CSIC 5011 / MATH 5473 Project 1
Story Exploration of Mammal Data

Yingshu CHEN
Department of Computer Science and Engineering
ychengw@connect.ust.hk

Wing Hei SUM
Department of Industrial Engineering and Decision Analytics
whsumaa@connect.ust.hk

Ho Pan IP
Department of Mathematics
hpipab@connect.ust.hk

Zipeng WU
Department of Physics
zwubp@connect.ust.hk

Abstract

Dimensionality reduction approaches are widely used in data exploration particularly high dimensional data. We conduct experiments on a small mammal dataset using typical dimensionality reduction approaches such as PCA, LLE, Modified LLE, MDS, t-SNE, etc. to get low dimensional data visualization. Reduced dimensional data visualization provides researchers with an intuitive way to display interrelation among high-dimensional features. From our visualized results, some implicit elements that may affect mammals' sleeping hours and life span are revealed. A simple experiment using PCA reduced dimensional data for logistic regression keeps equivalent accuracy shows potential of dimensional reduction in boosting data exploration. It is believed the our work can be a simple exemplar to explore other dataset via data dimensionality reduction. See detailed implementation and complete results in `https://github.com/chenyingshu/csic5011_project1`.

1 Background

There are animal mysteries behind interrelation among animals' sleeping time, life time, and their ecological and constitutional factors. To start with a mammal dataset of 62 mammalism species with 10 attributes, we explore the relationship among ecological and constitutional variables, and sleeping hours and lifespan. Typical dimensionality reduction approaches are utilized to obtain main principal components with particular attributes and visualize the data in terms of sleeping hours and life span. Various visualization and revealed patterns behind the data can provide researchers with more intuitive observations over high-dimension data.

2 Problem Definition

From the dataset "*Sleep in Mammals*" from Allison and Cicchetti (1976), we investigate 10 potential factors X_0 for sleep and life-span characteristics, such as constitutional variables, i.e., body weight, brain weight, life-span, gestation time, environmental or ecological variables i.e., sleep exposure level, overall danger level, predation index (level of being predated), and different sleeping hours including slow wave sleep, dream sleep and

CSIC 5011 / MATH 6437 Project 1 (NIPS 2016 template).

total sleep[1]. Let label set $Y = \{\{slowWaveSleep, dreamSleep, sleep\}, life\}$, and according variables $X = X_0 - y, y \in Y$. For example, when we explore sleep hours, we get high dimensional attributes $X = X_0 - \{slowWaveSleep, dreamSleep, sleep\} = \{body, brain, life, gestation, predation, sleepExposure, danger\}$. Here shows some sample data in Table 1.

Give target label $y \in Y$ and corresponding variables X, we aim at low dimensional representations $V(X, y)$ of cleansed data (without missing value), which embed condense features in the life of mammals. In this report, low-dimension (2 to 3 dimensions) visualized figures are obtained from dimensionality reduction methods for intuitive illustration of correlation of tagert label (e.g., sleep habits) and other attributes (e.g., danger levels). Refer to Section 3.1 for more implementation details.

Table 1: Sample Data

species	slowWaveSleep	dreamSleep	sleep	body	brain	life	gestation	predation	sleepExposure	danger
Arctic_Fox	NaN	NaN	12.5	3.385	44.5	14.0	60.0	1	1	1
Tree_hyrax	4.9	0.5	5.4	2.0	12.3	7.5	200.0	3	1	3
Vervet	9.7	0.6	10.3	4.190	58.0	24.0	210.0	4	3	4

3 Experiments

3.1 Implementation Details

3.1.1 Data Cleansing

Before the data are analysed, data quality check is performed and data are preprocessed through data cleansing to improve the data quality for further analysis. The given dataset is incomplete in the sense that there are actually some missing values (NaN), which could be handled through pandas data frame built in functions. All entities with missing values with corresponding features would be filtered out before analysis.

3.1.2 Exploratory Data Analysis

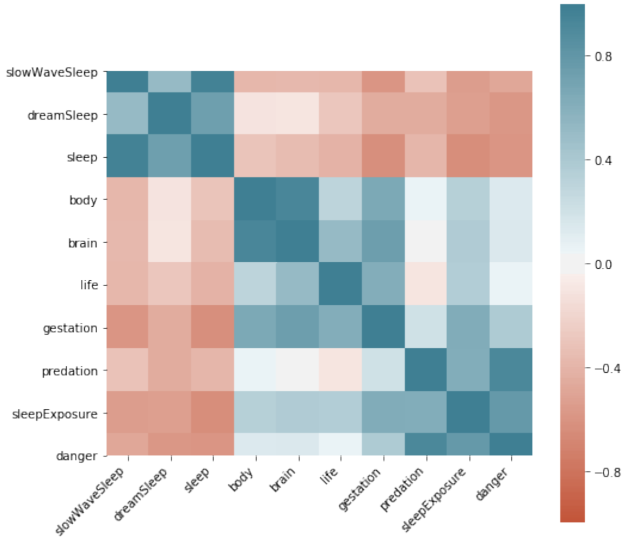


Figure 1: Heatmap of Pearson Correlation

We first explore our data through pairwise feature correlational analysis. Notice in the heatmap (Figure 1) that some features have higher correlation with features that describe similar characteristics,

[1] Slow wave sleep plus dream sleep equals total sleep.

such as ’sleep’ and ’slowWaveSleep’, shown with the darker blue color. Therefore, it makes sense to group those features with higher correlation together for further exploration and analysis in different angles to draw interesting conclusions, with the application of dimensionality reduction algorithms. Based on these findings, the data are processed with grouping by features in some categories, by grouping ’body’, ’brain’, ’gestation’ as size factors, ’slowWaveSleep’, ’dreamSleep’, ’sleep’ as sleep factors, and ’danger’,’predation’ ’sleepExposure’ as danger factors.

3.1.3 Methods

We utilize the *scikit-learn*, *pandas*, *numpy*, *matplotlib* python libraries in this project. Methodologies for dimensionality reduction we use include PCA, standard LLE, LTSA, Hessian LLE, Modified LLE, Isomap, MDS, Spectral Embedding, and t-SNE. The reasons why robust PCA and sparse PCA are not used are because those data are the representation of a specific species, without much Gaussian noise since the effects of outliers have been reduced. Also, there does not exists a high enough level of sparsity structure inherent in our dataset, as shown in figure 2, and thus it does not contribute much using the sparse PCA.

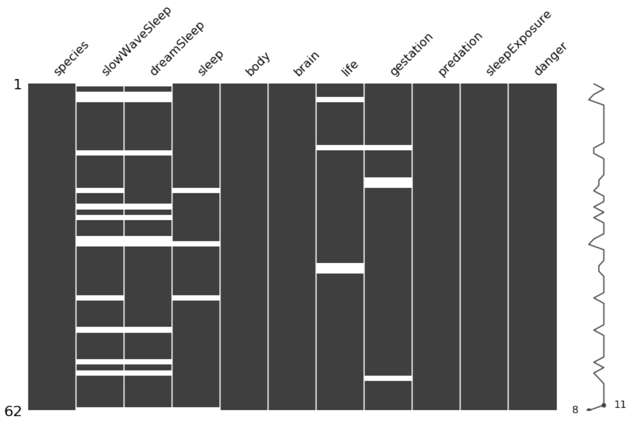


Figure 2: Missing Data Visulization

3.2 Results and Analysis

3.2.1 Sleep Hours

In this section, 7 ecological and constitutional factors are divided into 2 factor subgroups for further relation exploration in terms of sleep hours.

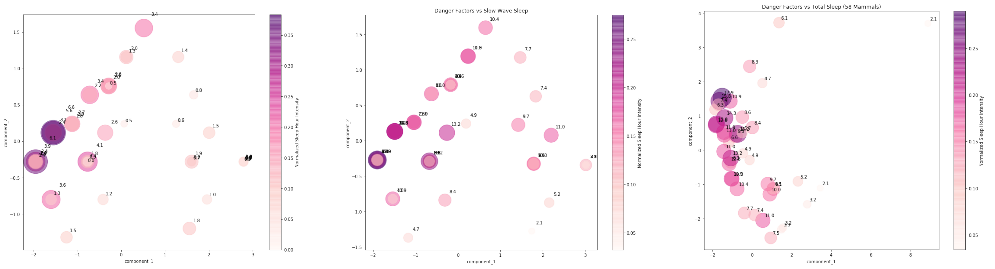
- Danger factors :{’danger’,’predation’ ’sleepExposure’}
- Size factors :{’body’, ’brain’, ’life’, ’gestation’}

We apply PCA and manifold learning methods to each group.

PCA method with the top two components. In this section, we explore the relationship between sleeping hours, and other factors by PCA.

Given danger factors (i.e., predation, sleep exposure, and danger), we compare the effect of these danger factors on the target labels, i.e. dream sleep, slow wave sleep and total sleep hours. From the experiments, the most important principal component explains over 85% of the information to the variables.

In Figure 3 (a) and (b), it shows that the “component 1” is negatively correlated to dream and slow wave sleeping hours. Most dark color and larger nodes are concentrated on the left with negative value of “component 1”. We also investigate the effect of the danger factors on total sleep. The most essential components among the danger factors explain more than 80% information of variables. Compared with slow wave sleep hours trend, dream sleep has a stronger relation to danger factors with more obvious value decrease from bottom left to top right direction in reduced dimensional

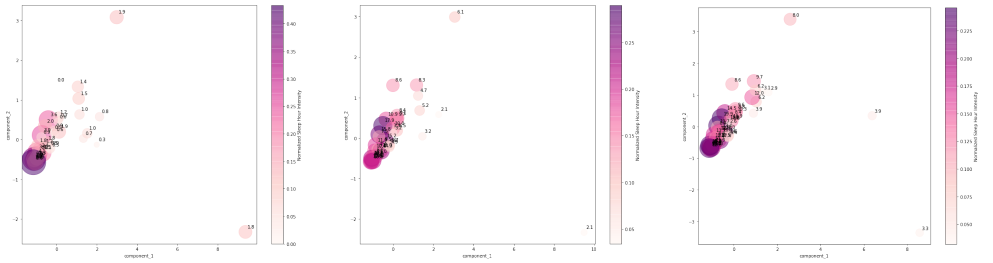


(a) Danger Factors vs Dream Sleep

(b) Danger Factors vs Slow Wave Sleep

(c) Danger Factors vs Total Sleep

Figure 3: PCA analysis for danger factors and sleep hours.



(a) Size Factors vs Dream Sleep

(b) Size Factors vs Slow Wave Sleep

(c) Size Factors vs Total Sleep

Figure 4: PCA analysis for size factors and sleep hours.

space. In Figure 3 (c), mammals with greater value in total sleep have a more negative value in "component 1" and mammals with smaller value in total sleep have a relatively more positive value in " component 1"

For size factors (i.e., body, brain, life, and gestation), we first inspect the correlation between the size factors and different types of sleeping hours. With the two most important principal components, over 90% of information is explained on the size factors. Animals with higher dream sleep hours possess a lower value in both "component 1" and "component 2", and vice versa (see Figure 4). Our results on dream sleep, slow wave sleep and total sleep are close to each other, where the components impose a negative correlation effect on sleep hours (See Figure 4 (a), (b) and (c)). From results in Fig.4, size factors can reveal sleep hours tendency in the low-dimension space, which means size factors and sleep hours have strong relationship.

Manifold learning. In the following analysis, we explore the effect by size factors and all factors in the data set on different sleep variables using manifold learning approaches.

From the Figure 5 and 7, we find that the sleep hours changes following an obvious direction in the graph of LTSA, Isomap, spectralEmbedding, and t-SNE. We can see a strong relation to the tendency of dream sleep hours and total sleep with the size factors. For the slow wave sleep, it still has a similar but not strong tendency to the above two sleep hours (see Figure 6).

More results and analysis about effect of all factors on sleep hours can be checked in Appendix.

3.2.2 Life Span

In this section, we study the effect of ecological and constitutional factors on life span. We divide the 9 factors into 3 factor groups:

- Danger factors :{'danger','predation' 'sleepExposure'}
- Size factors :{'body', 'brain', 'gestation'}
- Sleep factors :{'slowWaveSleep', 'dreamSleep', 'sleep'}

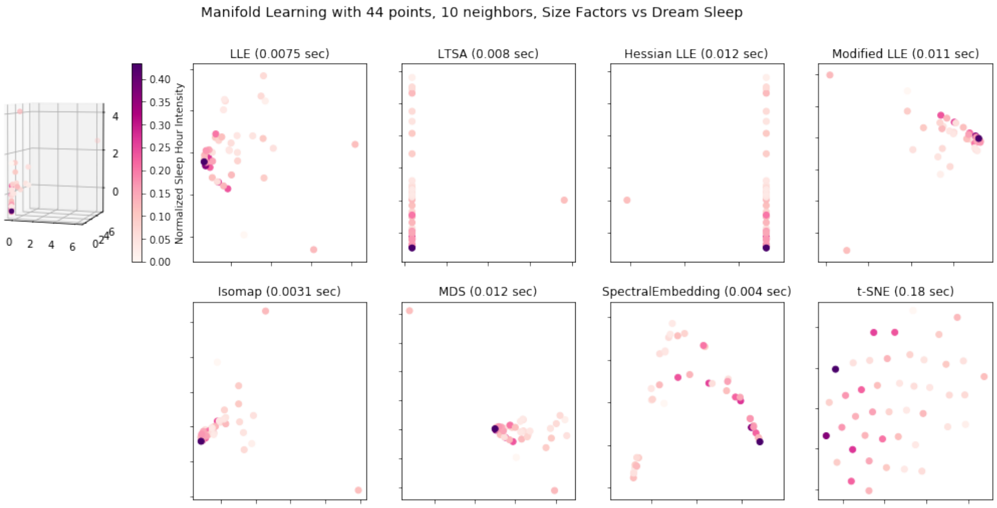


Figure 5: Size Factors vs Dream Sleep

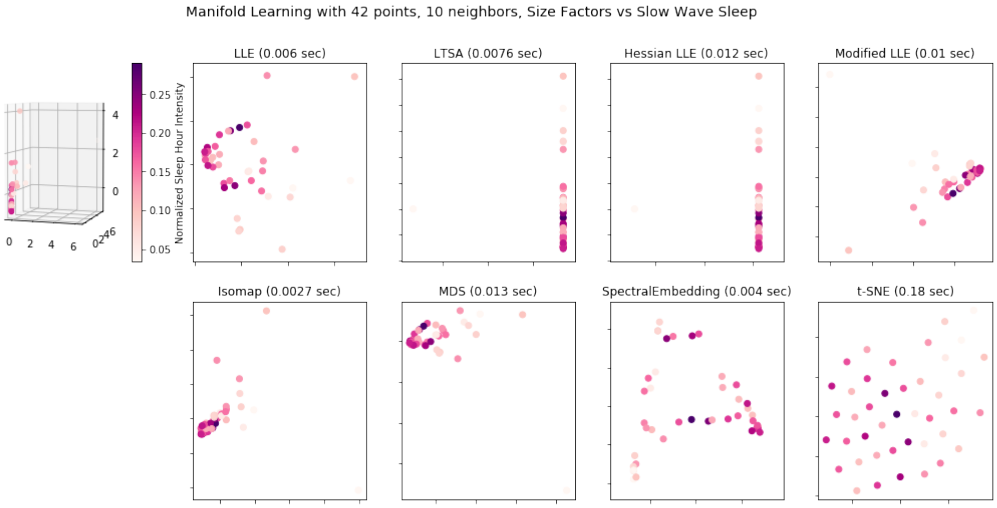


Figure 6: Size Factors vs Slow Wave Sleep

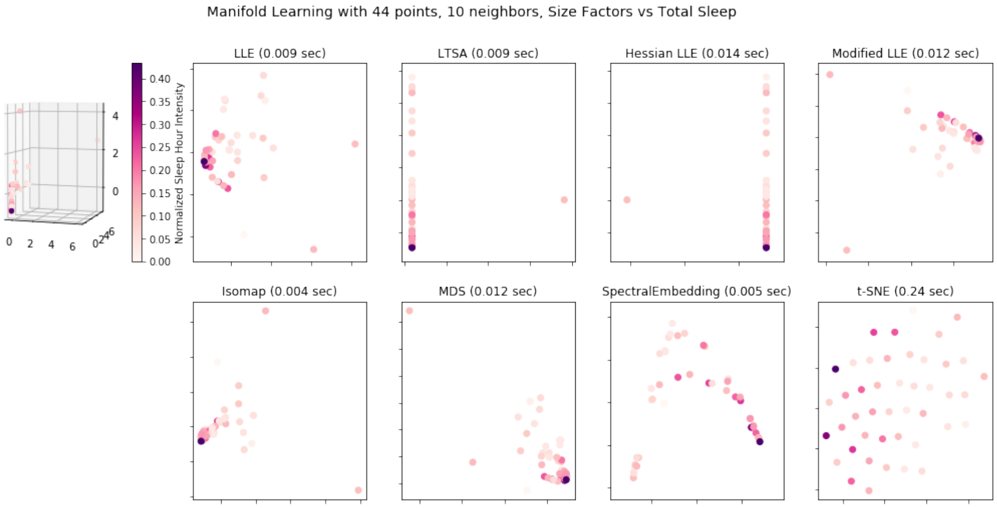
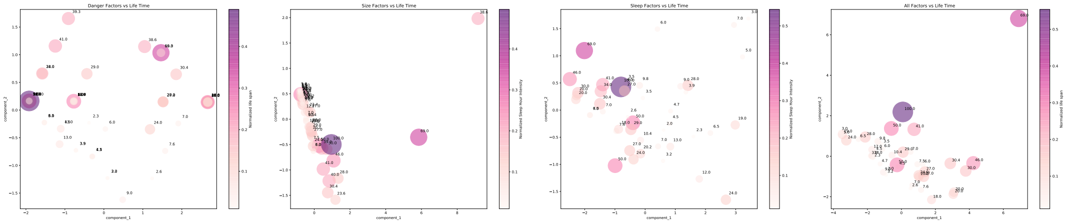


Figure 7: Size Factors vs Total Sleep



(a) danger factors (b) size factors (c) sleep factors (d) all factors

Figure 8: PCA analysis for 3 subgroup factors and all features in terms of life span.

We apply PCA and manifold learning methods to each group.

PCA method with the top two components.

The explained variance ratios of the top two components are [0.8599 0.1262] for the danger factors, [0.8542, 0.1274] for size factors and [0.8313, 0.1687] for the sleep factors. Thus for these three groups, the top two components are sufficient to describe the data.

We also apply PCA to all nine features. The explained variance ratio of the top three components are [0.5404, 0.2320, 0.1140]. It does not perform as well as the subgroup cases.

We further analyze the PCA plot for every group and all features.

Fig.8 (a) shows the PCA for danger factors. It indicates that the animals with smaller “component 2” is prone to have a longer life span. In the meantime, the life span has no direct tendency with the “component 1”.

Fig.8 (b) is PCA result for size factors which shows a narrow distribution for “component 1”. It means that most mammals have a similar value in “component 1”. Further looking into the data, we found out the two outliers in the figure are `African elephant` and `Asian elephant`. It’s reasonable because the size factors of the elephant is quite different from most mammals’, e.g., they have relatively large body weights.

Fig.8 (c) plots PCA for sleep factors. It indicates that the animals with larger “component 1” is prone to have a longer life span. In the meantime, the life span has no direct tendency with the “component 2”.

Fig.8 (d) plots the first two components of all 9 features. From both the figure and explained variance ratios, we can find that only two-dimensional data can poorly represent the data’s underlying structure. This figure shows similarity to Fig.8 (b). In the meantime, the outlier in this figure is also the `Asian elephant`. It is believed that “component 1” and “component 2” both include a great portion of the size factor, making the elephant so unique.

Manifold learning. Fig.9, 10, 11 show the result of manifold learning for danger factors,size factors and sleep factors respectively. The left 3D plot shows how the data points are distributed in the original high dimensional space. Every subplot shows how these data maps to low dimensional space using different manifold learning techniques. We also include the previously discussed PCA results here for comparison.

These three figures show that the data transform by LLE, modified LLE, Hessian LLE, Isomap and MDS have a very similar pattern to the PCA plot. It indicates that for these data, the above methods perform a transformation that is close to PCA. However, the SpectralEmbedding and the t-SNE plot have a somewhat different and precise pattern. In the SpectralEmbedding data, the life span increases as “component 1” of the size factor decrease. On the contrary, the life span has a positive relation to “component 1” of the sleep factor. A similar result can be found in the t-SNE method. We believe that SpectralEmbedding and t-SNE perform a non-linear transformation to the data while unveiling its underlying structure. It’s worth mention that, in SpectralEmbedding and t-SNE plot for the size factor, the `Asian elephant` is no longer an obvious outlier of the sample. Therefore making the non-linear transformation to the size factor data may have some improvement.

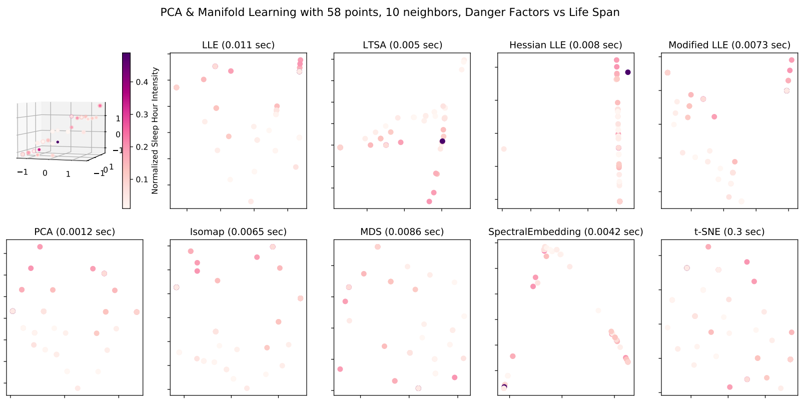


Figure 9: Manifold learning & PCA for the danger factors

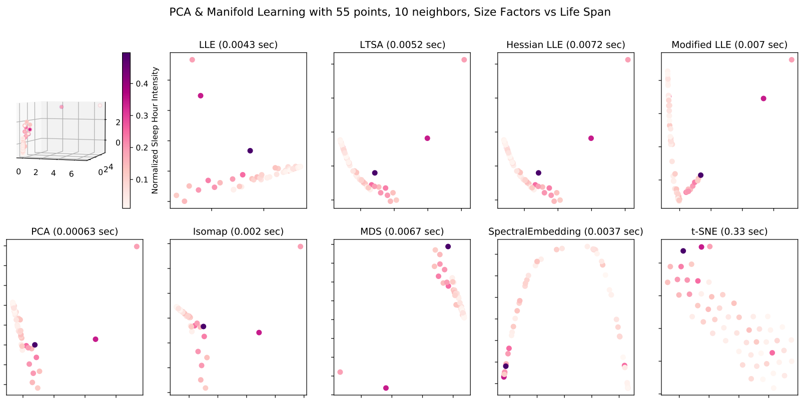


Figure 10: Manifold learning & PCA for the size factors

4 More Potential Application

In the future work, we desire to make use of PCA on other applications. We divide the data set into two different groups which are “long” and “short” total sleeping time. Applying the PCA method, we separate the animals into two different sides in two reduced dimensions with top-2 components (see Figure12 (a)).

Then, we apply the Logistics Regression to train the first 70% data and the remaining 30% data are used to test the accuracy of the prediction. We train the regression model by 2 components and 7 components individually. The R-squared score are the same in both cases, which are around 0.923. It implies that even we reduce the number of components while training the regression model, we still obtain a high accuracy result. We have the same experiment on the “long” and “short” lifetime (see Figure12 (b)). We observe a similar result as the total sleeping time.

We might also want to try different algorithms or models through using the PCA and manifold learning approaches to reduce the complexity in the model. In some larger data sets, it is expected to shorten the running time in the model training process by reducing the number of data dimensions without sacrificing the accuracy.

5 Conclusion

Our mammal data exploration results using diverse dimensionality reduction approaches intuitively discover some interrelations behind the high dimensional raw data. In general, size factors such as body weight, brain weight, gestation time, life time strongly indicate sleeping hours, while danger factors relate more to dream sleep hours. For mammal life span, all ecological and constitutional features (i.e., sleep hours, weights, gestation time, danger levels, etc.) embrace life span implicit

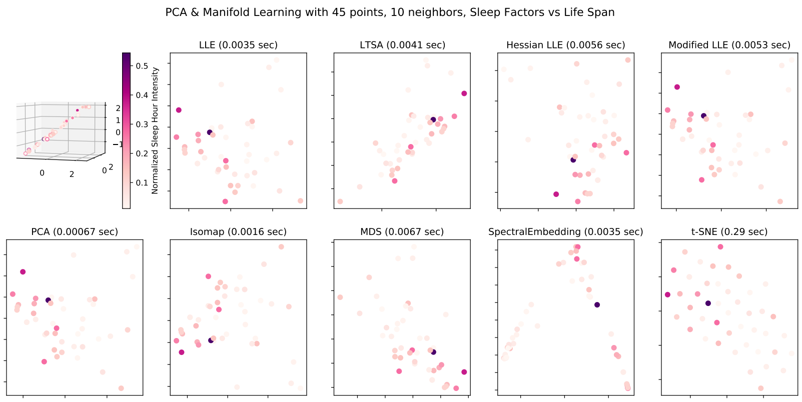
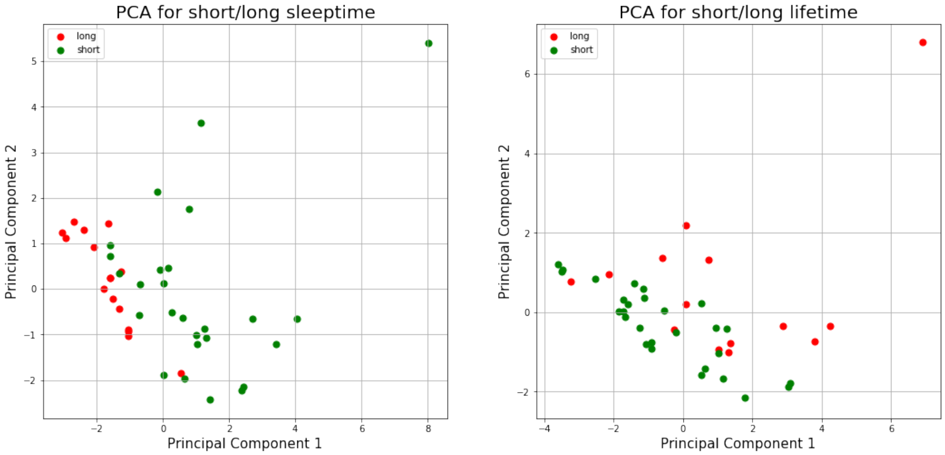


Figure 11: Manifold learning & PCA for the sleep factors



(a) PCA for short/long sleeping time

(b) PCA for short/long lifetime

Figure 12: PCA analysis for short/long sleeping time and lifetime.

conditions. 2 or 3 low-dimensional data visualization of primal high-dimensional raw data provides a straightforward preliminary results to further research exploration. For instance, biologists can directly further investigate the relationship between dream sleep and danger factors in a narrow area.

Dimensionality reduction also give us a hint on data simplification for other application, and thus potentially improve the algorithm performance and boost the research. Our simple experiment in Section 4 shows such a potential that even with reduced dimensional data some algorithms (e.g., logistic regression) still keep relative high accuracy.

References

Allison, T.; Cicchetti, D. V. Sleep in mammals: ecological and constitutional correlates. *Science* **1976**, *194*, 732–734.

Appendix

A Additional Results

Here are some supplementary results to Section 3.2.1. Considering all factors effect on different sleep hours, we have different results on various types of sleep hours. Although the sleep hours are strongly correlated to all other factors, the direction of the sleep hour tendency are not the same to each others. For instance, the intensity of dream sleep hours changes from high to low gives a upward trend (see Appendix). But the intensity of slow wave sleep hours changes from high to low gives a downward trend (see Appendix). Because the dream and slow wave sleep do not have a similar pattern in the analysis, the relation between the total sleep and size factors is also unlike to the the above two types of sleep hours (see Appendix).

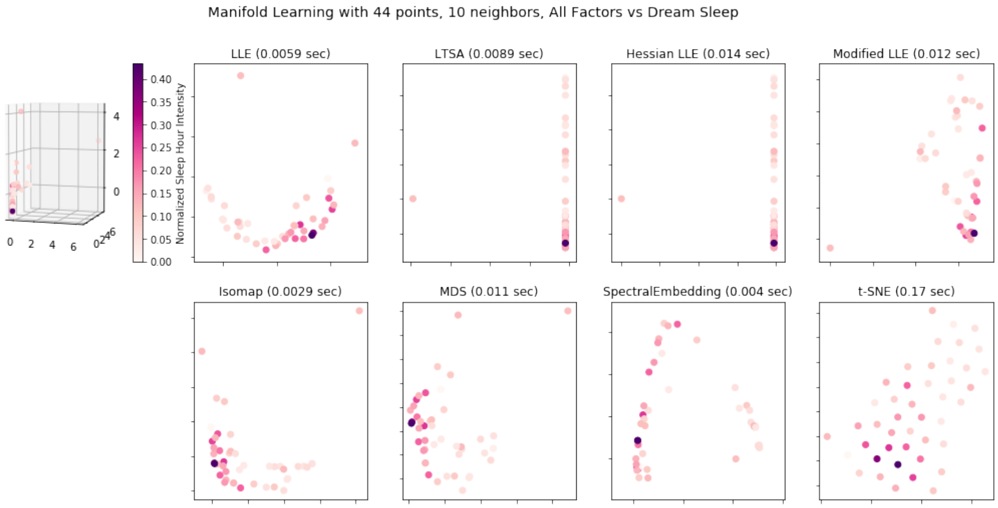


Figure 13: All Factors vs Dream Sleep

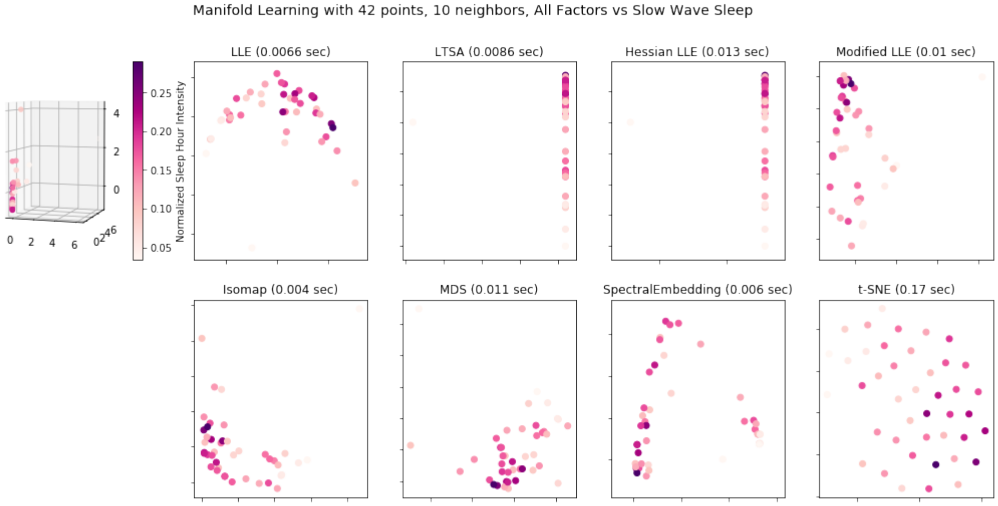


Figure 14: All Factors vs Slow Wave Sleep

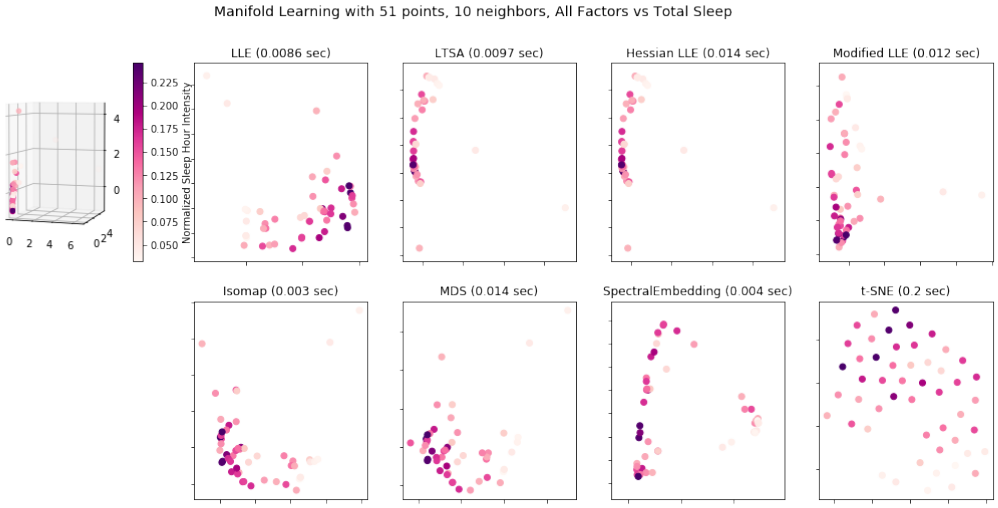


Figure 15: All Factors vs Total Sleep

B Contribution

Yingshu CHEN code implementation and integration, data visualization, report writing (Sec 1, 2, 5).

Wing Hei SUM code implementation, data visualization, report writing (Sec 3.2.1, 4).

Ho Pan IP code implementation, data visualization, report writing (Sec 3.1).

Zipeng WU code implementation, data visualization, report writing (Sec 3.2.2).