

Biplot Analysis

Introduction

- A **biplot** is a graphical representation of the information in an $n \times p$ data matrix.
- The **bi-** refers to the two kinds of information contained in a data matrix. The information in the rows pertains to samples or sampling units and that in the columns pertains to variables.
- Biplot display of n observations and p variables on the same two-dimensional diagram.
- A two-dimensional plot of the sampling units can be obtained by graphing the first two principal components. The idea behind biplots is to add the information about the variables to the principal component graph.
- The biplot simultaneously gives graphical information about the relationships between variables.
- The relative positions of variables and observations, which are plotted on the same diagram, can also be interpreted.

Constructing Biplot

A direct approach to obtaining a biplot starts from the singular value decomposition (SVD) which first expresses the $n \times p$ mean corrected matrix \mathbf{X}_c as

$$\mathbf{X}_c = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T$$

Where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ and \mathbf{V} is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{X}_c^T \mathbf{X}_c = (n - 1)\mathbf{S}$. That is, $\mathbf{V} = \hat{\mathbf{E}} = [\hat{\mathbf{e}}_1 \quad \hat{\mathbf{e}}_2 \quad \dots \quad \hat{\mathbf{e}}_p]$.

$$\mathbf{V}^T \hat{\mathbf{E}} = \hat{\mathbf{E}}^T \hat{\mathbf{E}} = \mathbf{I}$$

$$\mathbf{X}_c \hat{\mathbf{E}} = \mathbf{U} \mathbf{\Lambda}$$

where the j th row of the left-hand side,

$$\left[(\mathbf{x}_j - \bar{\mathbf{x}})^T \hat{\mathbf{e}}_1 \quad (\mathbf{x}_j - \bar{\mathbf{x}})^T \hat{\mathbf{e}}_2 \quad \dots \quad (\mathbf{x}_j - \bar{\mathbf{x}})^T \hat{\mathbf{e}}_p \right] = [\hat{y}_{j1} \quad \hat{y}_{j2} \quad \dots \quad \hat{y}_{jp}]$$

is just the value of the principal components for the j th item.

That is, $\mathbf{U} \mathbf{\Lambda}$ contains all of the values of the principal components, while $\mathbf{V} = \hat{\mathbf{E}}$ contains the coefficients that define the principal components.

The best rank 2 approximation to X_c is obtained by replacing Λ by

$$\Lambda * = \text{diag}(\lambda_1, \lambda_2, 0, \dots, 0)$$

This result, called the Eckart-Young theorem, the approximation is then

$$X_c = U \Lambda * V^T = [\hat{y}_1 \quad \hat{y}_2] \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \end{bmatrix}$$

where \hat{y}_1 is the $n \times 1$ vector of values of the first principal component and \hat{y}_2 is the $n \times 1$ vector of values of the second principal component.

In the biplot, each row of the data matrix, or item, is represented by the point located by the pair of values of the principal components.

The i th column of the data matrix, or variable, is represented as an arrow from the origin to the point with coordinates $[\hat{e}_{1i} \quad \hat{e}_{2i}]$, the entries in the i th column of the second matrix $[\hat{e}_1 \quad \hat{e}_2]^T$.

Biplot the first two principal component

The best two-dimensional approximation to the data matrix X approximates the j th observation x_j in terms of the sample values of the first two principal components. In particular,

$$x_j \cong \bar{x} + \hat{y}_{j1}\hat{e}_1 + \hat{y}_{j2}\hat{e}_2$$

\hat{e}_1 and \hat{e}_2 are the first two eigenvectors of S or, equivalently, of $X_c^T X_c = (n - 1)S$ and X_c denotes the mean corrected data matrix with rows $(x_j - \bar{x})^T$.

- The coordinates of the j th unit (row) are the pair of values of the first two principal components, $(\hat{y}_{j1}, \hat{y}_{j2})$.
- These eigenvectors \hat{e}_1 and \hat{e}_2 are the coefficient vectors for the first two sample principal components.

Interpretasi Biplot

Interpretasi posisi Variabel (vector):

- Panjang Vector dari suatu variable menggambarkan kontribusinya terhadap pembentukan komponen utama.
- Sudut yang terbentuk dari setiap vector terhadap sumbu axis komponen utama menggambarkan semakin lancip semakin kuat korelasi variable dengan komponen utamanya.
- Sudut antar vector yang semakin lancip menggambarkan adanya korelasi yang semakin tinggi antar variable dan positif.

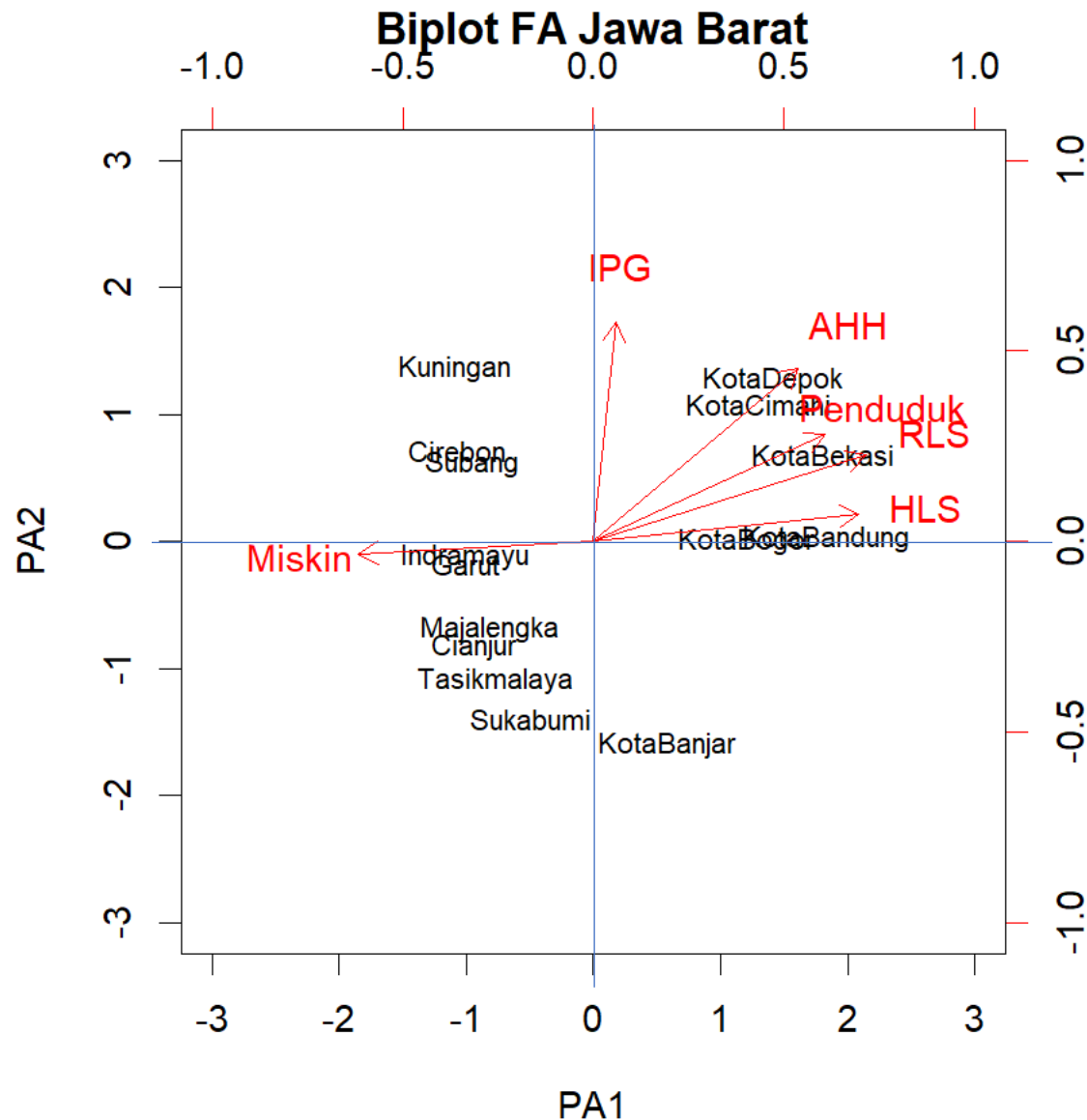
Interpretasi posisi unit observasi (titik koordinat):

- Titik-titik yang berdekatan menunjukkan bahwa unit-unit tersebut memiliki kesamaan/kemiripan karakteristik
- Karakteristiknya dapat dilihat dari jarak titik ke vector. Semakin dekat menunjukkan bahwa unit observasi tersebut memiliki nilai yang tinggi pada variable tersebut

Keragaman Data yang dijelaskan biplot:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i}$$

Contoh biplot analisis factor untuk data indicator social ekonomi di Provinsi Jawa Barat



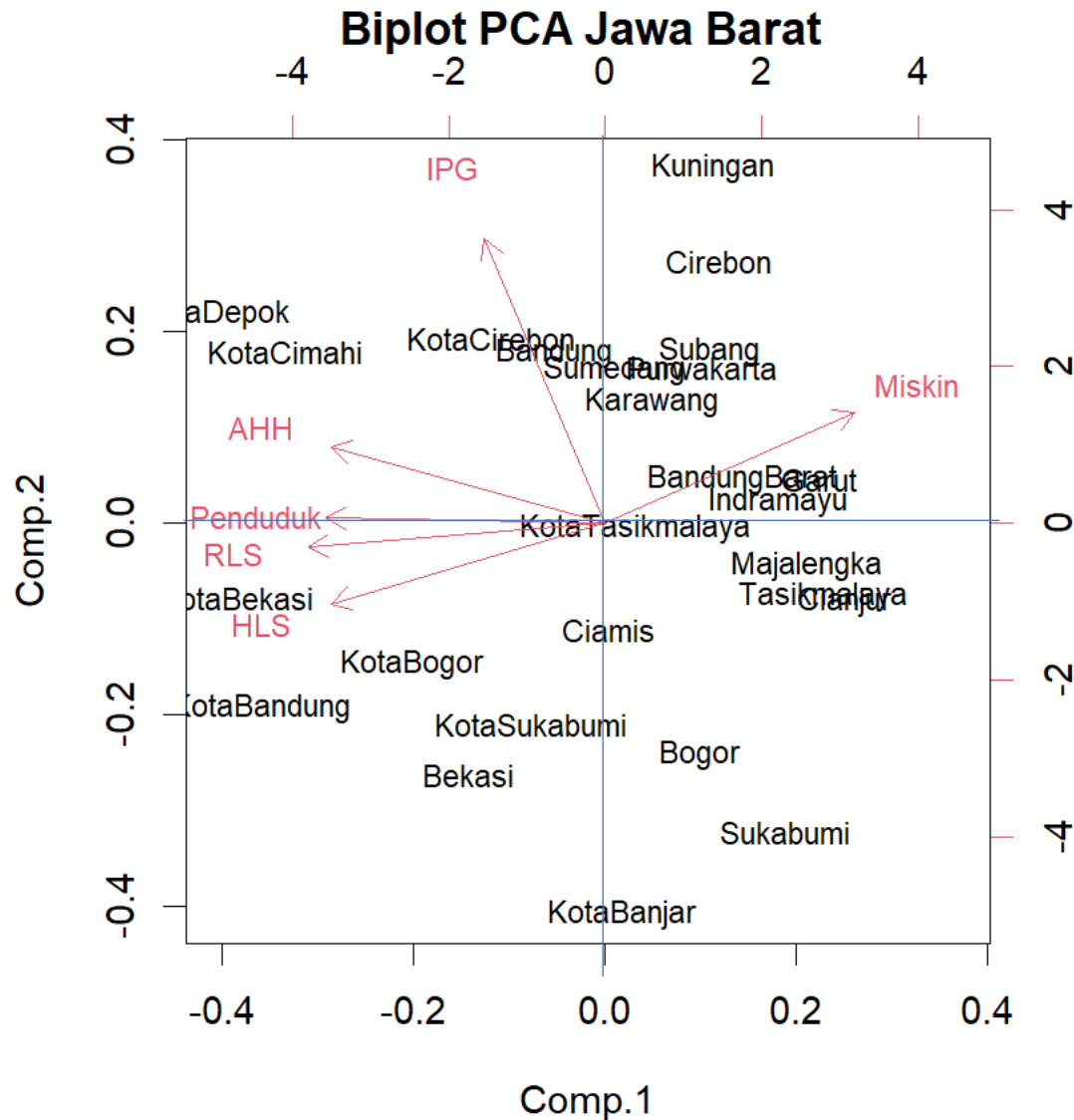
Faktor 1 berkorelasi erat dengan kemiskinan, HLS, RLS dan Kepadatan Penduduk

- Variabel HLS, RLS dan Kepadatan Penduduk berhubungan erat dan positif.
- Variable kemiskinan berhubungan negatif dengan variable lainnya

Faktor 2 berkorelasi erat dengan IPG

- Kota Bogor, Kota Bandung, Kota Bekasi, Kota Cimahi dan Kota Depok memiliki karakteristik yang hampir sama, tergolong daerah dengan kesejahteraan/kualitas hidup yang lebih baik (HLS, RLS & AHH tinggi, sedangkan kemiskinannya rendah) dibanding kelompok kab/kota lainnya.
- Kota Cimahi dan Kota Depok selain AHH yang tinggi juga memiliki IPG yang tinggi.
- Kota Banjar dan Kab. Sukabumi IPG relative lebih rendah kab/kota lainnya namun kesejahteraannya sedang.

Contoh biplot PCA untuk data indicator social ekonomi di Provinsi Jawa Barat



PC1 berkorelasi erat dengan kemiskinan, HLS, RLS dan Kepadatan Penduduk

- Variabel HLS, RLS dan Kepadatan Penduduk berhubungan erat dan positif.
- Variable kemiskinan berhubungan negatif dengan variable lainnya

PC 2 berkorelasi erat dengan IPG

Karakteristik kab/kota:

Dapat kita bagi menjadi kelompok daerah dengan tingkat kesejahteraan sosial-ekonomi tinggi, sedang dan rendah

Selain itu ada yang IPG tinggi, sedang dan rendah.

Interpretasi titik (karakteristik setiap daerah)

	Kesejahteraan Sosial-Ekonomi		
	Rendah	Sedang	Tinggi
IPG Rendah	-	Kota Sukabumi, Bekasi, Kota Banjar, Bogor, Sukabumi	
IPG Sedang	Subang, Purwakarta, Bandung Barat, Indramayu, Garut, Majalengka, Tasikmalaya, Cianjur	Kota Tasikmalaya, Ciamis	Kota Bekasi, Kota Bogor, Kota Bandung
IPG Tinggi	Kuningan, Cirebon	Kota Cirebon, Bandung, Sumedang, Karawang	Kota Depok, Kota Cimahi