Practical Teaching Methods of Linear Algebra
for Students in the Economics Course

Yukari Shirota*

ABSTRACT

To analyze the big data on web or so, a new type teaching methods of linear algebra is required. In the paper, three requirements for the new teaching method are discussed. They are realated to (1) Vector-notation based expression transformation, (2) eigenvalues of a variance-covariance matrix, and (3) visualization of multivariant analysis.

1. Introduction

In the paper, I will discuss a new type teaching methods of linear algebra. Suppose that the target student for that is a student in the economics course. There are many linera algebra textbooks for science and engineering courses. However I think a different teaching way of the linear algebra for the economics students is needed. Especially between students in a mathematics course and students in an economics course, a completely different textbook is required.

Studetns in the economics course need basic ideas of statistics and linear algebra, so that they can conduct multivariant analysis or so. Additionally, various knowledges such as a Bayesian inference and a time series analysis are required for them as basic mathematical ideas. For example, some economical or marketing decision making processes use the Bayesian inference in which using the observation value, the probability inference of a non-observation value is performed by a method such as a Bayesian network or so. Then, Bayesian inference concepts such as Bayesian predictive distributions, a maximum a posteriori probability (MAP) estimation and so forth would be needed[1, 2]. To meet needs of the time, a teaching method of linear algebra should also be changed.

In the paper, I shall propose the practical teaching method of linear algebra for economics students. In the next section, I shall propose three policies of the teaching method. After that, two concrete examples will be illustrated in Section 3 and 4 which are a MAP estimation by the Gaussian mixture model and the Principal Component Analysis (PCA) and the Latent Semantic Analysis (LSA). Finally, I shall conclude the paper.

---

*）　Faculty of Economics, Gakushuin University, Tokyo, Japan
　　yukari.shirota@gakushuin.ac.jp

2. Proposed Teaching Policies

In the section, I shall propose three requirements for the new teaching methodology of linear algebra. They are:

(1) **Vector-notation based expression transformation**: Students should be able to understand and translate the given math expressions in a vector-notation (or a matrix form), so that they can conduct the concrete algebraic operations. The transformation skills in a matrix form should be cultivated.

(2) **Eigenvalues of a variance-covariance matrix**: Students should understand deeply the invariance of eigenvalues. I think that the most important concept in linear algebra is eigenvalues. Many multivariant analysis methods such as Principle Component Analysis (abbreviated as PCA) need the knowledge of eigenvalues.

(3) **Visualization of multivariant analysis**: Students should have visual images of the mathematical processes such as PCA and the Mahalanobis distances to grasp the meanings.

I shall explain my proposal in more detail.

2.1 Vector-notation based expression transformation

Recently the complexity level of the given math problem has become quite high compared to the previous time. One of the reasons would be that we have to analyze big data on the web. The authors of the famous machine learning textbook say in the preface as follows[3]:

Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and understand "what the data says." We call this learning from data. While some mathematical details are needed, we emphasize the methods and their conceptual underpinnings rather than their theoretical properties. As a result, we hope that this book will appeal not just to statisticians but also to researchers and practitioners in a wide variety of fields.

I think that a new type teaching method for linear algebra is needed and that a suitable mathematical notation for that is needed to *emphasize the methods and their conceptual underpinnings*. Hence, I think that a vector-notation is essential. The reason is that an explanation without the vector-notaion, namely a step-by-step calculation expalnation is too long to grasp the whole flow and makes the whole process obscure.

In the first place, some mathematical software such as MATLAB or Mathematica is supposed to be used for the analysis. A problem we handle now is not a simple one that a manual labour can calculate. Therefore, some mathematical software is absolutely essential for the calculation. Of course, students should in advance understand the basic idea of the calculations. However that is not the goal and the goal is far away from that. We should assume from the beginning that students calculate using the mathematical tools. In addition, the mathematical notation should be carefully selected. A good example can be seen in famous textbooks [3-5].

The students should also get accustommed to expression transformation in the vector-notation. That

includes partial differentiation like the Lagrange multiplier method. I hardly think that many students have enough such vector-noted transformation skills. The new type teaching method of linear algebra should focus on a cultivation of the matrix transformation ability.

2.2 Eigenvalues of a variance-covariance matrix

I think that a textbook of linear algebra should teach the eigenvalue concept more directly. In many existing textbooks, the following step explanation is offered: First, given a matrix. Then find the eigenvalues and the eigenvectors. Finally obtain the diagonal matrix. The important concept is that eigenvalues are invariant under a change of a basis matrix. I hardly think that the above-mentioned explanations make students comprehend the invariance. If the taget is a student in a mathematics department, they would be able to grasp the concept for themselves. However, other course students have difficulties to understand the eigenvalues.

Therefore, I contrived means for the explanation for our students. When I teach eigenvalues in my classes, I use the fable story titled "Enlargement Factors of the Magnification Machine are Eigenvalues" (See the appendix). The key point is an enlargement factor of the magnification machine. By the story, many students have comprehended the eigenvalue concept. The point of the teaching is that the diagonal matrix should be offered first before other non-diagonal matrices. The viasulalization of PCA as illustrated in Section 4 is also effective to have students understand the eigenvalues.

2.3 Visualization of multivariant analysis

There are many maltivariant analyses. The student have to spot the differneces among the analytical approaches in order to select the most suitable one. I think that a linear algebra textbook for economics students needs to include many word math problems using multivariant analyses. Then before the algebraic calculations, the visual images of 2-dimension or 3-dimension should be offered to students so that they can easily understand the approach. Such visualization is helpful so that students can understand a general dimensional matrix case.

The author has been so far developing many visual teaching materials [6, 7]. For example, the Central Limit Theorem visual materials are published on the web page[1]. In my classes, the Wolfram CDF[2] materials I made are available; although a teacher needs to buy Mathematica to make the CDF material, a student can see and use the CDF material only by the CDF player which is a free software tool. Recently a well-known machine learning textbook has been published as an e-book[3]. The Bishop's textbook also offers beautiful visual images[5]. I have also published a basis financial math e-textbook, so that the readers can use the graphical tools interactively and visually[7]. To take the advantage of visualization, more and more math e-textbooks would be continuously published.

I have discussed three requirements for a new linear algebra teaching method. These should be reflected in educational policies of linear algebra.

1） Yukari Shirota: Mathematics Graphics Materials,
　　http://www-cc.gakushuin.ac.jp/~20010570/mathABC/ABC/

2） Wolfram CDF player: http://www.wolfram.com/cdf-player/

3. Gaussian Mixture Model and Maharanobis Distance

In the section, as an example of teaching materials, I will consider the Gaussian mixture model (GMM) and Maharanobis distance. What I would like to show is the matehamtical notation and its transformation style, and the visualization. Concerning the mathematical notation, I referred to [8]. The graphics are my original ones.

Suppose that the multivariate Gaussian distribution over a set of random variables {Xi} is characterized by an d-dimensional mean value μ, and a symmetic $n \times n$ variance-covariance matirix Σ. Then the density function can be defines as follows:

$$q(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$$

The matirix Σ needs to be a positive definite matrix. The two parameters μ and Σ of the Gaussian model are defined as follows:

$$\begin{cases} \mu = \mathbb{E}_x[x] = \displaystyle\int xq(x;\mu,\Sigma)dx \\ \Sigma = \mathbb{V}_x[x] = \displaystyle\int (x-\mu)(x-\mu)^{\mathrm{T}}q(x;\mu,\Sigma)dx \end{cases}$$

The mathematical formulas to handle the Gaussian distribution should also be noted as a matrix form. I will show the formula examples (cited from [9]):

$$x^T A x = \mathrm{tr}(A x x^T)$$
$$\frac{\partial}{\partial A}\log|A| = (A^{-1})^T$$
$$\frac{\partial}{\partial x}\log|A| = \mathrm{tr}\left(A^{-1}\frac{\partial}{\partial x}A\right)$$
$$\frac{\partial}{\partial A}\mathrm{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^T$$
$$|I + ab^T| = 1 + a^T b$$

In many statistical applications, it is more important to handle the logarithm of the determinant than the determinant. So transformation formulas concerning that should be taught in linear algebra.

Now let me conduct a maximum a posteriori probability (MAP) estimate using sample data. The MAP estimate is explained as follows[3]: *In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is a mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to Fisher's method of maximum likelihood (ML), but employs an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate.* Because the MAP estimate is a point estimate, the calculation becomes simpler than an estimate of the whole distribution function.

I shall estimate the following conditional density for the category y by the Gaussian model where $\widehat{\Sigma}_y$ is the estimated quantity corresponding to the data set {Xi} in the category y.

---

3) From Wikipedia, the free encyclopedia of a maximum a posteriori probability.

$$\hat{p}(x|y) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\widehat{\Sigma}_y)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \widehat{\mu}_y)^{\mathrm{T}}\widehat{\Sigma}_y^{-1}(x - \widehat{\mu}_y)\right)$$

We have the chain rule $p(y|x)p(x) = p(x, y) = p(x|y)p(y)$ and by using this, the Bayes' rule $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$ is obtained. Then we shall take a log of the both sides of the Bayes' rule expression:

$$\log p(y|x) = \log p(x|y) + \log p(y) - \log p(x)$$

By using the sample size of each category n_y, the distribution p(y) can be estimated as

$$\hat{p}(y) = \frac{n_y}{n}$$

Taking a log of that, the posterior probability can be calculated as follows:

$$\log \hat{p}(y|x) = \log \hat{p}(x|y) + \log \hat{p}(y) - \log p(x)$$
$$= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(\det(\widehat{\Sigma}_y)\right) \quad -\frac{1}{2}(x - \widehat{\mu}_y)^{\mathrm{T}}\widehat{\Sigma}_y^{-1}(x - \widehat{\mu}_y) + \log\frac{n_y}{n} - \log p(x)$$
$$= -\frac{1}{2}(x - \widehat{\mu}_y)^{\mathrm{T}}\widehat{\Sigma}_y^{-1}(x - \widehat{\mu}_y) - \frac{1}{2}\log\left(\det(\widehat{\Sigma}_y)\right) + \log n_y + C$$

As shown here, the posterior probability expression includes a term of the variance-covariance estimator $\widehat{\Sigma}_y$ and a term of the category y number n_y. Here the term $(x - \widehat{\mu}_y)^{\mathrm{T}}\widehat{\Sigma}_y^{-1}(x - \widehat{\mu}_y)$ is called a Mahalanobis distance.
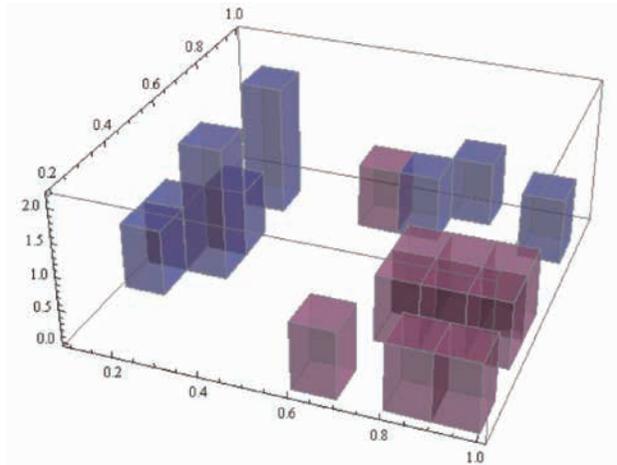

Figure 1: The histogram of the given sample data. The blue (pale) colored ones belong to the category "WHITE" and the red (dark) colored ones belong to the category "BLACK."

In the following, using the sample data of two categories named "WHITE" and "BLACK", I will visualize the Mahalanobis distance or so forth. The given data is cited from [10]. In Figure 1, the histogram is illustrated. An individual category has 10 sample data. Suppose that we have no idea concerning categories of the trained data. Without the categories information, I shall visualize the Gaussian model for the sample data (See Figure 2). The eigenvectors of the variance-covariance matrix are also shown here. The contour lines are corresponding to the Mahalanobis distance. In Figure 3, the Gaussian model of the distribution and the two eigenvectors are illustrated at the same time. Seeing that, we found that following the first principal component vector direction, the data are almost divided to two categories.

Next using the same sample data as trained data for two categories, I shall illustrate the two-top distributions (See Figure 4). Each category has 10 sample data. Using the above-mentioned logarithmic posterior probability equation, let me calculate the boundary curve on which the two logarithmic posterior probability values are equal. The resultant curve is a quadratic form of x (See Figure 5). In Figure 6, how the two categories are classified by the selected boundary curve is shown.
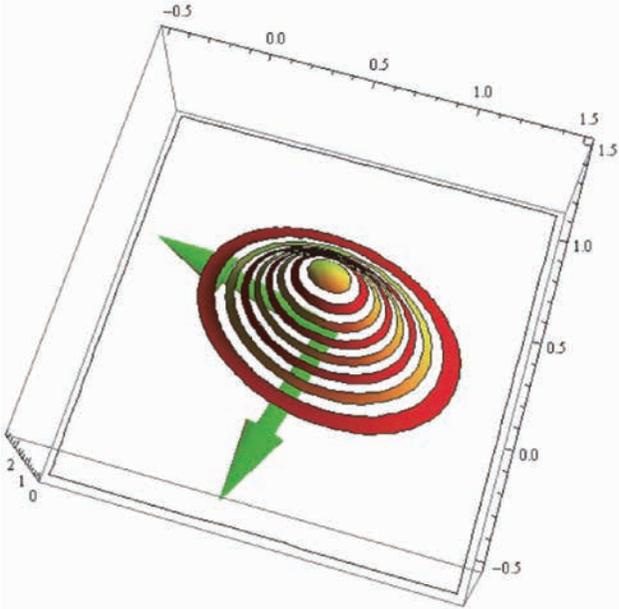

Figure 2: The Gaussian model for the sample data of the two categories WHITE and BLACK. The eigenvectors of the covariance matrix are also shown here. The contour lines show the Mahalanobis distance.
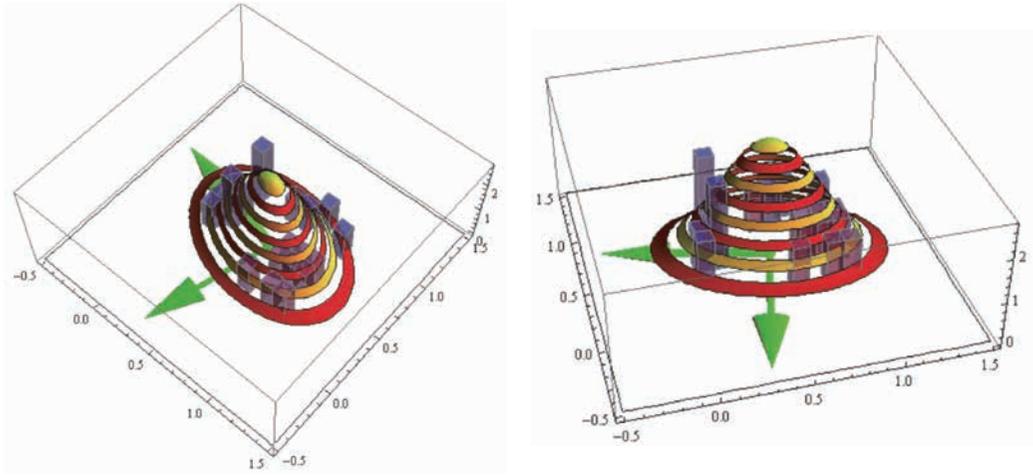
Figure 3: The Gaussian model of the distribution and the two eigenvectors. The two figures are the same one from different point of views.
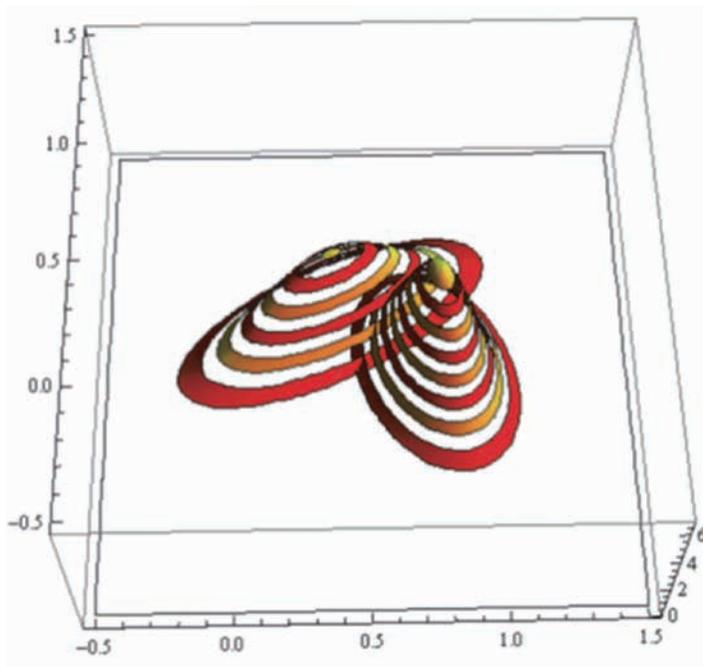

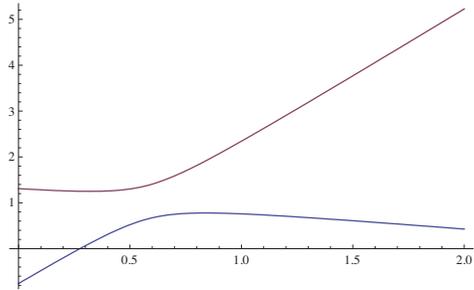Figure 4: The two Gaussian model of the given sample data.

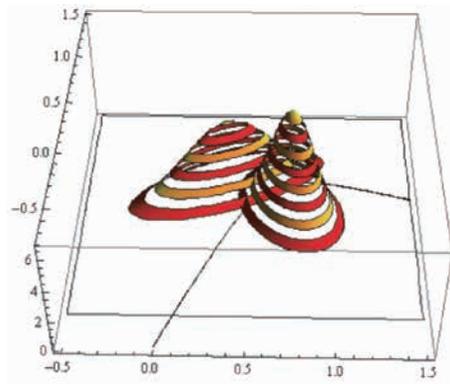Figure 5: The boundary curves which are the solution of the boundary equation.

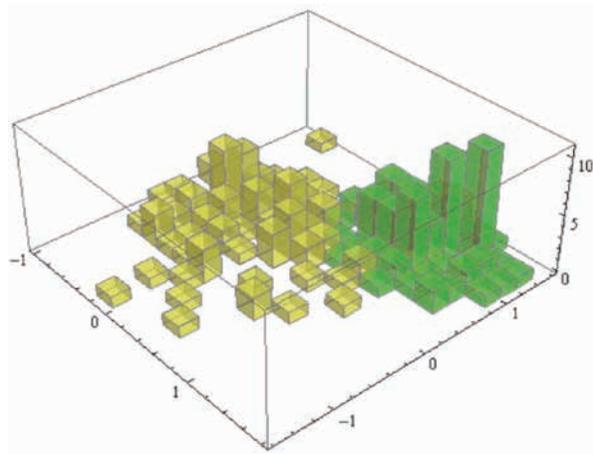Figure 6: The selected boundary curve and the two Gaussian models.

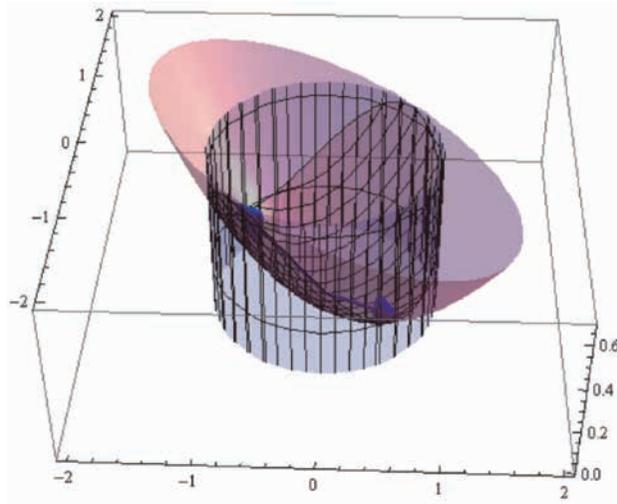Figure 7: The sample data given for the PCA.

Figure 8: The surface here shows the variance of the projected data. The cylinder shows the constraint. The two arrows correspond to the eigenvectors of the covariance matrix.
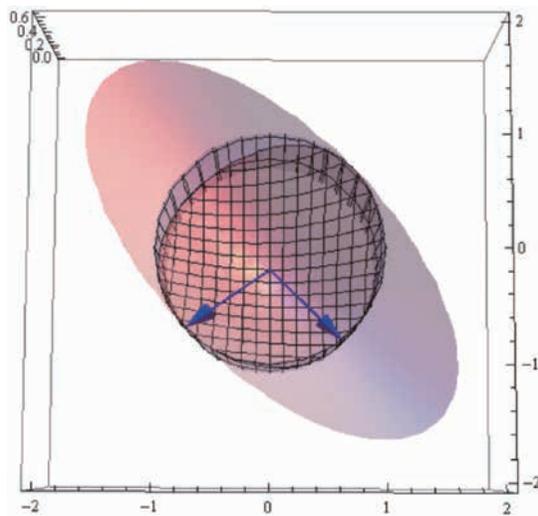

Figure 9: The two eigenvectors show the direction of two principal component directions. They are orthogonal.
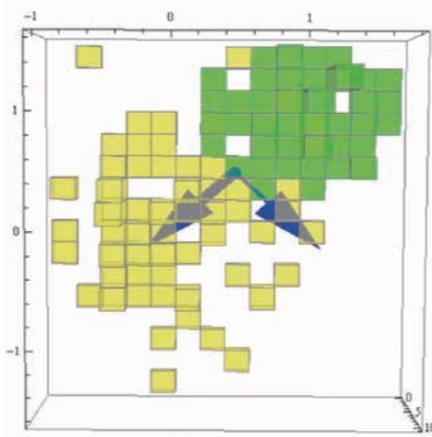
Figure10: The given sample data are divided to categories following the first principal componet direction.

4. Principal Componect Analysis and Latent Semantic Analysis

In the section, I shall describe the PCA and the LSA as its extension. In PCA, we consider the projection onto reduced-dimensional space. The variance of the projected data is expressed as follows (where Z is the projected data of the given data X and Σ represents the covariance matrix of X) :

$$\frac{1}{n-1}Z\,Z^{\mathrm{T}} = \frac{1}{n-1}\ AXX^{\mathrm{T}}A^{\mathrm{T}}$$

We would like to maximize the variance of the projected data. Using the Lagrange multipliers as folows (where Λ is the matrix of Lagrange multipliers):

$$J\ \ = A\Sigma A^{\mathrm{T}}\ \ -\Lambda(AA^{\mathrm{T}}-I)$$

Solving the following equation,

$$\frac{\partial}{\partial\mathbf{A}}J = \mathbf{0}$$

we obtain the relationship $\Sigma A = \Lambda A$.

As shown here, the PCA becomes a problem to find the eigenvalues and eigenvectors of the covarience matrix Σ. The process is visualized as shown in Figure 7 to 10.

Next let me consider LSA. LSA is widely used for a topic extraction in text minng. Analysis of text data starts with compilation of a term-frequency matrix A, counting the occurrencies of each term in each document. The structure of the term-frequency matrix is perused with the purpose of extracting latent components of meaning through Singular Value Decomposition (SDV). A matrix operation that extracts simultaneous least-square principal components of two sets of variables, namely the set of terms and the set of documents, is executed [11].

SVD is an extension of PCA. The PCA finds eigenvalues and eigenvectors of a covariance-matrix

which is symmetrical. However, a tem-frequency matrix is not symmetrical. Therefore, we use SVD instead. By SVD, we obtain $\mathbf{A}=\mathbf{U\Sigma}\ \boldsymbol{V}^{\mathrm{T}}$ where U is the term eigenvectors, V is the document eigenvectors, and $\mathbf{\Sigma}$ is the diagonal matrix of singular values.

Next consider the dimension reduction in LSA visually. For example, suppose that the term-frequency matrix A is given and that after SVD we have $\mathbf{A}=\mathbf{U\Sigma}\ \boldsymbol{V}^{\mathrm{T}}$ where

$$\mathbf{\Sigma} = \begin{bmatrix} 4.09887 & 0. & 0. \\ 0. & 2.36157 & 0. \\ 0. & 0. & 1.27366 \end{bmatrix}.$$

The sample data is cited from [12]. Implement a rank 2 approximation on that by keeping the first two columns of U and V and the first two columns and rows of $\mathbf{\Sigma}$. In the dimension reduction, the SVD is used to estimate what words actually appear in each context and more important components are selected and left.

SVD is an extension of PCA. Therefore, if students have clear images of a PCA process, they will be able to extend the images to ones of SDV. Therefore I think that it is important to have the students grasp the graphical images of PCA.
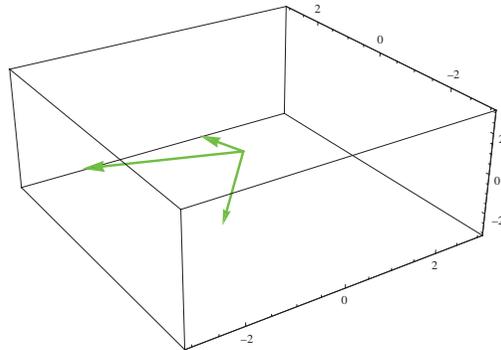

Figure11: The three vectors of V which is the document eigenvectors. The length of each eigenvector is set to be the corresponding eigenvalue here.
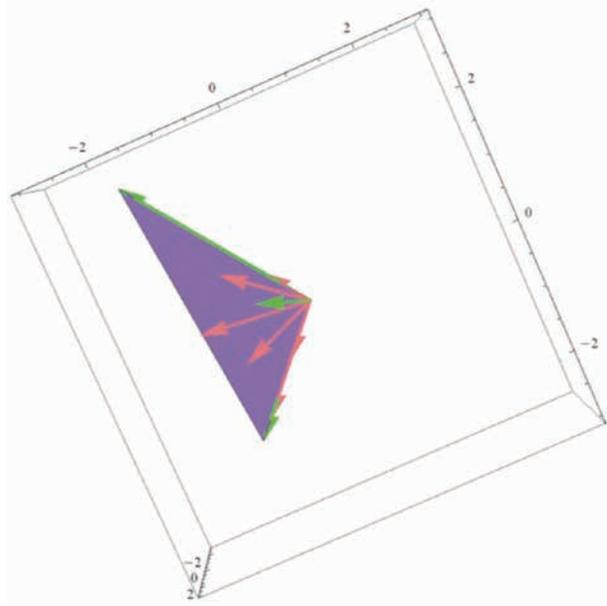
Figure12: The pink vector shows each row of a matrix U Σ which is a term-component matrix. The three green (pale) vector show three vectors of V. Here, the projection to the plane with the first and second component vectors can be seen.
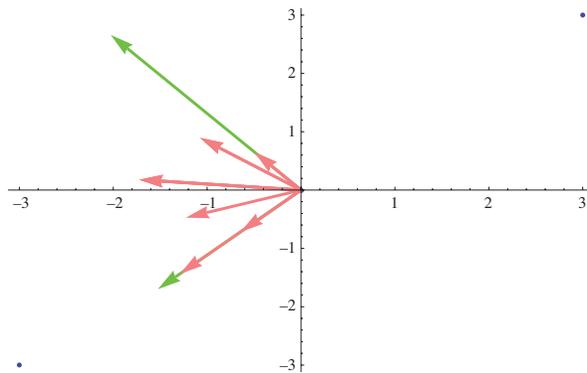

Figure13: After the rank 2 approximation, the 11 vectors in U which is the term eigenvectors are projected as shown here. The green (pale) two vectors corresponding to the first two eigenvectors of V.

5. Conclusion

In the paper, three requirements for a new teaching method of linear algebra are described. Now we have to handle vast amounts of data, for example the data on web. Then, a mathematical tool is an essential tool. By that, we will be able to execute the complicated matrix operations such as the Cholesky

Decomposition and the SVD. Now, a textbook of linear algebra also needs to be changed. There, a new mathematical notation is needed and a matrix expression should be described in a vector notation or a matrix form so that we can read them easily. The matrix expressions include differentiation operations for example the Jacobian and the Hessian matrix. The formulas taught in linear algebra should also be changed. For example, formulas to handle the logarithm of determinant operations should be taught explicitly. In addition, the concept of the invariance of eigenvalues should explained repeatedly because that is the most important concept in linear algebra. The textbook should include many graphical images so that students can easily understand the meanings of the operation processes. Furthermore I recommend students input the programme code of the mathematical tool and verify the same resultant data as the ones illustrated in the textbook. Reading the textbook and input the expressions with the mathematical tool is an effective instructive method. Continuously I will study and contrive an easy-to-understand teaching method.

ACKNOWLEDGEMENT

References

[1]   A. Innocenti and A. S. (eds), *Neuroscience and the Economics of Decision Making*: Routledge, 2012.

[2]   S. Stojanovic, *Computational financial mathematics using Mathematica : optimal trading in stocks and options*. Boston: Birkhäuser 2003.

[3]   T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second ed)*: Springer, 2009.

[4]   M. Sugiyama, T. Suzuki, and T. Kanamori, *Density ratio estimation in machine learning*: Cambridge University of Press, 2012.

[5]   C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2006.

[6]   Y. Shirota and S. Suzuki, "Visualization of the Central Limit Theorem and 95 Percent Confidence Intervals," *Gakushuin Economics Papers*, Vol.50, 2014.

[7]   Yukari Shirota, T. Hashimoto, and S. Iitaka, 感じて理解する数学入門 *"Introduction to Financial Mathematics" (e-Book written in Japanese)* O'Reilley JAPAN, 2012.

[8]   M. Sugiyama, *Statistical Machine Learning (written in Japanese)*. Tokyo: Ohmsha, 2009.

[9]   S. Mitsunari, *Learning Pattern Recoginiton and Machine Learning (written in Japanese)*: Ankoku-tsushindan（暗黒通信団）, 2012.

[10] M. Tsujitani and K. Takezawa, *Machine Learning (written in Japanese)*. Tokyo: Kyoritsu, 2009.

[11] Evangelopoulos, N. and L. Visinescu, *Text-Mining the Voice of the People*. Communications of the ACM, 2012. 55(2): p. 62-69.

[12] Grossman, D.A. and O. Frieder, *Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) (2nd Edition)*: Springer, 2004.

Appendix: **Enlargement Factors of the Magnification Machine are Eigenvalues**

Long time ago, there was an animal kingdom. The five small countries named CENTRE, SOUTH, EAST, NORTH, and WEST are segments of the animal kingdom. As you know well, the space in which we live is three dimensional. However, suppose that the animal kingdom is two dimensional, for simplification of the following description. In the two dimensional world, the world form is a plane. There is no concept of the vertical height. They cannot feel the height, as we cannot feel the fourth dimension. (Please do not think of details of the two dimensional world. I adopted and used a two dimensional world just because of simplification of the matrix description.) As the centre of the kingdom, there is a country named CENTRE which is also the cultural central hub. So many foreign students from other countries go to CENTRE to study innovative technologies there. Here in CENTRE, a genius dog named Mr Orthom born in CENTRE and a genial cat named Ms Diagonary from WEST study together in University of CENTRE.

Now let me explain a bit a basis of a vector space to kindle your memory about that. In linear algebra, a basis is a set of linearly independent vectors which define a coordinate system. Given a basis of a vector space, every element of the vector space can be expressed uniquely as a finite linear combination of basis vectors. Let's consider a basis of CENTRE. A basis of CENTRE is an orthogonal basis: namely the two basis vectors are mutually orthogonal. On the other hand, the basis of WEST is diagonal.

Figuratively speaking, Mr Orthom wears the orthogonal glasses an Ms Diagonary wears the diagonal glasses. So when they express the same point, they represent the point by different coordinates. This is because the bases they use are different. The coordinates are expressed in this world by an ordered pair (a, b) of real numbers. Although you may be anxious about their gathering, there is no trouble. Even if the coordinates are different ones, they can get together at the spot. (Skip the explanation.)

One day, after many improvements, Mr Orthom accomplished the magnifying machine of which enlargement factors can be defined independently for each basis vectors. The matrix is expressed as a diagonal matrix. The convenient point of the M machine is its constant magnification operation which is called scaling, not shearing. The technology that they eagerly wanted was scaling.

Hearing this great news, Ms Diagonary rushed to his laboratory to watch the demonstration. In the demonstration, the input image is a face of Ms Diagonary. Her face image is here scaled three times and twice orthogonally. Ms Diagonary felt proud to be a friend of Mr Orthom. At the same time, she got a bit cross to see her enlarged face. She delivered a monologue "I don't gain a weight." Then Ms Diagonary tried to conduct this demonstration. But the transformation was not done well, sheared and rotated. She said "That was not a scaling. What is the trouble reason?"

After several minute thinking, Mr Orthom found the reason. Because the basis vector of the WEST country is different, the M machine did not work when Ms Diagonary input her coordinate-oriented image. Hearing the reason, Ms Diagonary asked him to change the basis vectors for her country WEST. Mr Orthom complied with her request. He said "We can solve the problem taking the following steps." (Skip the explanation.)

Mr Orthom said to Ms Diagonary, "This matrix is the Magnification Machine for the WEST country. In other words, this is the M Machine with the interpretation function for the WEST country." "Thank you so much, Mr Orthom." She said to Mr Orthom. Then, she soon tried the magnification using this new version machine. This time, the transformation has been succeeded. Now then, we will verify the correctness by the eigenvalues of the matrix B. (I skipped over the explanation.) In conclusion, the enlargement factors 3 and 2 of the M machine are

corresponding to the eigenvalues of the magnifying transformation matrix. Figuratively, let's consider the eigenvectors of the matrix corresponding to the M Machine with interpretation function. Then, the eigenvectors are to be each country's basis vectors.

The change-of-basis changes the matrix and the eigenvectors. But the eigenvalues are invariant.

Mr Orthom customized the M Machine for the WEST. Then, for exports of the M Machine to other countries, customized interpretation functions have to be added. In other words, we have to find the change-of-basis matrix from the CENTRE country to the target country. This is the end of the story.

<div align="right">The END</div>