SVD movies and plots for
*Singular Value Decomposition and its Visualization*

Lingsong Zhang
January 26, 2007

**Notions:**

- **SSVD:** Simple SVD, i.e., applying SVD with the original data matrix $X$.
- **CSVD:** Column SVD, i.e., applying SVD after removing the column mean of $X$.
- **RSVD:** Row SVD, i.e., applying SVD after removing the row mean of $X$.
- **DSVD:** Double SVD, i.e., applying SVD after removing the double mean of $X$.
- **OSVD:** Overall SVD, i.e., applying SVD after removing the overall mean of $X$.

The pdf version of this website is available now. ([the pdf file](#))

This website is designed to provide the MATLAB functions for Zhang et al. (2006), along with the related plots and movies. The following materials are arranged in several sections, including the MATLAB functions, toy examples and the real applications. Further comments are discussed at the end.

- Abstract and Introduction part of Zhang et. al (2006)
- Presentation slides at JSM, 2005, Minneapolis, MN
- Tiny manuscript, which won the 2005 Student Paper Award
- Long manuscript, as the technical report online (cite as Zhang et al, 2006 Long)
- Long manuscript, as the potential published version (still under revision) (short manuscript) (cite as Zhang et al, 2006 Short)

If you want to cite this paper, please use


If you want to cite this website, please use


**MATLAB functions**

This website is designed to provide MATLAB functions for the SVD surface plots, the SVD image plots and two SVD movies (SVD curve movie and SVD rotation movie). In addition, related visualization pictures and movies for Zhang et al. (2006) are also provided in the following sections.
The function, `svd3dplot.m`, can be used to generate the SVD surface plots, the SVD image plots and the SVD curve movie (for a given component) with appropriate options. Use `help svd3dplot` (or `doc svd3dplot`) to check all possible options. For a matrix with large dimensionality (for example, either row number or column number is greater than 100), it is not recommended to generate the curve movie using this function.

The function, `svd3dzoomplot.m`, provides a zoomed version of the SVD curve movie and the SVD surface plots, with appropriate options. If the inputting data matrix has either a large number of rows, or a large number of columns, it is recommended to generate the zoomed version of the SVD surface plots or curve movie instead of the full version.

The function, `svdviewbycomp.m`, provides a SVD rotation movie for a given component (or a subpart of the component). In addition, this function can also be used to generate some conventional Functional Data Analysis (FDA) plots, such as the time series plots for column vectors and row vectors of a given component.

The following personal functions are required for the above SVD visualization functions:

- `svdls.m`: personal SVD function, give a unique sign for the singular columns and rows.
- `matview.m`: personal function for visualization of a data matrix.
- `imagels.m`: personal image visualization for a data matrix.
- `rowmean.m`: personal row mean matrix function.
- `columnmean.m`: personal column mean matrix function.
- `overallmean.m`: personal overall mean matrix function.
- `doublemean.m`: personal double mean matrix function.

See related mathematical details in the paper of Zhang et al. (2006) and the above online technical report.

All the above MATLAB functions can be downloaded together through a compressed file: `svdvisual.zip`. Separate manual files for the above functions will be available soon.

**Analysis of the network traffic data set**

The network traffic data set is collected at the main Internet link of the UNC campus network, as packet counts per half hour over a period of 7 weeks, which covers part of two sessions of UNC summer school in 2003. Detailed analysis are in Section 2 and Section 5 of Zhang et al (2006).

1. Time Series of the network traffic packet counts data set. Bin size is half-an-hour
   This plot shows clear weekly pattern and daily shapes.
   
   Figure 1(a) in Zhang et al (2006 Short and Long).
2. **Mesh plot for the network traffic data set.** The rows are corresponding to the days, while the columns represent half-hour intervals. It also shows the clear weekly pattern and daily shapes. Note that it is more straightforward than the former time series plot.  
   *Figure 1(b) in Zhang et al (2006 Short and Long).*

3. **General Scree plot for the network traffic data set.** It suggests that Simple SVD (SSVD) with two components is the “best” model. However, all surface plots here show that the SSVD model with three components has the best interpretation.  
   *Figure 5 in Zhang et al (2006 Short)*

4. **Surface plots of the SSVD model for the network traffic data set** with three SSVD components, reconstruction and corresponding residual.  
   *Figure 2 in Zhang et al (2006 Short and Long), note that the layout is different.*

5. **Surface plots of the CSVD model for the network traffic data set** with two additional CSVD components.

6. **Surface plots of the RSVD model for the network traffic data set** with two additional RSVD components.

7. **Surface plots of the DSVD model for the network traffic data set** with two additional DSVD components.

8. **Surface plots of the OSVD model for the network traffic data set** with three additional OSVD components.

9. **SVD curve movie for the first three SVD components.**  
   *Figure 3 in Zhang et al (2006 short and Long) shows some carefully chosen snapshots.*

10. **SVD rotation movie for the first three SVD components.**  
    *First SVD, second SVD, third SVD*  
    *Figure 8 in Zhang et al (2006 Short), and Figure 12 in Zhang et al (2006 Long), show some snapshots of these movies.*

11. Scatter plots among $u_1$ vs. $u_2$, $u_1$ vs. $u_3$.  
    *Figure 9 in Zhang et al (2006 short), and Figure 13 in Zhang et al (2006 long), are the scatter plot of $u_1$ vs $u_2$.*

### Analysis of the Chemometrics data set

The Chemometrics data considered here consists of 70 Infrared (IR) spectra of various samples of a polymeric material measured over a 27-day cooling period. Each IR spectra has 1556 measurements representing integers frequency numbers. More details about this data can be found in Marron et al. (2004)

1. **Mesh plot for the Chemometrics data set.**

2. **General Scree plot for the Chemometrics data set** (original plot, log scale plot),

3. **Surface plot for the CSVD of the Chemometrics data set** with 2 more CSVD components.

4. **Surface plot for the RSVD of the Chemometrics data set** with 2 more RSVD components.
5. Surface plot for the SSVD of the Chemometrics data set with 2 more SSVD components.
6. Surface plot for the OSVD of the Chemometrics data set with 2 more OSVD components.
7. Surface plot for the DSVD of the Chemometrics data set with 2 more DSVD components.

The above surface plots show that all decompositions provide similar decomposition. From the generalized scree plot, we will use CSVD as the final model to analyze the data set, which is the same model as in Marron et al (2004).

8. Mesh plot for the first CSVD component.
9. Zoomed Mesh plot for the first CSVD component.
10. Zoomed curve movie for the first CSVD component.

Analysis of the Spanish mortality data set

The Spanish mortality data set is provided by Dr. Andrés M. Alonso Fernández of Departamento de Estadística in Universidad Carlos III de Madrid. This analysis follows the first functional data analysis (FDA) of this data set by him. The data can be accessed from HMD (2005). The data set is a logarithm of the mortality. Each column represents the mortality of a year from 1908 to 2002, and each row is the mortality of a age group from age 0 to age 110. SSVD is a natural choice to analyze the variations of mortality of years and age groups, which is also suggested by the generalized scree plot.

1. Mesh plot for the Mortality data set
   Figure 10 in Zhang et al (2006)

2. General scree plot for the mortality data set (original plot, log scale plot)

3. Surface plot for the Three SVD components of the mortality data set. The first component shows the general decreasing trend in mortality across years; the second component shows the infant effect; and the third component shows some clustering information. Note that these interpretations are more obvious in the image view. In addition, there are some interesting features in the residual part, which is not obvious in the surface view.

4. Image plot for the three SSVD components of the mortality data set
   Figure 11 in Zhang et al (2006)

5. Image plot for the CSVD model with three CSVD components.
6. Image plot for the RSVD model with two RSVD components.
7. Image plot for the DSVD model with two DSVD components.
8. Image plot for the OSVD model with three OSVD components.

Four types of SVD decomposition

The following three examples are designed as $49 \times 48$ matrices, which are the same as the network traffic data set. In this setting, for the example 1 and 3, the rows can be viewed as days, and columns can be treated as times within a day, and the data sets have clearly weekly patterns.
Example 1

The first example is based on the following model,

\[ h_1(i, j) = \mu_c(j) + f_1(i)g_1(j) + \varepsilon(i, j) \]  

(1)

where

\[ \mu_c(j) = \sin \left( \frac{j\pi}{24} \right), \quad g_1(j) = -\cos \left( \frac{j\pi}{24} \right), \quad f_1(i) = \begin{cases} 1 \text{ mod}(i, 7) \neq 0, \text{ and } 6, \\ 2 \quad \text{otherwise}. \end{cases} \]

1. General scree plot for example 1 (original plot, log scale plot).

   The generalized scree plot suggests that the CSVD is the best model among the four types of decomposition in terms of model complexity and approximation performance. However, the surface plots (of these different types) show that the SSVD model provides better interpretation.

2. SSVD for example 1 with 2 SSVD components

   The second row of Figure 6 in Zhang et al (2006 Short)

3. CSVD for example 1 with 1 CSVD component

   The first row of Figure 6 in Zhang et al (2006 Long)

4. RSVD for example 1 with 1 RSVD component

5. DSVD for example 1 with 1 DSVD component

6. OSVD for example 1 with 2 SVD component

   Note that the overall mean for this example is close to zero, so the OSVD model with 2 SVD components are very similar to the SSVD model with 2 components. Detailed discussion about other models are in Zhang et al (2006 short).

Example 2

The model of the second example is set up as following

\[ h_2(i, j) = f_2(i) + f_3(i)g_3(j) + \varepsilon(i, j) \]  

(2)

where

\[ f_2(i) = \cos \left( \frac{i\pi}{24} \right), \quad f_3(i) = \sin \left( \frac{i\pi}{24} \right), \quad g_3(j) = \begin{cases} 1 \quad 1 \leq j \leq 12 \text{ or } 25 \leq j \leq 36, \\ -1 \quad 13 \leq j \leq 24 \text{ or } 37 \leq j \leq 48. \end{cases} \]

Note that the two components \((f_2(i) \text{ and } f_3(i)g_3(j))\) are orthogonal to each other in both the row and the column spaces.

1. General scree plot for example 2 (original plot, log scale plot).

2. SVD for example 2 with 2 SVD components

3. CSVD for example 2 with 2 CSVD component

4. RSV for example 2 with 1 RSV component
5. **DSVD for example 2 with 1 DSVD component**

6. **OSVD for example 2 with 2 SVD component**

Note that the overall mean is also close to 0, thus, the OSVD model is very similar to the SSVD model.

**Example 3**

The model for example 3 is defined as

\[ h_3(i, j) = \mu + f_4(i)g_4(j) + f_5(i)g_5(j) + \epsilon(i, j) \]  
(3)

Where

\[ f_4(i) = \begin{cases} 2 & \text{mod}(i, 7) \neq 0 \text{ and } 6, \\ 0 & \text{otherwise}. \end{cases} \]

\[ g_4(j) = \sin \left( \frac{j \pi}{48} \right), \]

\[ f_5(i) = \begin{cases} 1 & \text{mod}(i, 7) = 0 \text{ or } 6, \\ 0 & \text{otherwise}. \end{cases} \]

\[ g_5(j) = \cos \left( \frac{j \pi}{48} \right), \]

and \( \mu = 5 \). Under this setting, all the elements in the data matrix are greater than 0.

1. General scree plot for example 3 (original plot, log scale plot),

2. **SSVD for example 3 with 2 SVD components**

3. **CSVD for example 3 with 1 CSVD component**

4. **RSVD for example 3 with 2 RSVD component**

   *Figure 7 in Zhang et al (2006 short)*

5. **DSVD for example 3 with 1 DSVD component**

6. **OSVD for example 3 with 2 SVD component**

Note that after removing the overall mean, the first SVD here is similar to the first CSVD component, while the second SVD here is similar to the column mean. The RSVD model provides the best separation of the curves, as discussed in Zhang et al (2006 short).

**Further work**

1. R packages for the SVD surface plot, image plot etc. (under development),

2. Surface plot and related visualization methods with regularized SVD,

3. Surface plot and related visualization methods with pre-smoothing options,

4. Surface plot and related visualization methods with interpolation options.
Reference

- Human Mortality Database, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de
