measure of dispersion** if Δ is not identically zero function which satisfies the following axioms for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$:

(A1) $\Delta(\mathbf{x}, \ldots, \mathbf{x}) = 0$

(A2) Δ is symmetric, i.e.

$$\Delta(\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(n)}) = \Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

for any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$,

(A3) Δ is translation invariant, i.e.

$$\Delta(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a}) = \Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

for any $\mathbf{a} \in \mathbb{R}^k$,

(A4) Δ is rotation invariant, i.e.

$$\Delta(\mathbf{R}\mathbf{x}_1, \ldots, \mathbf{R}\mathbf{x}_n) = \Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

for any rotation matrix $\mathbb{R}^k$ (i.e. $\mathbf{R}$ is an orthogonal matrix and such that $det\ \mathbf{R} = 1$).

Sometimes one more axiom is also considered:

(A5) there exists a function $\rho : \mathbb{R} \to [0, \infty)$ such that

$$\Delta(a\mathbf{x}_1, \ldots, a\mathbf{x}_n) = \rho(a)\Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

for $a \in \mathbb{R}^+$.

Usually adding another observation to a data set under study we expect changes in the dispersion measure value, no matter where the new point is located. However, there also exists a class of measures for which by adding new observations we do not change the scatter of the data set (provided those observations belong to some area). To clarify the situation in further sections we indicate two important subfamilies of dispersion measures.

3. Unstable dispersion measures

Definition 3.1. A measure of dispersion $\Delta : \bigcup_{n=1}^{\infty} \left(\mathbb{R}^k\right)^n \to [0, \infty)$ is called **unstable** if

$$\Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}) \neq \Delta(\mathbf{x}_1, \ldots, \mathbf{x}_n), \tag{1}$$

for almost all $\mathbf{x}_{n+1} \in \mathbb{R}^k$.

In other words, for any unstable dispersion measure and any data set there exist a set which has the k-dimensional Lebesgue measure zero and such that joining any its point to the data set do not change a value of the dispersion obtained for the initial data set. Let us now discuss some examples and basic properties of the unstable dispersion measures.

Example 3.2. A one-dimensional sample, i.e. $x_1, \ldots, x_n \in \mathbb{R}$ provides many examples of well-known unstable dispersion measures, like different sample variances: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $s_b^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ or corresponding sample standard deviation.

Example 3.3. Having a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ let us define the following function

$$G_e(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_e^2(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

where $d_e(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance in $\mathbb{R}^k$. It can be shown that (2) is an unstable dispersion measure. To prove it, let us firstly assume that $A_m = \sum_{i=1}^{n}(x_i^m - \bar{x}^m)^2$, where x_i^m denotes the mth component of $\mathbf{x}_i$ and $\bar{x}^m = \frac{1}{n}\sum_{i=1}^{n} x_i^m$. Then for any j we get

$$\begin{aligned} A_m &= \sum_{i=1}^{n}\left(x_i^m - x_j^m + x_j^m - \bar{x}^m\right)^2 \\ &= \sum_{i=1}^{n}\left(x_i^m - x_j^m\right)^2 + 2\sum_{i=1}^{n}\left(x_i^m - x_j^m\right)\left(x_j^m - \bar{x}^m\right) \\ &= \sum_{i=1}^{n}\left(x_i^m - x_j^m\right)^2 - n\left(x_j^m - \bar{x}^m\right)^2. \end{aligned}$$

Summing up both sides over j we get $nA_m = \sum_{i,j=1}^{n}(x_i^m - x_j^m)^2 - nA_m$, which implies that

$$A_m = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^m - x_j^m\right)^2 = \frac{1}{n}\sum_{1 \leqslant i < j \leqslant n}\left(x_i^m - x_j^m\right)^2$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_e^2(\mathbf{x}_i, \mathbf{x}_j) = n\sum_{m=1}^{k} A_m = n(n-1)\sum_{m=1}^{k} S_m^2,$$

where S_m^2 denotes the sample variance of $x_1^m, \ldots, x_n^m$. It is clear that a linear combination of unstable measures is also an unstable dispersion measure.

Lemma 3.4. *Let $\Delta_1, \ldots, \Delta_m$ denote unstable dispersion measures such that for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$ a set $\{\mathbf{x}_i : \Delta_j = 0\}$ has a zero Lebesgue measure and let f: $[0, \infty)^m \to [0, \infty)$ be a function which is not constant, is continuous and $f(\mathbf{0}) = 0$. Then $\Delta' = f(\Delta_1, \ldots, \Delta_m)$ is also an unstable dispersion measure.*