Applied Mathematical Sciences, Vol. 10, 2016, no. 20, 959 - 971
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2016.6121

Three-Dimensional Representation of a Multidimensional Data Set

Miguel Diogenes Matrakas

Graduate Program of Nuimerical Methods in Engineering (PPGMNE)
Paraná Federal University (UFPR) - Brazil

Sérgio Scheer

Graduate Program of Nuimerical Methods in Engineering (PPGMNE)
Paraná Federal University (UFPR) - Brazil

Abstract

This study aims to describe a technique for performing a visualization of a multi-dimensional data volume such that the entire dataset is represented in the image, not just a subset of the dimensions that are part of the data. The results indicate that it is possible to perform the visualization using a dimensional reduction technique to create a code for the data.

Keywords: Computer Vision, Scientific Visualization, Dimensional Reduction

1 Introduction

A large portion of the actual processes or phenomena that are studied have multidimensional data, that is, have a large set of distinctive features. In order to analysts can understand and draw conclusions from this data, it is necessary to create representations or projections thereof in two or three dimensions. Ideally, projections should use the maximum possible dimensions or characteristics, to represent the original set, which is not always possible as the used devices typically represent only two or three dimensions, beyond what, according Wright [14], understanding of a set with more than three dimensions is quite difficult [13].

Multidimensional sets are the result of processes in different areas, such as files containing digital audio or digital image signals, the results of a MRI scan, DNA sequences, among many others. To facilitate viewing sets with such characteristics, it is necessary to adjust the number of dimensions of data to the device capabilities, and for that, there are existing techniques that perform a Dimensional Reduction (DR) on the dataset. The classification, understanding and visualization of multi-dimensional sets are some of the processes facilitated by applying the DR methods [13].

This paper presents the characteristics of some of the Dimensional Reduction algorithms, along with a taxonomy of criteria for evaluating their results. In a second step, using one of these DR techniques to map a n-dimensional set such that a representation containing all the features present in the input data can be displayed composing an interactive solution for data analysis.

This paper is organized as follows. The next section presents the concepts and short descriptions about some Dimensional Reduction (DR) methods. In the sequence, it is described some aspects of Scientific Visualization with the characteristics of imaging methods used to represent data volumes. In the *Visualization of a n-dimensional data volume* section is presented a description of a method for treating a volume of n-dimensional data and generating a graphical representation. In the last section are the final considerations about the methods and algorithms discussed in the paper.

2 Dimensional Reduction

The DR procedure is to map the elements of a set with n dimensions for a representation that maintains in the best possible way, the relationships between elements and their groupings in a set with m dimensions, with $m << n$ [13, 1]. Therefore, for a set of h elements $X^n = \{x_i \in \mathbb{R}^n\}_{1 \leq i \leq h}$, an DR algorithm can be interpreted as a function defined as

$$f : \mathbb{R}^n \times T \to \mathbb{R}^m \tag{1}$$

that maps each x_i elements into a new element y_i in space $\mathbb{R}^m$ [8].

A multidimensional set after performing the size reduction should keep the neighborhood relations between the vectors, i.e., a set of points near the n-dimensional space must also form a set of neighbors in the projection data. Each of the DR methods presents a peculiarity regarding the disposal of the vicinity of the projected points [8].

2.1 Dimensional Reduction Methods

There are dozens of methods used to make DR in the literature, classified according to the algorithm used to calculate the function f, presented in equation 1. Some of the most used methods are the Classical Scaling or Multidimensional Scaling (MDS), Principal Components Analysis (PCA), Isomap, Maximum Variance Unfolding (MVU), Locally Linear Embedding (LLE), Stochastic Neighbor Embedding (SNE), Stochastic Proximity Embedding (SPE) and several configurations and models of Artificial Neural Networks. This list does not claim to be complete or qualify the methods, but present a set of the most cited, to expose the diversity of approaches to the DR problem. Some of these methods are briefly described below:

Classical Scaling (MDS) according to Borg [2], consists of an array of elements X in the n-dimensional space, calculating a Δ^2 matrix containing the squares of these elements dissimilarities to then apply the operation called double centering consisting in calculating the matrix B_Δ given by

$$B_\Delta = -\frac{1}{2}\, J\, \Delta^2\, J$$

J being the centralization matrix given by: $J = I \,-\, n^{-1}\, U$, where I is the identity matrix, U is a matrix whose elements are equal to 1 and n is the number of dimensions of the original set.

Δ^2 matrix must be decomposed into its eigenvalues and eigenvectors, so that:

$$B_\Delta = Q\,\Lambda\,Q' = (Q\Lambda^{1/2})(Q\Lambda^{1/2})' = YY'$$

After the decomposition, it is considered the matrix formed by the first m eigenvalues greater than zero, called Λ_+ and Q_+ is a matrix formed by the first m columns of Q, making the resulting array of coordinates to be:

$$Y = Q_+\Lambda_+^{1/2}$$

.

This method minimizes the loss function given by

$$L(Y) = \|YY' - B_\Delta\|^2$$

SMACOF it is an algorithm to minimize the stress function, whose acronym means “Scaling by Mojorizing a Complicated Function”, as described in [2]. This algorithm solves the Multidimensional Scaling by an iterative process. Therefore, from an array X of elements in the n-dimensional space and the Δ matrix which is formed by the dissimilarity of these elements, the stress function represents the difference between the measures of the dissimilarities represented in the Δ matrix and the values of distance between the projections of the elements of X in m-dimensional space

The stress function is written as:

$$\sigma_r(Y) = \sum_{i<j} w_{i\,j}(\delta_{i\,j} - d_{i\,j}(Y))^2 \tag{2}$$

where w is a weight matrix, $\delta_{i\,j}$ are the elements of the dissimilarity matrix, $d_{i\,j}$ are the elements of the matrix formed by the distances between the elements of Y, which is the matrix formed by the projection of the X matrix.

The algorithm consists of, from an initial, nonrandom projection Y, calculate the difference between the distances using the Stress function, and while its value is greater than a precision limit, or a maximum of iterations is not reached, refresh the Y matrix using $Y^u = n^{-1}B(Y)Y$ if the weight matrix has all elements equal to 1, or $Y^u = V^+B(Y)Y$ otherwise. The matrix V^+ is the inverse of the weighted sums of the distances between the elements of Y, and $B(Y)$ is the matrix formed by the weighted ratio of the dissimilarity of the elements of X and Y.

Principal Components Analysis (PCA) is described by Van Der Maaten [13] as being mathematically equivalent to Classical scaling method, since in both methods the aim is to minimize the loss function, or search a representation for the data in which the difference value for a particular distance measure applied between the pairs of elements, is as low as possible. For the Classical Scaling it is used the Euclidean distance and for PCA it is used the covariance matrix.

In the same manner as in MDS, PCA solve an eigenvalue decomposition as follows $cov(X)M = \lambda M$ where M is the matrix that maps the elements of the n-dimensional space to the m-dimensional space, and λ is the matrix formed by eigenvalues of $cov(X)$.

Isomap Method proposed by Tenenbaum [12] that works with the Geodetic distances, not the Euclidean distances between the points of the set of elements in n-dimensional space. Its goal is to capture the points' distribution geometry in the n-dimensional set and keep it on the projection.

The DR is carried out using the Classical Scaling, but to work with the geodesic distances it is necessary to calculate for each element its k nearest neighbors, thus forming a connected graph, and from this, the shortest path between each pair of elements corresponds to their distance, which can be calculated by algorithms such as Djikstra's or Floyd's shortest path [13].

Stochastic Proximity Embedding (SPE) It is described in the work of Najim [9] as a nonlinear, iterative method which consists in updating the projections of each element taking into account their distances to other components of the set that are smaller than a radius r_C.

From an initial random projection a point i is drawn and used to adjust the coordinates of all the other elements of the projection, using the following rule:

$$y_j = y_j + \lambda(t_k)\, S(\delta_{i\,j})\, \frac{\delta_{i\,j} - d_{i\,j}}{d_{i\,j} + \epsilon}\, (y_j - y_i) \tag{3}$$

$$S(\delta_{i\,j}) = \begin{cases} 1 & se(\,(\delta_{i\,j} \leq r_C) \wedge (\,(\delta_{i\,j} > r_C) \vee (d_{i\,j} < \delta_{i\,j})\,)\,) \\ 0 & \text{caso contrrio} \end{cases} \tag{4}$$

where x_i and x_j represent coordinates of i and j elements, $\lambda(t_k)$ is the learning rate, ϵ is a constant to prevent division by 0, $\delta_{i\,j}$ is the distance between the elements in the n-dimensional space, and $d_{i\,j}$ is the projections distance in m-dimensional space.

3 Scientific Visualization

Wright defines Visualization as an interactive process to understand what generated or produced the data, and not just a technical presentation of these data [14]. It also states that human beings naturally understands three dimensions, hence the understanding of spaces with more dimensions, except for the special case of time is limited. Therefore, if it is necessary to represent more variables than can be accommodated with these restrictions, other resources should be used, such as colors, sounds, animation, or whatever else is available.

The dimensional reduction, accomplished by methods such as those discussed in section 2, is also part of the toolkit available for viewing and analyzing data in which the number of dimensions exceeds the capacity of human understanding or representation of a particular device. To illustrate this situation it is represented in Figure 1 an artificially generated three-dimensional block comprising three scalar fields. Each scalar field represents an existing variable in the data set with different variation ranges and rates. In Figure 1a one realizes that the values vary in only one axis, the variation of yellow color to red, as in Figure 1b the variation is in another direction and with a greater range of values represented in the 1a. In Figure 1c besides the change on the

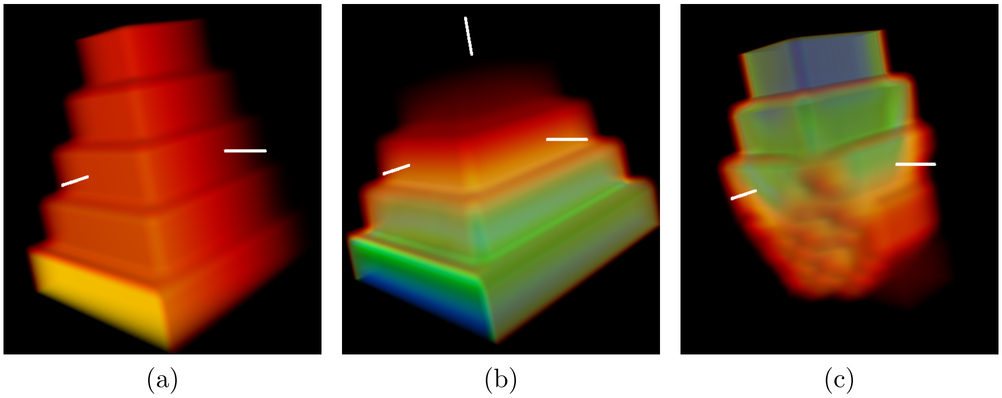


Figure 1: Visualization of three scalar fields that form a volume corresponding to the representation of an artificially generated block

variations direction, part of the block is not represented, indicating that there there is no values for the scalar being displayed.

In this line, the authors Pao and Meng [11] address the problems of getting to understand a set of multidimensional and multivariate data, presenting as main tool the DR methods, allowing data to be viewed in 2D charts (projections), which allows analysts to understand more easily the relationships in the data. According to the authors there are three aspects in the understanding of multidimensional data:

Distribution of n-dimensional points Knowing how the data points occupy the space by answering questions such as: Data distribution is uniform or in clusters? It follows the same distribution throughout the space, or is regular in a region and irregular in another?

Functional relationship Whether there is a match between the values of the vector field of the input space and the space of the property values.

Categories creation Clusters creation in the properties space. How the points in the data space relate to the categories? Elements close in the data space correspond to the same category in the properties space? Inconsistencies are studied.

Dos Santos and Brodlie perform the visualization of multidimensional and multivariate data using filters to select the set of dimensions to be displayed, the described tool gives good results for navigation in large data sets, however viewing occurs separately, with a set of dimensions being shown at each moment, which are chosen by the user [3].

The authors Guo, Xiao and Yuan describe an iterative method for defining a Transfer Function used to generate images from multidimensional data [6].

The transfer function is created with the help of data visualization in a diagram of parallel coordinates and projections of the data clusters resulting from the MDS method.

In another line of research Lawrence et al. describes a multi-dimensional image display method which unifies the multiple bands of a multispectral image in the visible band with a high degree of realism and maintaining the distribution characteristics of the original data values showing reduction examples from 4D, 8D and 31D spaces to the 3D space with variations in the source data that reach 1000:1 [7]. The objective is to develop a method to map an image with a large number of bands of scalar values in an image with few dimensions, or bands. Simply put, the purpose of this mapping is to generate an image with few dimensions whose scalar values are consistent with the original values in order to preserve (within the possibilities) the relative distances between the magnitudes present in the input image pixels pairs and "multidimensional data visualization is crucial to analyze these data sets".

3.1 Volume Visualization

According to Engel [4], generating graphical representations of volumes require that the participant medium is modeled with the light energy transport mechanism and both the representation of gaseous phenomena as scientific visualization of volume data share the same mechanism of light energy propagation.

In the model used to perform the volume rendering, it is assumed that light travels in straight lines if there is no interaction with the medium. According to Engel et al. and Glassner [4, 5], the three main types of interactions that can occur between a ray of light and the medium through which it is propagating are:

Emission Case in which the material emits light effectively, thus increasing the amount of energy that propagates in the medium.

Absorption Occurs when the material through which the light beam is traveling can convert radioactive energy into heat, effectively decreasing the amount of light energy.

Scattering It is a situation in which the direction of the light path is changed by the medium that it is going through. This change can be *elastic*, in which case the wavelength of the light radiation is not altered by the change of direction, or *inelastic*, when there is a change in the wavelength of light. The scattering can also be addictive, when the direction of another light beam is changed and coincides with the direction of the current beam, increasing it's energy, or subtractive, when part of the energy of the light beam is deflected in another direction.

Also according to Engel et al. [4], the energy of a beam of light can be described by its radiance I, which is defined by the amount of radiative energy Q per unit area A, which is measured as projected along the direction of light indicated by $\perp$, per solid angle $Omega$ and per unit time t:

$$I = \frac{dQ}{dA_{\perp}\, d\Omega\, dt} \tag{5}$$

The traditional method for visualizing data from simulations is the generation of a matrix with the data and then perform the visualization using traditional methods based on linear interpolation. In his work, Nelson proposes a method for performing the visualization of such data accurately and interactively [10]. The problem is to calculate and obtain an approximation of the volume rendering integral value, which has no analytical solution and must therefore be solved by numerical techniques, which introduce errors in the resulting image or take too long to calculate.

The method's optical model is the emission-absorption, for which the radiation along a beam segment is given by [10]:

$$I(a,b) = \int_{a}^{b} k\left(f(t)\right) \tau\left(f(t)\right) e^{-\int_{a}^{t} \tau(f(u))\, du}\, dt \tag{6}$$

where a and b are the segment boundaries, and k and τ are the color and density transfer function. $f(t)$ is the scalar field value at t discontinuity point, along the segment that represents the ray of light.

Given the function F, the composition of the transfer function (convolution) results in a continuous and differentiable function only in a finite set of discontinuity points. The calculation of visualization integral requires knowledge of these points by taking the following form, with t_i being the ith point of discontinuity:

$$I = \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} k\left(f(t)\right) \tau\left(f(t)\right) e^{-\sum\limits_{j=0}^{i-1} \int_{t_j}^{t_{j+1}} \tau(f(u))\, du - \int_{t_i}^{t} \tau(f(u))\, du}\, dt \tag{7}$$

which, despite being a possible and attractive solution, is computationally prohibitive. The convergence of high order quadrature methods assumes smooth functions (functions that have derivatives of all orders), which is violated by the discontinuity points in this case.

Engel et al. defines the equation for volume visualization as [4]:

$$I(D) = I_0 e^{-\int\limits_{s_0}^{D} k(t) dt} + \int\limits_{s_0}^{D} q(s) e^{-\int\limits_{s}^{D} k(t) dt} ds \tag{8}$$

In which I_0 is the light that enters the volume at point $s = s_0$, $I(D)$ represents the amount of light coming out of the volume through the point $s = D$ and reaches the camera. The set of operations to generate an image representative for a data volume comprises:

Data Traversal definition of the points in data volume, it provides the basis for the discretization of the visualization integral.

Interpolation Normally the sampling points are different from the data grid, so it is necessary to reconstruct the continuous space from the grid to obtain the sample values.

Gradient Computation The gradient of a scalar field is typically used to determine the local illumination.

Classification Held usually by transfer functions is used to map properties of the data on optical characteristics, typically as a set of color values and opacity.

Shading and Illumination Shading can be incorporated to the process by adding a term in the lighting visualization integral - Equation (8).

Compositing It is the iterative process to determine the value of the visualization integral, which can be calculated either starting from the observer or in his direction.

4 Visualization of a n-dimensional data volume

According to the announcement, what is expected to achieve it is to apply the techniques presented in the previous sections for, from a data volume, from whatever process, generate an image in which are represented all the features present in the data, not only the relationship between two or three of its dimensions. As an example, in Figure 2 is shown a three dimensional array whose cells have two scalar values.

To accomplish the visualization of a matrix representing a number of factors, such as that shown in Figure 2, the procedure adopted the following steps:

1. Normalize the data individually to each dimension by using the range of 0 to 255 (with the objective of representing the volume of data optimally);

2. Check for more than one occurrence for each of the elements, recording the position in which they occur in the input matrix into an array of mapping and form a new array of data with unique elements;

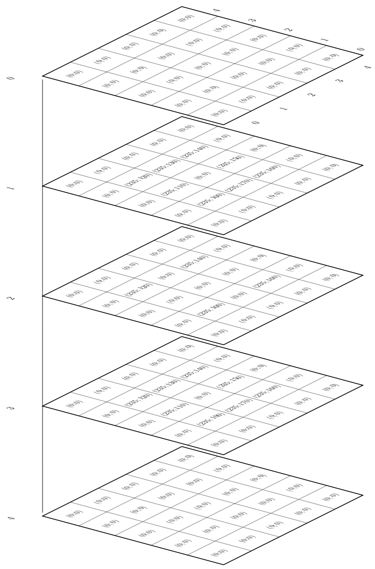


Figure 2: 3D representation of a data matrix containing a 2D vector in each position.

3. Calculate the of distance matrix between each of the elements from the unique elements matrix;

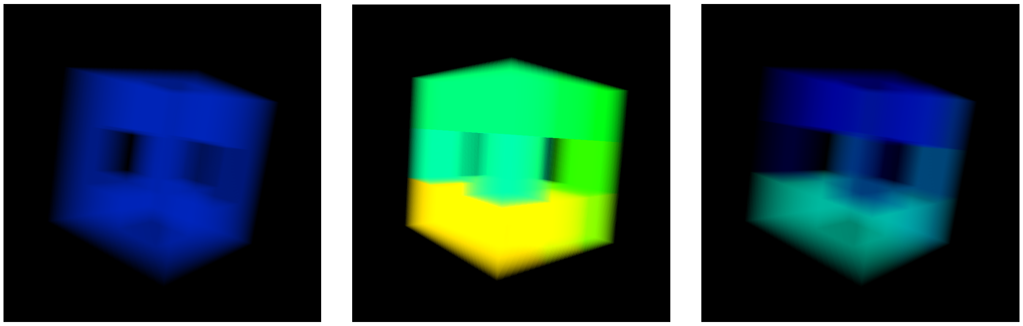
4. Carry out the projection of the unique elements matrix to the target one-dimensional space m;

5. Apply the SMACOF algorithm using the distance matrix, from step 3 and the projection, found in step 4, as input parameters;

6. Considering the results of SMACOF and applying the values in the corresponding positions of the mapping matrix of step 2;

7. Using the reduced data matrix obtained in step 6 as input for the visualization algorithm for generating the data volume representation with reduced dimensions

The result of the algorithm applied in the Figure 2 data matrix is shown in Figure 3. Figure 3a correspond to the first dimension of the original matrix, and Figure 3b presents the contents of its second dimension. Figure 3c is the result of the algorithm with the combined data characteristics present in the two input dimensions.



(a) First dimension (b) Second dimension (c) MDS dimension

Figure 3: Rendering the dimensions of the matrix shown in Figure 2

5 Final considerations

The present study shows that the idea of carrying out the visualization of a set of n-dimensional data is feasible, since the dimensions are reduced so that in each volume position being represented there is a single scalar value, which is used as a scalar field to generate the output image.

The result obtained, according to considerations regarding the illustrated in Figure 3, demonstrates the feasibility of using a dimensional reduction technique in order to achieve a representation of a set of multidimensional data, enabling the generation of an image with the representation of all the information in the original set.

Tests with other dimensional reduction techniques and the possibility of using a projection plane, is work still to be performed, along with the application of this technique in real data sets for the results to be evaluated by experts, confirming its practical usefulness.

Acknowledgements. The authors would like to thank the support given by the Itaipu Binational (the hydroelectric power plant and dam company) to the Graduate Program in Numerical Methods for Engineering (PPGMNE - *Programa de Ps Graduao em Mtodos Numricos em Engenharia*) at UFPR , as well as to thank the Center for Advanced Studies on Dam Safety (CEASB - *Centro de Estudos Avanados em Segurana de Barragens*) of the Itaipu Technological Park (PTI - *Parque Tecnolgico Itaipu*).

References

[1] L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, N.R. Tallent, HPCTOOLKIT: Tools for performance analysis of optimized parallel programs, *Concurrency Computation Practice and Experience,* (2013), 662 - 682.

[2] I. Borg, P.J.F. Groenen, *Modern Multidimensional Scaling - Theory and Applications,* Springer, New York, 2005.
http://dx.doi.org/10.1007/0-387-28981-x

[3] S. dos Santos, K. Brodlie, Gaining understanding of multivariate and multidimensional data through visualization, *Computers & Graphics,* **28** (2004), no. 3, 311 - 325. http://dx.doi.org/10.1016/j.cag.2004.03.013

[4] K. Engel, M. Hadwiger, J.M. Kniss, C.R. Salama, D. Weiskopf, *Real-Time Volume Graphics,* A K Peters Ltd., 2006.
http://dx.doi.org/10.1201/b10629

[5] A.S. Glassner, *Principles of Digital Image Synthesis,* Morgan Kaufmann Publishers Inc., San Francisco, CA, 1995.

[6] H. Guo, H. Xiao, X. Yuan, Multi-Dimensional Transfer Function Design based on Flexible Dimension Projection Embedded in Parallel Coordinates, *IEEE Pacific Visualization Symposium,* (2011), 19 - 26.
http://dx.doi.org/10.1109/pacificvis.2011.5742368

[7] J. Lawrence, S. Arietta, M. Kazhdan, D. Lepage, C. O'Hagan, A user-assisted approach to visualizing multidimensional images, *IEEE Transactions on Visualization and Computer Graphics,* **17** (2011), 1487 - 1498.
http://dx.doi.org/10.1109/tvcg.2010.229

[8] R.M. Martins, D.B. Coimbra, R. Minghim, A.C. Telea, Visual analysis of dimensionality reduction quality for parameterized projections, *Computers and Graphics,* **41** (2014), 26 - 42.
http://dx.doi.org/10.1016/j.cag.2014.01.006

[9] S.A. Najim, I. S. Lim, Trustworthy dimension reduction for visualization different data sets, *Information Sciences,* **278** (2014), 206 - 220.
http://dx.doi.org/10.1016/j.ins.2014.03.048

[10] B. Nelson, R.M. Kirby, R. Haimes, GPU-based volume visualization from high-order finite element fields, *IEEE Transactions on Visualization and Computer Graphics,* **20** (2014), no. 1, 70 - 83.
http://dx.doi.org/10.1109/tvcg.2013.96

[11] Y. Pao, Z. Meng, Visualization and the understanding of multidimensional data, *Engineering Applications of Artificial Intelligence,* **5** (1998), 659 - 667. http://dx.doi.org/10.1016/s0952-1976(98)00031-1

[12] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science,* **290** (2000), 2319 - 2323. http://dx.doi.org/10.1126/science.290.5500.2319

[13] L. Van Der Maaten, E.O. Postma, J. Van Den Herik, Dimensionality Reduction: A Comparative Review, *Journal of Machine Learning Research,* **10** (2009), 1 - 41.

[14] H. Wright, *Introduction to Scientific Visualization,* Springer, UK, 2007. http://dx.doi.org/10.1007/978-1-84628-755-8

Received: January 19, 2016; Publsihed: March 18, 2016