5
Some Other Dimension Reduction Methods

Dimensionality reduction techniques beyond PCA are often necessary when the underlying data structure is nonlinear or when variance-based methods like PCA do not capture enough of the essential relationships in the data. The goal of this chapter is to show what are the possible ways to handle non-linearity with preserving some of the computational advantages of linear methods. We do it by introducing two state-of-the-art techniques: Uniform Manifold Approximation and Projection (UMAP). This part of the lecture will be presented on slides and will be relatively high-level.

Exercise 5.5.5. *Let p, q be two probability distributions over some space $\mathcal{X}$. Define the Kullback-Leibler divergence $\mathrm{KL}(p,q) = \mathbb{E}_p \log \frac{p(X)}{q(X)}$. Show that $\mathrm{KL}(p,q) \geq 0$ and it is equal to zero if and only if p, q define the same distribution. Hint: Use the Jensen's inequality and the fact that $-\log$ is a strictly convex function.*

Exercise 5.5.6. *Suppose $X_i \sim \mathrm{Bern}(p_i)$ and $Y_j \sim \mathrm{Bern}(q_j)$ are all independent. Show that the Kullback-Leibler divergence between the distribution p of $X = (X_1, \ldots, X_m)$ and the distribution q of $Y = (Y_1, \ldots, Y_m)$ is given*

KL Divergence Measure the difference between two probability function p, q over a space X

$$\underline{KL(p,q)} = E_p\left(\log \frac{p(x)}{q(x)}\right) = \sum p(x) \frac{p(x)}{q(x)}$$

property ① $KL(p,q) \geq 0$ ✓ = ∫

② $KL(p,q) = 0 \iff p = q$ ✓

Recall, Jensen's Inequality states

for a function f and a R.V Y.

we have $f(E(Y)) \leq E(f(Y))$

equality holds only if $\exists$ a constant c s.t. $Y = c$.

$-\log(x)$ is convex function $f''(x) = \frac{1}{x^2}$.

$$KL(p,q) = E_p\left(\log\frac{p(x)}{q(x)}\right) = E_p\left(-\log\frac{q(x)}{p(x)}\right)$$

Here we Let $Y = \frac{q(x)}{p(x)}$

So that we have $-\log(E_p(Y)) \le E_p(-\log(Y))$

$$\Rightarrow KL(p,q) \ge -\log(E_p(Y)) = 0$$

Exercise 5.5.6. *Suppose $X_i \sim \mathrm{Bern}(p_i)$ and $Y_j \sim \mathrm{Bern}(q_j)$ are all independent. Show that the Kullback-Leibler divergence between the distribution p of $X = (X_1, \ldots, X_m)$ and the distribution q of $Y = (Y_1, \ldots, Y_m)$ is given by the formula*

$$\mathrm{KL}(p,q) = \sum_{i=1}^{m}\left(p_i \log\frac{p_i}{q_i} + (1-p_i)\log\frac{1-p_i}{1-q_i}\right).$$

$$E_p(Y) = \int p(x)\frac{q(x)}{p(x)}dx = 1$$

$$\Rightarrow E_p(-\log(Y)) = E_p(-\log(1)) = 0$$

$$\frac{q(x)}{p(x)} = 1 \iff KL(p,q) = 0$$

given $X_i \sim Ber(p_i)$ $Y_j \sim Ber(q_j)$ are all independent

we have $p(x) = \prod_{i=1}^{n} p(x_i)$ $\quad q(y) = \prod_{j=1}^{m} q(y_j)$

where $p(x_i) = p_i^{x_i}(1-p_i)^{1-x_i}$ $\quad q(y_i) = q_j^{y_j}(1-q_j)^{1-y_j}$

$$KL(p,q) = E_p\left[\log\frac{\prod_{i=1}^{m} p(x_i)}{\prod_{i=1}^{m} q(x_i)}\right]$$

$$= E_p\left[\sum_{i=1}^{m}\log\frac{p(x_i)}{q(x_i)}\right]$$

$$= \sum_{i=1}^{m} E_p\left(\log\frac{p(x_i)}{q(x_i)}\right)$$

$$= \sum_{i=1}^{m} KL(p_i, q_i))$$

$\log(a) + \log(b) = \log(ab)$

where $KL(p_i, q_i) = p_i\log\frac{p_i}{q_i} + (1-p_i)\log\frac{1-p_i}{1-q_i}$

Exercise 7.7.4 (Identifiability in FA). *Show that if (W, Ψ) satisfies the FA model, then for any orthogonal matrix $U \in \mathcal{O}(r)$, the pair (WU, Ψ) defines the same observed distribution for X. What does this imply for parameter estimation?*

FA Model

$$X = \mu + WF + \varepsilon$$

where X is Observed Random Vector

μ is Mean Vector

W factor loading Matrix

F latent vector Factor $F \sim N(0, I_r)$

ε error term $\varepsilon \sim N(0, \phi)$

diagonal Matrix

$$\mathrm{Cov}(X) = W \,\mathrm{Cov}F\, W^T + \mathrm{Cov}(\varepsilon)$$
$$= WW^T + \phi$$

Assume $U \in O(r)$

Let $W' = WU$

$F' = U^T F$ $\quad F' \sim N(0, I_r)$

$$X = \mu + W'F' + \varepsilon$$
$$= \mu + (WU)(U^T F) + \varepsilon$$
$$= \mu + WF + \varepsilon \quad \text{unchanged}$$

$$\mathrm{cov}(X) = W'W'^T + \phi$$
$$= (WU)(WU)^T + \phi$$
$$= WW^T + \phi \quad \text{unchanged}$$

FA is NOT identifiable up to orthogonal transformation of factor loading.

Tut_STA437

Shupeng Chen

2025-03-27

Package

We need the "sm" package for accessing the dataset and the "energy" package for calculating distance correlation, which is a measure of dependence that captures non-linear relationships

```
install.packages("sm")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)

install.packages("energy")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)

library(sm)

## Package 'sm', version 2.2-6.0: type help(sm) for summary information

library(energy)

# Load the aircraft dataset
data(aircraft)

# Extract the relevant variables
X <- aircraft$Span  # Wing span
Y <- aircraft$Speed # Speed

# Compute the Pearson correlation
correlation <- cor(X, Y)
correlation

## [1] -0.01042982
```

Correlation

The correlation is very close to zero (≈−0.01), indicating no significant linear relationship between wing span and speed.

Test for Independence Using Correlation-Based Tests

Perform a hypothesis test for the correlation to confirm whether the observed correlation is statistically different from zero. The p-value is approximately 0.78, which is much greater than the typical significance level (e.g., 0.05). This suggests that we fail to reject the null hypothesis of zero correlation. In other words, there is no evidence of a linear relationship between Span and Speed.

```
# Perform a correlation test
cor_test_result <- cor.test(X, Y)
cor_test_result$p.value

## [1] 0.7816014
```

Compute Distance Correlation

Distance correlation is a measure of dependence that captures both linear and non-linear relationships. It ranges from 0 (no dependence) to 1 (perfect dependence). Use the dcor.test function from the energy package to compute the distance correlation and test for dependence. The p-value is approximately 0.00099, which is much smaller than the typical significance level (e.g., 0.05). This indicates strong evidence against the null hypothesis of independence. In other words, Span and Speed are highly dependent, even though their correlation is close to zero.

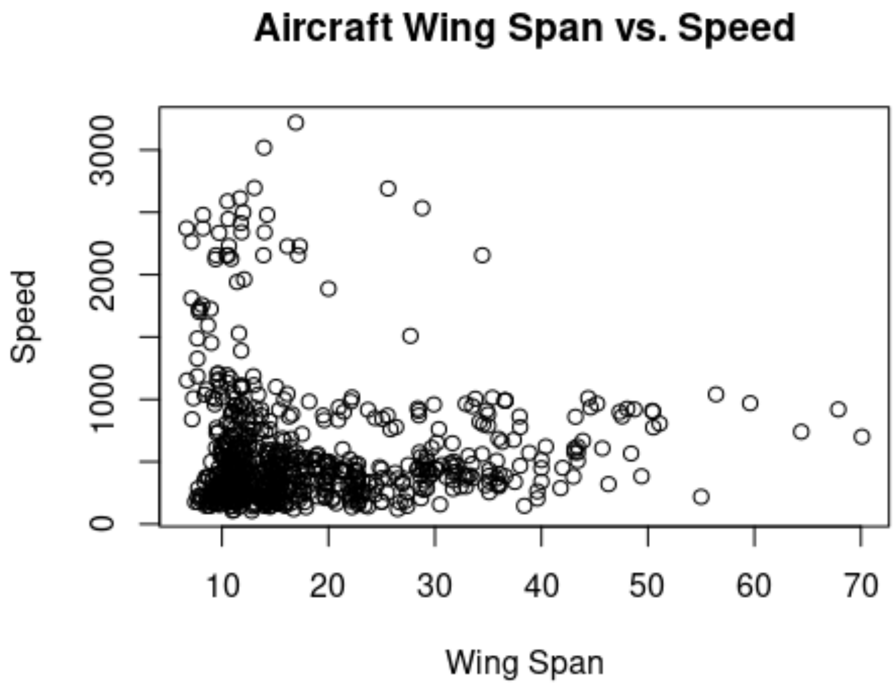
```
# Perform a distance correlation test
distance_correlation_test <- dcor.test(X, Y, R = 1000)
distance_correlation_test$p.value

## [1] 0.000999001
```

Visualize the Data

Plot the data to visualize the relationship between Span and Speed. This will help us understand why the correlation is close to zero but the variables are still dependent. The scatter plot may show a pattern that is not linear. For example, there could be a quadratic or other non-linear relationship. Even though the points do not form a straight line, they may still exhibit a clear structure, indicating dependence.

```
# Plot the data
plot(X, Y, xlab = "Wing Span", ylab = "Speed", main = "Aircraft Wing Span vs.
Speed")
```



Summary

Correlation: The Pearson correlation is close to zero (≈−0.01), indicating no significant linear relationship. Distance Correlation: The distance correlation test has a very small p-value (≈0.00099), indicating strong evidence of dependence. Visual Inspection: The scatter plot may reveal a non-linear pattern, confirming that the variables are dependent in a non-linear way

Conclusion

This example demonstrates that variables can be highly dependent despite having zero correlation. The key takeaway is:

Zero correlation does not imply independence