Computers & Graphics ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Computers & Graphics

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

Special Section on SIBGRAPI 2016

Depth functions as a quality measure and for steering multidimensional projections

Douglas Cedrim [a], Viktor Vad [b], Afonso Paiva [a], M. Eduard Gröller [b,c], Luis Gustavo Nonato [a], Antonio Castelo [a]

[a] ICMC–USP, São Carlos, Brazil
[b] TU Wien, Austria
[c] VRVIS Research Center, Austria

ARTICLE INFO

Article history:
Received 29 February 2016
Received in revised form
23 August 2016
Accepted 23 August 2016

Keywords:
Visual analytics
Depth functions
Non-parametric statistics
Dimensionality reduction
Quality measures

ABSTRACT

The analysis of multidimensional data has been a topic of continuous research for many years. This type of data can be found in several different areas of science. A common task while analyzing such data is to investigate patterns by interacting with spatializations of the data in a visual domain. Understanding the relation between the underlying dataset characteristics and the technique used to provide its visual representation is of fundamental importance since it can provide a better intuition on what to expect from the spatialization. In this paper, we propose the usage of concepts from non-parametric statistics, namely depth functions, as a quality measure for spatializations. We evaluate the action of multidimensional projection techniques on such estimates. We apply both qualitative and quantitative analyses on four different multidimensional techniques selected according to the properties they aim to preserve. We evaluate them with datasets of different characteristics: synthetic, real world, high dimensional; and contaminated with outliers. As a straightforward application, we propose to use depth information to guide multidimensional projection techniques which rely on interaction through control point selection and positioning. Even for techniques which do not intend to preserve any centrality measure, interesting results can be achieved by separating regions possibly contaminated with outliers.

1. Introduction

The importance of data analysis has grown tremendously in the last years. It has become a challenging task for many different reasons. First of all nowadays *data sources are ubiquitous* and are available *for a broader audience*, one of the reasons is that high-definition sensors have become less expensive (e.g., high-definition cameras, 3D scanners). Secondly, the *scalability* since the volume of user generated data in a small scale of time (e.g., hours) can easily achieve the range of gigabytes (or terabytes). At last but not least, the *complexity* of the data itself is also an important aspect to be taken into account. Some examples of such data are collection of images [1] and textual data [2], computational simulations [3], gene data [4] and so on.

While dealing with multidimensional, possibly high dimensional, data there have been many efforts in the visualization on providing techniques and tools to allow for data analysis. Common approaches are visual exploration through linked views [3,5,6] and multidimensional projections [1,7]. In order to improve the effectiveness of the analysis, *quality metrics* can be defined to aid a particular visualization, which allows for quantifying how much a particular visual design conveys relevant patterns of the data (e.g., cluster information). The works by Bertini et al. [8] and more recently Sedlmair and Aupetit [9] provide a comprehensive and systematic analysis of such quality metrics.

On the other hand, a common approach for data analysis comes from machine learning, which automatically generates hypotheses from the input data [10]. These hypotheses have many different uses. Commonly they delimit meaningful regions of the input space (i.e., identifying cluster regions) by defining decision boundaries between data from different categories or classes. Although algorithms that perform such operations in an automatic fashion have proven quite useful, they rely on the definition of a proper similarity metric among all pairs of the input data.

The idea of *proper* can vary with the type of data being analyzed. It is rather complex to define feature extractors, which are discriminative enough for a broad class of datasets. This is mainly due for two reasons: the *diverse* sources of the data (e.g., collection of images vs car engine design [3]) and because of *ambiguities*.

Multidimensional scaling techniques play a particularly important role in the context of such general datasets. They decrease the complexity of the analysis by reducing the dimensionality of the data. This enables, for instance, projecting the input data onto a visual space aiming to preserve constraints

http://dx.doi.org/10.1016/j.cag.2016.08.008
0097-8493/

(e.g., pairwise distances) as much as possible. Scatter plots have been used to effectively convey absolute and relative distances between points projected into the visual space, and to allow for interactions within exploratory environments [11].

Reasoning on spatializations through scatter plots involves two different aspects: the definition of objective quality measures (e.g., distance preservation) as well as subjective ones (e.g., user-specific metrics). One of the main challenges in this process is the possible mismatch between the objective measures and the user-made quality judgments [12,13]. As the human perception is highly based on pattern finding, the design of visual metaphors should provide ways of handling both aspects without much effort. In the literature, there are some works aiming to perform a quantitative analysis of different patterns that can be found on these spatializations. For example they analyze separation factors of clusters [12] or more general graph-based measures [14]. Wilkinson et al. [14] evaluate the presence of outliers as one important property to characterize scatter plots. They propose to use a minimum spanning tree to quantify the appearance of outliers, defined for the whole scatter plot rather than performing a point-wise analysis. Moreover, *outlier detection* itself is an active interdisciplinary area of research [15].

On the other hand, scalar fields defined over the data are interesting, because they allow for various types of analysis. For instance, finding patterns in different dimensions of the data can be handled by a visual inspection of the scalar field coordinate by coordinate. Also the scalar field can be analyzed quantitatively by using topological tools, such as persistent homology [16].

Data depth is a particular interesting scalar field which comes from order statistics and non-parametric multivariate statistical analysis. In order statistics no – or as few as possible – assumptions from the underlying data distribution are made beforehand [17,18]. It is tightly related to multivariate median estimation, since the latter does not have a single generalization from the unidimensional situation. In the multivariate setting, different generalizations for the median are provided by data depth functions [17].

Data depth functions convey the notion of centrality concerning the data. They also relate to methods of *extreme value analysis* in the outlier detection literature, as outliers can be seen as the least central points in the data. Such points might contain useful information about the data such as an abnormal behavior during their acquisition or synthesis; data with different underlying distributions mixed together, and abnormal patterns introduced while processing the data (e.g., multidimensional projection) [19].

Once the data depth distribution is calculated before and after some processing on the data (e.g., multidimensional projection), a global analysis through statistical tests could be done. Alternatively, a visual analysis provides a qualitative way of understanding how the depth distribution is given on the input space and how it has changed for individual data points. This also allows for including user knowledge in the process, since the way users perceive centrality on scatter plots could be taken into account, although this is out of the scope of this work.

Other statistics measures such as mean and variance, although widely used, can lead to misleading interpretations of the data distribution since they are easily influenced by outliers and also by non-symmetric data distributions [17].

The usage of the median as a more robust location estimation is considered an interesting alternative. It has an asymptotic *breakdown point* of 0.5, which is a robustness estimator. Only if half of the data were modified the location estimation would become completely corrupted [20]. To put this into perspective, the mean has a breakdown point of 0, i.e., a single outlier can completely modify the estimation. Robust statistical estimates have been shown to be successfully applied on different research areas outside statistics, such as image and geometry processing [21,22].

1.1. Paper outline and contributions

The paper is structured as follows: in Section 2 we discuss the literature of quality measures for multidimensional projections and also some approaches using statistics for multivariate analysis. In Section 3 depth functions are described more formally and the choice for a particular one is motivated. In Section 4 its use as a quality measure for multidimensional projections is discussed with some quantitative and qualitative experiments. Moreover a comparison with another quality measure is performed. In Section 5, we describe how data depth can be applied for control point selection and we explore some strategies for steering multidimensional projections. At last, we point out limitations, in Section 6.

Taking what has been previously described into consideration, the main contributions of this paper are:

- Using order statistics as a quality measure for spatializations of multidimensional data.
- A qualitative analysis tailored for visually inspecting candidate regions of multidimensional outliers.
- A quantitative analysis through a quality metric defined by individual point – data depth – variations.
- A framework for steering multidimensional projection techniques by different sampling strategies using data depth information.

2. Related work

One of the main challenges in information visualization is to increase the suitability of multidimensional data representation for data analysts [23]. Within this context, several quality measures have been proposed, in order to evaluate patterns in multidimensional data visualization. A common evaluation characteristic is the ability of the quality measure to identify clusters and relate that to how humans perceive scatter plots. For a comprehensive review of such quality measures, we refer to Albuquerque et al. [24] and to Tatu's thesis [13]. However, the proposed pipeline by Tatu [13] is rather different from ours, since it uses quality measures, selected by the user, to steer the multidimensional projection. Afterwards, appropriate dimensions are selected where the data is then projected onto.

Different lines of investigation evaluate spatializations by how much specific quantities are preserved after the projection procedure [25]. Etemadpour et al. [26] introduce the concept of *density-based motion* in order to evaluate the point density of clusters that can be lost when a multidimensional dataset is projected.

Three common categories of methods for measuring distortions are distance-based, topology-based (neighborhood), and perception-based methods.

Some distance-based approaches are not scale-invariant (i.e., standard stress measures), meaning that even identical spatializations might indicate totally different preservation situations. Moreover, even after a normalization procedure, datasets with outliers introduce a bias on the analysis. By definition, the distance of outlying points to non-outlying points is high. This affects how variations among non-outlying points are perceived, since they contribute less to the distance deviation measure [7].

Topological approaches mainly investigate the mismatch between neighborhoods on the input space and neighborhoods on the visual space [27,28]. This measure might be too strict, since small perturbations on the neighborhood of a point might considerably affect its neighborhood topology, although points still remain close. Some works try to address this problem.

Please cite this article as: Cedrim D, et al. Depth functions as a quality measure and for steering multidimensional projections. Comput Graph (2016), http://dx.doi.org/10.1016/j.cag.2016.08.008