Information Engineering Express
International Institute of Applied Informatics
2016, Vol.2, No.1, 87 – 96

Visual Explanation of Eigenvalues and Math Process in Latent Semantic Analysis

Yukari Shirota [*], Basabi Chacraborty [†]

Abstract

Latent Semantic Analysis (LSA) is a widely used method in text mining field to extract the underlying concepts in the text document. The mathematical technique behind LSA is Singular Value Decomposition (SVD) in which the key concept is the eigenvalues. It is difficult to understand the underlying mathematics for general people, not proficient in mathematics. One reason might be that the linear algebra textbooks available in the market are not written for non–mathematics majors such as economics students. We believe that there is better teaching method to explain the eigenvalues and eigenvectors to our students. In this paper, we would like to illustrate the method. In the main part of the paper, we have proposed a visualization of the mathematical process behind LSA to make it easily understandable to general people, novice in mathematics. In addition, to understand the SVD process more deeply, another example which is a time series data analysis by SVD is also presented.

Keywords: LSA, text mining, SVD, visualization, mathematics; multi-variant analysis, eigenvalue

1 Introduction

Learning mathematics is originally enjoyable. However, it is not simple. We teach multi-variant analysis methods to our university students. Our job is to motivate them to work hard in order to understand the mathematical philosophy before working with statistical tools such as SPSS and SAS. From over 15-year experience of teaching mathematics to undergraduate students, we found that text books of linear algebra are in general too difficult for them to understand. Some simple interpretation is needed so that the students become excited about learning mathematics. We found that visualization of the mathematical process behind any technique might help them growing their interests in mathematics. This paper describes visualization of the mathematical process in Latent Semantic Analysis (LSA). Currently many researchers use text mining techniques in various fields because of the recent popularity of the analysis of big data around the world in various application areas. Especially LSA is widely used in the area of text mining [1]. For example, text mining technologies are used to extract the content of the communication

[*] Gakushuin University, Tokyo, Japan
[†] Iwate Prefectural University, Iwate, Japan

between the citizens and their governments as voices of the people and these are used as feedback for government decision support. Evangelopoulos et al. presented how LSA is implemented on real E-Democracy data so that the politicians can efficiently interpret the feedback by citizen [2]. We think that understanding the mathematical process behind LSA would lead to skills for understanding various other multi-variant analytical methods. The mathematical process behind LSA is Singular Value Decomposition (SVD). SVD includes the significant essence of eigenvalues and eigenvectors. So at the beginning we would like to visually explain eigenvalues and eigenvectors. Then SVD is explained visually through an application of LSA.

In addition, another SVD application which is a time series data analysis problem is described.
between the citizens and their governments as voices of the people and these are used as feedback for government decision support. Evangelopoulos et al. presented how LSA is implemented on real E-Democracy data so that the politicians can efficiently interpret the feedback by citizen [2]. We think that understanding the mathematical process behind LSA would lead to skills for understanding various other multi-variant analytical methods. The mathematical process behind LSA is Singular Value Decomposition (SVD). SVD includes the significant essence of eigenvalues and eigenvectors. So at the beginning we would like to visually explain eigenvalues and eigenvectors. Then SVD is explained visually through an application of LSA. In addition, another SVD application which is a time series data analysis problem is described.

2 Review of LSA and SVD Teaching Methods

In this section, a review the existing teaching methods of LSA and SVD is presented. Analysis of text data starts with the generation of a term-frequency matrix which is analyzed with the purpose of extracting latent components of meaning through SVD. A matrix operation that extracts simultaneous least-square principal components of two sets of variables, namely the set of terms and the sets of documents, is executed in [1, 3]. LSA (or Latent Semantic Indexing LSI) is explained in the textbooks such as [3-5]. SVD is an extension of PCA (Pricipal Component Analysis). The PCA extracts eigenvalues and eigenvectors of a covariance-matrix when the matrix is symmetrical. However, a term-frequency matrix is not symmetrical. Therefore, we use SVD instead of PCA. PCA is explained in the textbooks such as [6-9]. We also have developed various interactive and visual teaching materials in [10-12] for PCA. Knowledge of PCA is helpful before learning SVD. Let us suppose that the term-document matrix is represented by $\mathbf{X}$. By SVD, we obtain $\mathbf{X} = \mathbf{U\Sigma}\ \ \boldsymbol{V}^{\mathrm{T}}$ where $\mathbf{U\Sigma}$ is the term-eigenvectors, $\mathbf{\Sigma}\ \ \boldsymbol{V}^{\mathrm{T}}$ is the document-eigenvectors, and $\mathbf{\Sigma}$ is the diagonal matrix of singular values. SVD is explained in the textbooks such as [4, 13].

The mathematical definition of SVD can also be found in many web sites such as Wikipe-dia[14]. The wikipedia article offers the visual explanation concerning the matrix operation. However, we hardly think that seeing and reading the explanation will make students understand the meaning of SVD, because the visualized example is two-dimensional and it illustrates only two canonical unit vectors. For students who want to learn text mining, much more application-based explanation is required. The existing textbooks and papers which explain LSA offer only the example with calculation and its analysis. They do not focus on explanation of the meaning of the mathematical process behind LSA. Before learning SVD, students need to understand the concept of eigenvalues, because to understand SVD, the most significant concept is the eigen-values. The eigenvalue is invariant under the change of the basis matrix. To express the invariance of eigenvalues, the terms "characteristic value" or "proper value" are sometimes used instead of

"eigenvalue" in some linear algebra textbook [15]. However, we hardly think that the term can give a concrete image of eigenvalues to our students. If the target is students in a mathematics department, they would be able to grasp the concept for themselves. However, students having non mathematics major could have difficulties to understand the eigenvalue concepts.

Generally speaking, a textbook of linear algebra should teach the eigenvalue concept in a more easily understandable way. In many existing textbooks, the following explanation is offered[16, 17]: *First, given a matrix. Then find the eigenvalues and the eigenvectors. Finally obtain the diagonal matrix.* We think that it should be described more directly what is invariant after the change of a basis matrix. Therefore, we propose a method to explain the process for our students. When I teach eigenvalues in my classes, I use the fable story titled "Enlargement Factors of the Magnification Machine are Eigenvalues" [11]. The summary of this story is that the magnification machine which corresponds to the diagonal matrix was invented in the CENTRE country and to export the machine to another country which uses a different basis, the linear transformation was executed. The key point is an eigenvalue is expressed as an enlargement factor of the magnification machine. By the story, many students could comprehend the eigenvalue concept. The visual teaching materials of eigenvectors and eigenvalues through PCA has been illustrated (See Figure 1). The data used is the agriculture and industry values added percentage of GDP from World Bank Data 2012. It is ordinary two dimensional data around us. Using the data, three-dimensional histogram has been drawn. Then we mold the shape by the Gaussian distribution model as shown in Figure 1. The Gaussian model shape is the approximation of the given data. The eigenvalues and eigenvectors on the graphics are then added as the two arrows. The arrows' directions correspond to the eigenvectors and the arrow lengths correspond to the eigenvalues. How much the data is expanded to the respective directions conveys the meaning of eigenvalues. The concrete values of the covariance matrix are as follows:

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix} \begin{bmatrix} -0.55 \\ 0.83 \end{bmatrix} = 233 \begin{bmatrix} -0.55 \\ 0.83 \end{bmatrix}$$

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix} \begin{bmatrix} -0.83 \\ -0.55 \end{bmatrix} = 108 \begin{bmatrix} -0.83 \\ -0.55 \end{bmatrix}$$

From the two eigenvectors, we can get the rotation matrix of the data as follows:

$$\begin{pmatrix} -0.55 & -0.83 \\ 0.83 & -0.55 \end{pmatrix}$$

When we use the above-mentioned fable story, the magnificent operation by 233 times and 108 times in CENTRE expression is divided to three step operations as follows: (1) interpretation from CENTRE one to WEST country one, (2) magnificent operation in WEST, and (3) interpretation from WEST one to CENRE one. The point of the teaching method is that the diagonal matrix which expresses the enlargement factors should be offered first before other non-diagonal matrices. Another point is the usage of Gaussian distribution model. By these, our students get to understand the eigenvalue concepts more deeply. We shall apply the eigenvalue-centered way of explanation to LSA/SVD explanation in the next section.

3 Visualization of Mathematical Process of SVD

In this section, we would like to explain our proposed teaching materials for the mathematical process behind SVD. The target students are supposed to learn LSA. Here we emphasize the

conceptual underpinnings of the mathematics by using stories and illustrations and interactive visual graphics.

Let us suppose that a term-document (or term-frequency) matrix X in Figure 2 is given. There are three documents and the six significant words selected from the documents. The words are "kindness", "tenderness", "dream", "glory", "promise", and "confidence". We suppose in advance that there are three essential concepts in the given documents/books; they are love, hope, and trust. We selected the three concepts because these are the strongest concepts without which a man cannot survive and that we think students can easily associate the latent concept to an eigenvector. The impact factor of the concepts are considered to be the eigenvalues. Although this explanation is not correct from a mathematical standpoint, this analogy would make students easily understand the mathematical process behind LSA.

Looking at the matrix $\boldsymbol{X}$, we can see that the document #1 is related to the concept "love", the document #2 is related to the concept "hope", and the document #3 is related to the concept "trust". In real text mining, at first we generally have no idea of such concepts or topics. After SVD, looking at the results of SVD, the interpretation of latent semantics would be conducted. The given term-document matrix in Figure 2 is supposed to be divided into three matrices by SVD, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\ \boldsymbol{V}^T$, as shown in Figure 3. In general, results of SVD is given as follows:

Σ: an $r \times r$ diagonal matrix.

λ_i (Diagonal elements of the i rows and i columns) is $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$

A column vector of $\boldsymbol{U}$ is an unit vector and orthonormal basis of the space spanned by the column vector of $\boldsymbol{X}$. A row vector of $\boldsymbol{V}^T$ is an unit vector and orthonormal basis of the space spanned by the row vector of $\boldsymbol{X}$. An eigenvalue λ_i shows the importance of as the base of the ith column vector of $\boldsymbol{U}$ (or the ith row vector of $\boldsymbol{V}^T$).

For interpretation of the above explanations, we use a story (See Figure 4 and 5); where there are six-dimensional term world and three-dimensional document world as shown in the figures. The transformation from a document to a word is executed by the matrix $\boldsymbol{X}$ and the transformation from a word to a document is executed by the matrix $\boldsymbol{X}^{\mathrm{T}}$. As shown in Figure 3, the first unit column vector in $\boldsymbol{U}$ is magnified by the eigenvalue 81.63 and then the resultant vector is the first term-eigenvector which corresponds to the first column $\boldsymbol{U\Sigma}$. Similarly the first unit column vector in $\boldsymbol{V}$ is magnified by the eigenvalue 81.63 and then the resultant vector is the first term-eigenvector which corresponds to the first column of $\boldsymbol{V\Sigma}$, because $(\boldsymbol{V\Sigma})^T = \boldsymbol{\Sigma}^T\boldsymbol{V}^T = \boldsymbol{\Sigma}\ \boldsymbol{V}^T$

The first term-eigenvector in Figure 4 corresponds to the latent concept "**trust**" and the eigenvector keeps the same vector direction even after a return trip to and from the document world. Any term vector other than the term-eigenvector is skewed by the transformation. The to-and-from transformation can be expressed by the matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ of which eigenvalues are approximately 6664, 4908, and 3153. We found eigenvalues of $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ are equal to square of eigenvalues in $\boldsymbol{\Sigma}$. For example, the square of 81.63 equals to 6664. We could consider that the three latent concepts such as **love** can be expressed as the three term-eigenvectors in the term world (See Figure 4) and that the three latent concepts can be also expressed as the three document-eigenvectors in the document world (See Figure 5). The term-eigenvector in the term world will be transformed to the corresponding document; for example the book is supposed to be a book defining the concept "**love**". After a transformation, the latent concept is invariant. The eigenvector changes in the space. However, the latent concept is invariant and the impact factor (eigenvalue) is invariant.

The canonical unit document-eigenvectors are shown in Figure 6. The scaled document-eigenvectors by the eigenvalues are also shown in Figure 6. This kind of geometrical presentations have already used in the existing textbooks and papers. Therefore we skip the visual explanations like this such as skewed vectors and a dimension reduction projection in this paper.

4 SVD in Time Series Data Analysis

The SVD is used in various application fields. In this section, we present another SVD application which is the stock price fluctuation analysis. In the financial field, SVD is used to make portfolios. Using SVD, we can find the similar kind of industry group like telephone companies and banking companies [18,19,20]. For example, Bank A and Bank B tend to move similarly. The given data of the application is stock price time series data as shown in Figure 7. In the example, there are seven companies and their 22 days stock price data. The results of SVD are seven eigenvalues. The two types of eigenvectors correspond to one eigenvalue and in the application they are a 7-dimensional vector and a 22-dimensional vector. In [20], the vectors are called Brand-Eigenvector and Motion-Eigenvector. The paper also uses the notations. The Brand-Eigenvector shows the group of companies with similar stock movement. On the other hand, the Motion-Eigenvector shows an average fluctuation of the group.

In Figure 8, the transformation of Brand-Eigenvectors from the Brand World to the Daily motion world and again to the Brand World is shown. The Brand-Eigenvector has the same direction as before, only being multiplied by the eigenvalue squared. In Figure 9, Motion-Eigenvectors are shown. The Motion-Eigenvector is 22-dimensional and each value illustrates the group average fluctuation.

5 Conclusions

We discuss visual explanation methods of SVD in LSA. In this paper, we presented two applications of SVD. Through two examples, we can see that SVD shows two different standpoints of an intrinsic meaning of one eigenvalue. In LSA, they are expressed by words and documents. In the stock price analysis, they are expressed by companies (brands) and fluctuations. As SVD is used in various application fields, understanding the mathematical process would be very helpful in many fields. Currently the multi-variant analysis methods have becoming more complicated. To teach the mathematical process, visual approach is helpful. We use various visualization materials; they are interactive and visual materials (See our other published materials), simulations, stories, and illustrations as shown in Figure 1. As shown in the paper, we present that mathematical illustrations have the strong descriptive power and they are useful in mathematical education. In our statistics classes, we found that these mathematical illustrations have made many students easily understand the essence of the theories.

The people need more visualized explanation as we use more complicated mathematical or statistical analysis. In statistics education, we have to make students convinced the meaning of the theory by using visual materials before making students calculate the expressions. We would like to continuously develop visual teaching materials of multi-variant analytical methods for text mining researchers.

References

[1] N. Evangelopoulos, and L. Visinescu, “Text-mining the voice of the people,” Communications of the ACM, vol. 55, no. 2, pp. 62-69, 2012.

[2] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” Discourse processes, vol. 25, no. 2-3, pp. 259-284, 1998.

[3] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Handbook of latent semantic analysis: Psychology Press, 2013.

[4] C. D. Manning, P. Raghavan, and H. Schuetze, Introduction to Information Retrieval: Cambridge University Press, 2008.

[5] D. A. Grossman, Information retrieval: Algorithms and heuristics: Springer, 2004.

[6] T. H. Wonnacott, and R. J. Wonnacott, REGRESSION: John Wiley & Sons, Inc., 1981.

[7] S. Konishi, Introduction to Multivariate Analysis: Linear and Nonlinear Modeling: Chapman & Hall/CRC, 2014.

[8] I. Koch, Analysis of Multivariate and High-Dimensional Data: Cambridge University Press, 2013.

[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2 ed.: Springer, 2009.

[10] Y. Shirota, and T. Hashimoto, “Knowledge Visualization of Reasoning for Statistical Problems,” Annual Report of Gakushuin University Research Institute for Economics and Management (GEM Bulletin), vol. 28, pp. 45-54, 2014/12, 2014.

[11] Y. Shirota, “Practical Teaching Methods of Linear Algebra for Students in the Economics Course,” Gakushuin Economics Papers, vol. 51, no. 2, pp. 133-147, 2014/07, 2014.

[12] Y. Shirota, and T. Hashimoto, “Web Publication of Three-Dimensional Animation Materials for Business Mathematics : 10 Graphics for Economics Mathematics (Part 2),” Annual Report of Gakushuin University Research Institute for Economics and Management (GEM Bulletin), no. 26, pp. 13-22, 2012/12, 2012.

[13] B. Mirkin, Core Concepts in Data Analysis: Summarization, Correlation and Visualization (Undergraduate Topics in Computer Science): Springer, 2011.

[14] Wikipedia. "Singular Value Decomposition," 2015; http://en.wikipedia.org/wiki/Singular_value_decomposition.

[15] S. Lipschutz, Theory and Problems of Beginning Linear Algebra: McGraw-Hill, 1997.

[16] B. Kolman, and D. R. Hill, Introductory Linear Algebra, 8 ed.: Pearson, 2005.

[17] W. K. Nicholson, Linear Algebra With Applications, 6 ed.: McGraw-Hill, 2003.

[18] V. Plerou et al., “Random matrix approach to cross correlations in financila data,” Physical Review E, Vol. 65, No. 6, pp. 066126-1-066126-18, 2002.

[19] V. Plerou et al., “A random matrix theory approach to financial cross-correlations,” Physica A: Statistical Mechanics and its Applications, Vol. 287, No. 34, pp. 374-382, 2000.

[20] M.F. Lubis, Y. Shirota, and R.F. Sari, “Thailand's 2011 Flooding: its Impacts on Japan Companies in Stock Price Data,” Gakushuin Economics Papers, Vol. 52, No. 3, pp. 101-121, 2015.

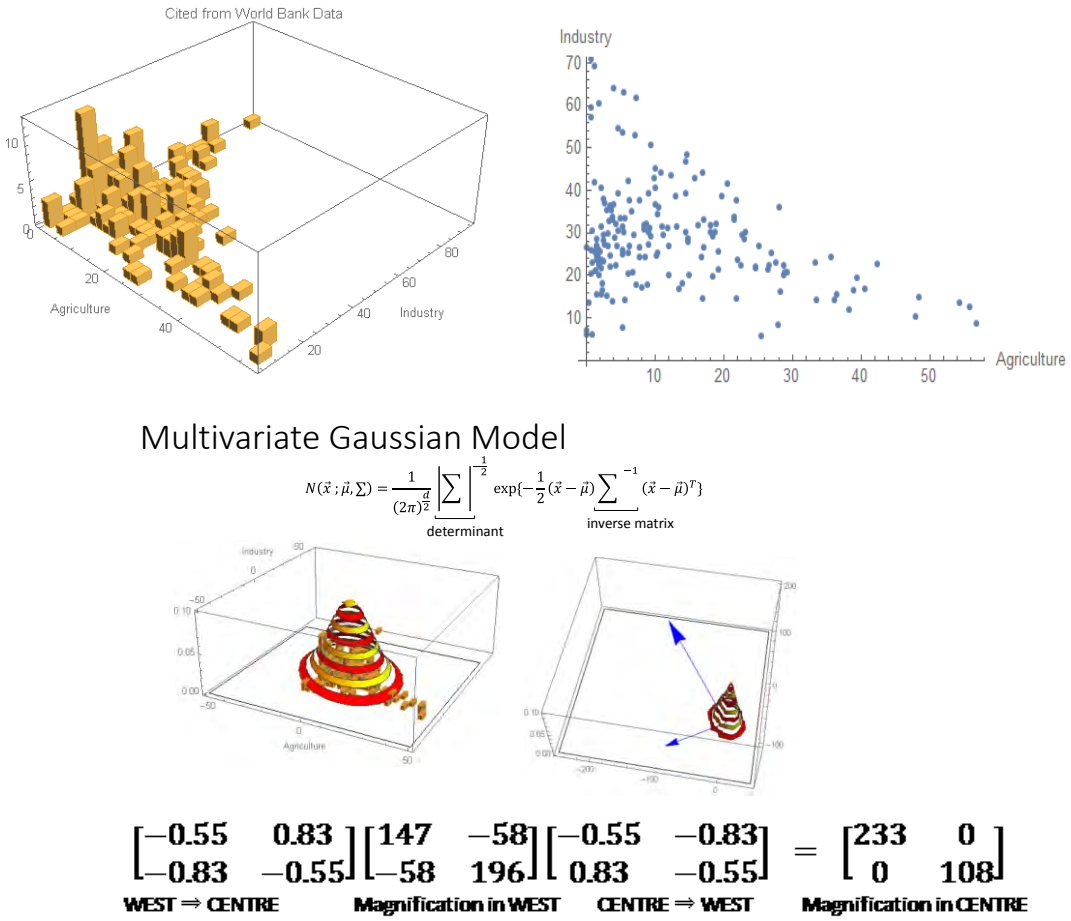

Figure 1: PCA visual teaching materials which explain the eigenvalues and eigenvectors.

6 Keywords By TFIDF

3 Documents

	Document #1	Document #2	Document #3
affection	56	2	1
passion	42	1	0
expectation	0	38	0
desire	0	35	3
promise	0	24	18
believe	0	0	79

Unknown

This group latent semantic may be [Love]

This group latent semantic may be [Hope]

This group latent semantic may be [Trust]

Figure 2: The given term-document matrix which was artificially created sample data with

three latent topics “love”, “hope”, and “trust”.

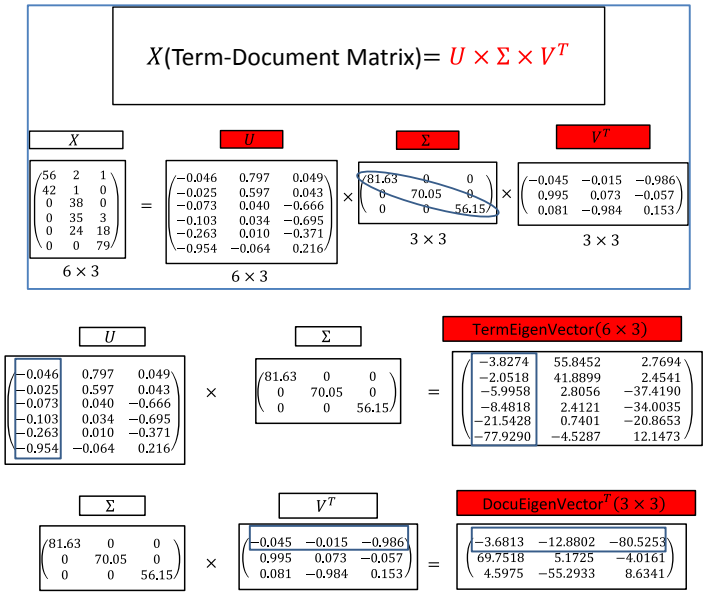

Figure 3: The SVD has made the term-document matrix divided into the three matrices.

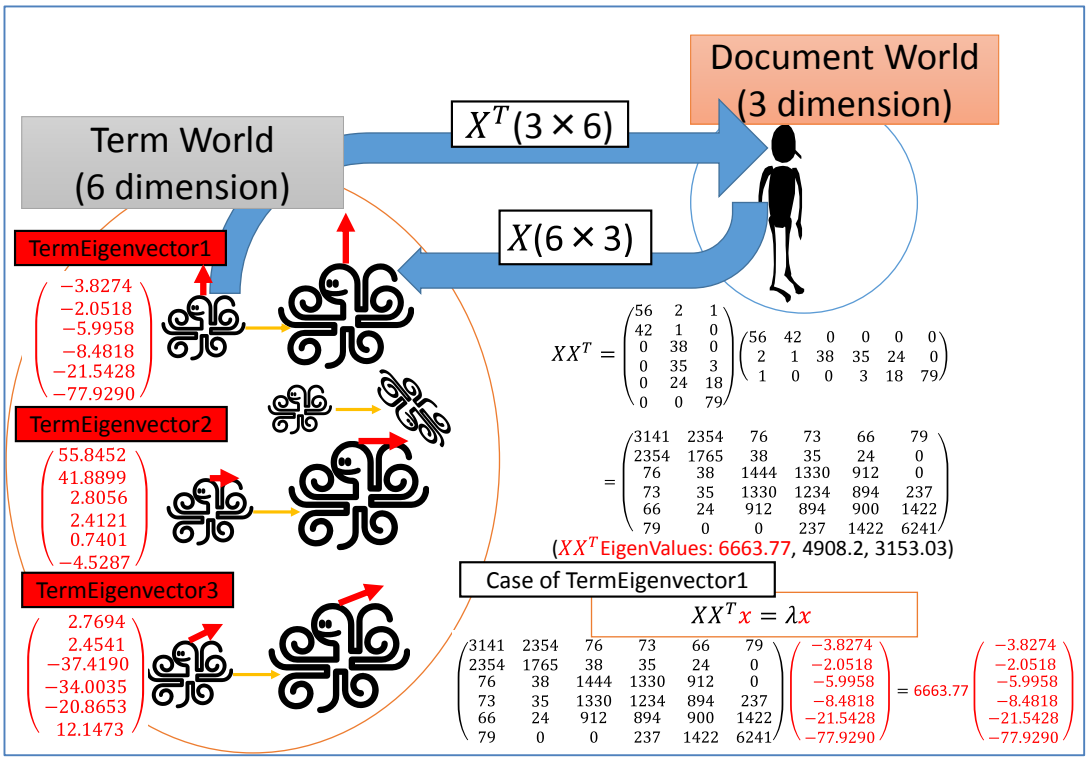

Figure 4: The SVD has made the term-document matrix divided into the three matrices.

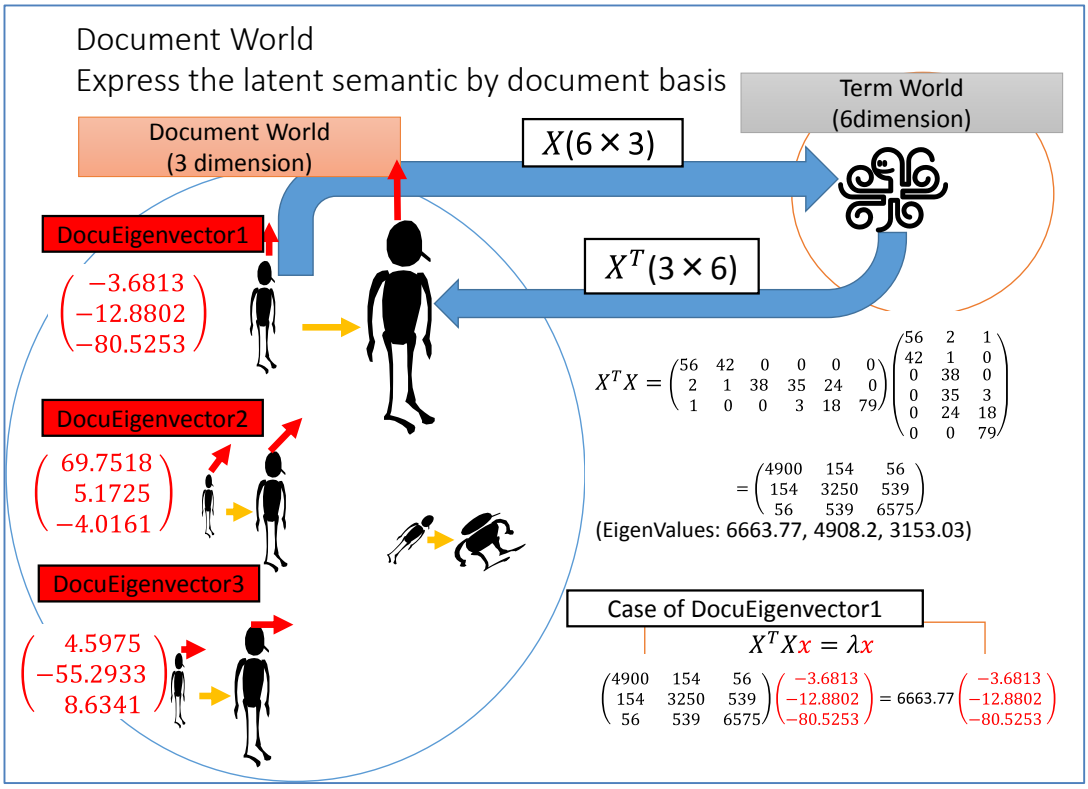

Figure 5: The mathematical illustration of term-eigenvectors and eigenvalues in SVD.

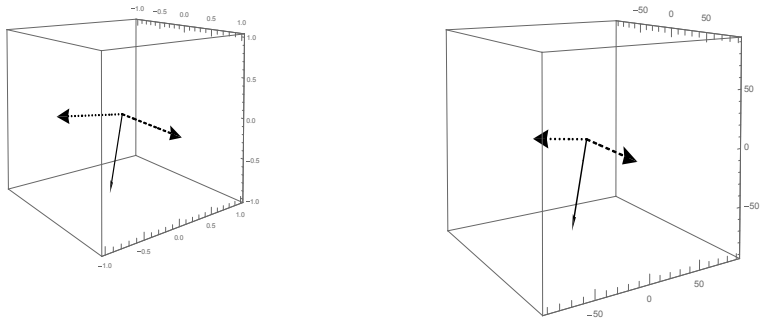

Figure 6: The unit document-eigenvectors and scaled one which is scaled by its eigenvalue.

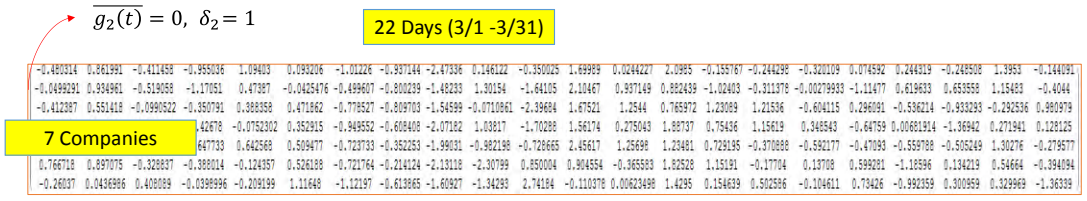

Figure 7: The given time series data of the stock price fluctuation. In advance, the data has been standardized.

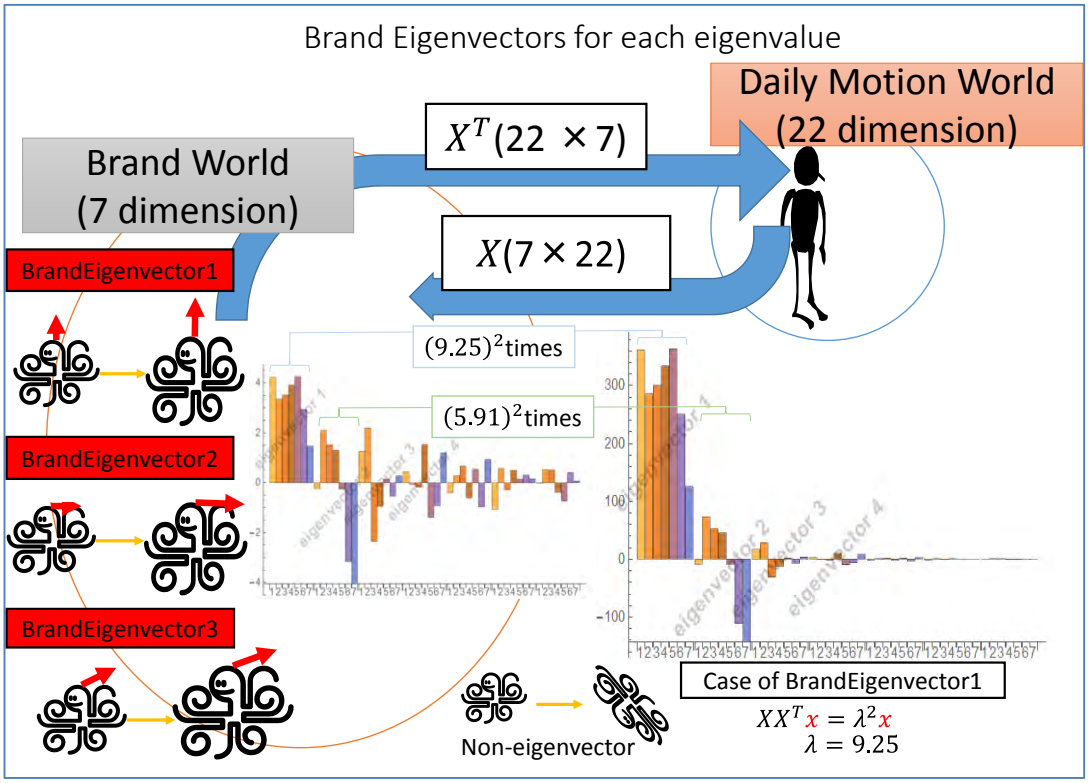

Figure 8: The BrandEigenvectors transformation between the two expression worlds.

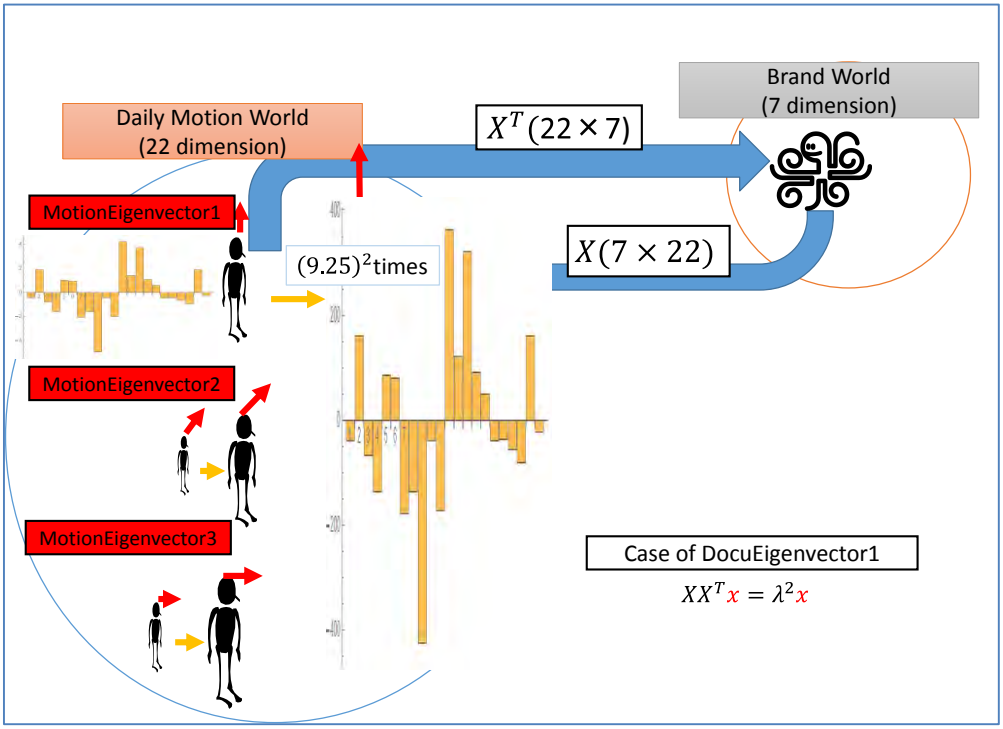

Figure 9: The MotionEigenvectors transformation between the two expression worlds.