VI-2. Structuration of data.

VI-2-1. Principle component analysis (PCA)

VI-2-1-1. What is PCA

Most simplified technical explanation of PCA is redrawing of scatter diagram from diagram plotted in the coordinate space of observed variables to diagram plotted in the coordinate space of eigenvector. It means reconstruction of picture of a phenomenon from the relation among visible things to relation among invisible latent components. For this moment, we considered that latent components are independent each other, and variables are explained by combination of latent components. Honestly, we do not have any method for confirmation of independency among latent components using only obtained data, and it may have relation of causal connections among them in upper layer of structure. However, we can recognize the structure of phenomenon by more abstract ideas. It may simplify our understanding of phenomena.

Precondition for application of PCA is normal distribution of data, though this precondition is not rigid. We can apply PCA when the frequency distributions are unimodal.

VI-2-1-2. Analytical operation of PCA

Originally the data are expressed as map of data projected on original variables. Mathematically, PCA is reforming the map of data projected on eigenvectors. It can be simplified as calculation of distance between orthogonal plane to eigenvector on which data is exist and origin.

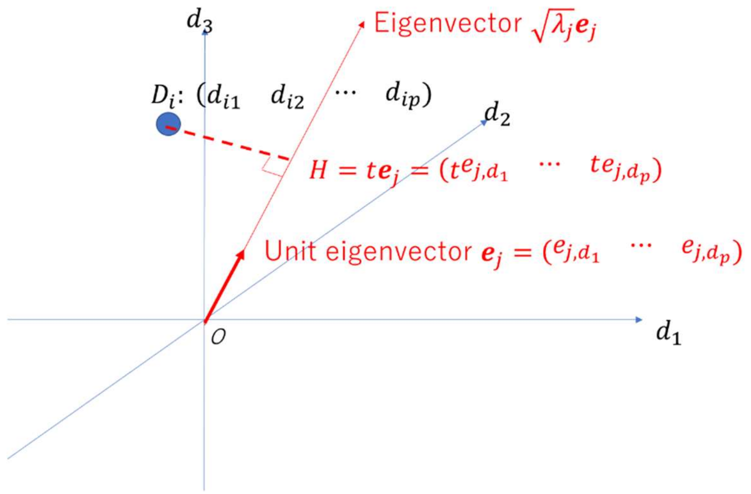

Fig 72. Coordinate transfer from observed variables to eigen vectors.

Figure 72 shows the coordinate transfer. In this case, size of data is n and there are p observed variables. $D_i\colon(d_{i,1} \quad \cdots \quad d_{i,p})$ is a data $((i = 1,\cdots,n))$, $\boldsymbol{e}_j\colon(e_{j,d_1} \quad \cdots \quad e_{j,d_p})$ is unit eigenvector, and H:is foot of perpendicular from origin. $\overrightarrow{\boldsymbol{OH}}$ is eigenvector of which length is t. In this figure, the author shows the eigenvector j (principal component j) of which length is standard deviation as $\sqrt{\lambda_j}\boldsymbol{e}_j$.

$$\text{Vector } \overrightarrow{\boldsymbol{OH}} = t\boldsymbol{e}_j = (te_{j,x_1} \quad \cdots \quad te_{j,x_p})$$

$$\text{Vector } \overrightarrow{\boldsymbol{HD_\iota}} = \left(d_{i,1} - te_{j,x_1} \quad \cdots \quad d_{i,p} - te_{j,x_p}\right)$$

$$\overrightarrow{\boldsymbol{HD_\iota}} \perp \overrightarrow{\boldsymbol{OH}}$$

$$\text{Inner prodduct } \overrightarrow{\boldsymbol{HD_\iota}} \cdot \overrightarrow{\boldsymbol{OH}} = 0$$

$$\left(d_{i,1} - te_{j,x_1} \quad \cdots \quad d_{i,p} - te_{j,x_p}\right)\begin{pmatrix} te_{j,d_1} \\ \vdots \\ te_{j,d_p} \end{pmatrix} = td_{i,1}e_{j,d_1} - t^2 {e_{j,d_1}}^2 + \cdots + td_{i,p}e_{j,d_p} - t^2 {e_{j,d_p}}^2$$

$$= t\left(d_{i,1}e_{j,d_1} - t{e_{j,d_1}}^2 + \cdots + d_{i,p}e_{j,d_p} - t{e_{j,d_p}}^2\right)$$

$$= \ t\left(d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p} - t\left({e_{j,d_1}}^2 + \cdots + {e_{j,d_p}}^2\right)\right)$$

$$= \ t\left(d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p} - t\right)$$

$$\because \ \boldsymbol{e}_j \text{ is unit vector}$$

$${e_{j,d_1}}^2 + \cdots + {e_{j,d_p}}^2 = 1$$

$$t\left(t - d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}\right) = 0$$

$$t = 0 \ or \ t = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

From prior condition

$$t \neq 0$$

So,

$$t = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

This is principle component score of principle component of j of data i.

$$\mathrm{PCS}_{i,j} = t_{ij} = d_{i,1}e_{j,d_1} + \cdots + d_{i,p}e_{j,d_p}$$

From this we can make following table.

Principle component score

	PC1	PC2	⋯,	PCp
Data				
1	$\mathrm{PCS}_{1,1}$	$\mathrm{PCS}_{1,2}$	⋯	$\mathrm{PCS}_{1,p}$
2	$\mathrm{PCS}_{2,1}$	$\mathrm{PCS}_{2,2}$	⋯	$\mathrm{PCS}_{2,p}$
⋮	⋮	⋮	⋱	⋮
n	$\mathrm{PCS}_{n,1}$	$\mathrm{PCS}_{n,2}$	⋯	$\mathrm{PCS}_{n,p}$

We can draw scatter diagram taking 2 or three PC as orthogonal axes. This is PCA.

VI-2-1-3. Diagonalization of variance covariance matrix and PCA.

There two different methods in PCA. Meaning and interpretation of the results are different between two methods. The eigenvectors of variance covariance matrix are used in one method, and the eigenvectors of correlation matrix is used in the other method. Both variance covariance matrix and correlation matrix are symmetric and quadratic form. We use space geometry of quadratic form and nature of symmetric matrix in PCA. Quadratic matrixes are equation of oblique hyperelliptic in multidimensional space when the matrix is positive definite, and the eigenvalues are expressing axes of hyperelliptic. This was explained in V-2-4. Here we use the character of symmetric matrix that we can use transposed matrix ($\boldsymbol{P}^T$) as inverse matrix ($\boldsymbol{P}^{-1}$) of diagonalizing matrix ($\boldsymbol{P}$) in diagonalization of symmetric matrix (V-2-2)

More concretely, we can express this nature as follow using equation

General operation of diagonalization

$$\boldsymbol{P}^{-1}\boldsymbol{V}\boldsymbol{P} = \boldsymbol{\Lambda}$$

In case when $\boldsymbol{V}$ is quadratic form,

$$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$$

we can diagonalize symmetric matrix $\boldsymbol{V}$ as follow

$$\boldsymbol{P}^T\boldsymbol{V}\boldsymbol{P} = \boldsymbol{\Lambda}$$

From definition of variance covariance matrix

$$\boldsymbol{V} = \boldsymbol{D}\boldsymbol{D}^{\boldsymbol{T}}$$

$$\boldsymbol{P}^T\boldsymbol{V}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{D}\boldsymbol{D}^{\boldsymbol{T}}\boldsymbol{P} = (\boldsymbol{P}^T\boldsymbol{D})(\boldsymbol{P}^T\boldsymbol{D})^{\boldsymbol{T}}$$

$$\boldsymbol{P}^T\boldsymbol{D} = \begin{pmatrix} \boldsymbol{e_1}^{\boldsymbol{T}} \\ \boldsymbol{e_2}^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{e_p}^{\boldsymbol{T}} \end{pmatrix} \begin{pmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1p} & d_{2p} & \cdots & d_{np} \end{pmatrix}_{p\times n}$$

$$= \begin{pmatrix} e_{11} & \cdots & e_{j1} & \cdots & e_{p1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{1k} & \cdots & e_{jk} & \cdots & e_{pk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{1p} & \cdots & e_{jp} & \cdots & e_{pp} \end{pmatrix} \begin{pmatrix} d_{11} & d_{21} & \cdots & d_{n1} \\ d_{12} & d_{22} & \cdots & d_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1p} & d_{2p} & \cdots & d_{np} \end{pmatrix}_{p\times n}$$

$$
= \begin{pmatrix} \sum_{j=1}^{p} e_{j1} d_{1j} & \sum_{j=1}^{p} e_{j1} d_{2j} & \cdots & \sum_{j=1}^{p} e_{j1} d_{nj} \\ \sum_{j=1}^{p} e_{j2} d_{1j} & \sum_{j=1}^{p} e_{j2} d_{2j} & \cdots & \sum_{j=1}^{p} e_{j2} d_{nj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{p} e_{jp} d_{1j} & \sum_{j=1}^{p} e_{jp} d_{2j} & \cdots & \sum_{j=1}^{p} e_{jp} d_{nj} \end{pmatrix}_{p\times n}
$$

When we rewrite $e_{jk} d_{ij}$ as $d_{i,j} e_{j,k}$,

$t_{ik} = d_{i,j} e_{k,d_1} + \cdots + d_{i,p} e_{k,d_p} = \sum_{j=1}^{p} e_{jk} d_{ij} = \mathrm{PCS}_{i,k}$

(distance from hyperplane orthogonal to eigenvector including point d.)

$$
\boldsymbol{P}^T\boldsymbol{D} = \begin{pmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ t_{12} & t_{22} & \cdots & t_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1p} & t_{2p} & \cdots & t_{np} \end{pmatrix}_{p\times n} = (\boldsymbol{t_1} \quad \boldsymbol{t_2} \quad \cdots \quad \boldsymbol{t_n})
$$

$$
\boldsymbol{t_i} = \begin{pmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{ip} \end{pmatrix} = \begin{pmatrix} \mathrm{PCS}_{i,1} \\ \mathrm{PCS}_{i,2} \\ \vdots \\ \mathrm{PCS}_{i,p} \end{pmatrix}
$$

$$
\boldsymbol{D}^T\boldsymbol{P} = (\boldsymbol{P}^T\boldsymbol{D})^{\boldsymbol{T}} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{np} \end{pmatrix}_{n\times p} = \begin{pmatrix} \boldsymbol{t_1}^T \\ \boldsymbol{t_2}^T \\ \vdots \\ \boldsymbol{t_2}^T \end{pmatrix}
$$

$$
\boldsymbol{P}^T\boldsymbol{V}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{D}\boldsymbol{D}^T\boldsymbol{P} = \begin{pmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ t_{12} & t_{22} & \cdots & t_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1p} & t_{2p} & \cdots & t_{np} \end{pmatrix}_{p\times n} \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{np} \end{pmatrix}_{n\times p}
$$

$$
= \begin{pmatrix} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i1}{}^2 & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i1} t_{i2} & \cdots & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i1} t_{1p} \\ \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i2} t_{i1} & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i2}{}^2 & \cdots & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i2} t_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{1p} t_{i1} & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{ip} t_{i2} & \cdots & \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{i2}{}^2 \end{pmatrix}_{p\times p}
$$

Eigenvectors are orthogonal each other, and projection of data to eigenvectors are orthogonal each other among eigenvectors.

$$
\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{ij} t_{ik} = \delta_{jk} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} t_{ij} t_{ik}
$$

$$\delta_{jk} = \{ \begin{matrix} 1\ (j = k) \\ 0\ (j \neq k) \end{matrix}$$

δ_{jk} is Kronecker's delta

$$\boldsymbol{P}^T\boldsymbol{V}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{D}\boldsymbol{D}^T\boldsymbol{P} == \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}_{p\times p}$$

Therefor λ_i is sum of square of PC_i (SS of eigenvector i).

VI-2-1-4. Notation of results of PCA.

One of the meaning of PCA is aggregation of data. PCA gives same number of PC as number of variables in original data. Several PCs have larger variation than original variances data and several PCs have smaller variances than original variances. We do not need to consider the impact of PCs which have weak influence to the phenomenon. For this reason, we have to select important PCs at first. In of the criteria of the selection is contribution ratio of principle component to total valiance. Variance of PC is λ, and total variance is trace of diagonalized matrix: sum of diagonal elements of diagonalized matrix.

$$V_{total} = \sum_{j=1}^{p} \lambda_j$$

	Contribution ratio	Cumulative contribution ratio
PC1	$\frac{\lambda_1}{V_{total}}$	$\frac{\lambda_1}{V_{total}}$
PC2	$\frac{\lambda_2}{V_{total}}$	$\frac{\lambda_1+\lambda_2}{V_{total}}$
⋮	⋮	⋮
PCp	$\frac{\lambda_p}{V_{total}}$	$\frac{\lambda_1+\lambda_2+\cdots+\lambda_p}{V_{total}}$

When we want to explain the 70% of mechanism of phenomena, we pick up principle components included in 0.7 cumulative contribution ratio, and we consider the other principle component as meaningless fluctuation. This judgement is simple, though it is too mechanical. When several similar contributions exist around 0.7 cumulative contribution ratio, we debate which principle component should be cut off. Another method for selection of principle component is scree plot. Scree plot is a method to draw line graph of eigenvalue in descending order as shown in figure 73.

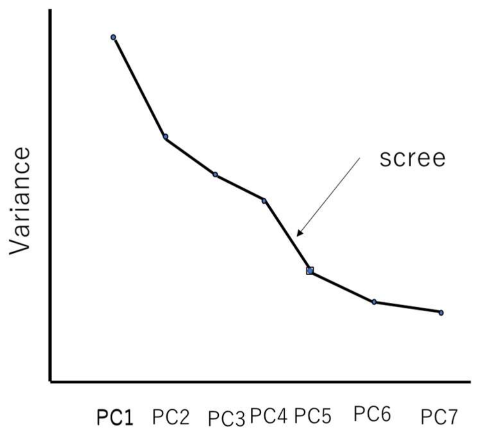

Fig. 73 Scree plot of variances.

We call this figure as scree plot. We can observe large difference between PC4 and PC5. This point is scree and we cut of PC after PC5.

This selection method includes ambiguities. When we started PCA from correlation matrix, we can make more mathematical discussion.

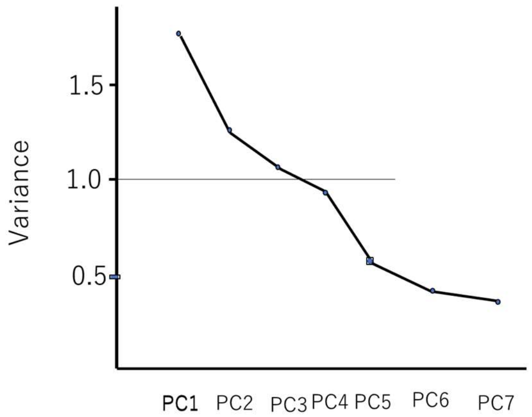

Fig. 74. Scree plot of variances in PCA started from correlation matrix.

When we started PCA from correlation matrix, variances of all variables are 1. PCA reallocates variances of original variables to PCs, though the average is still 1. This means that PCs which have variance higher than 1 have strong relation to absorb variances, and PCs which have variance lower than 1 have week relation to the phenomena. We select PCs which have variance larger than 1 (See figure 74).

Actually, opened soft wares for public use have additional functions for support of interpretation of the result. Most commonly used indicator for interpretation of meaning of PSs is principle component loading. This indicator means strength of interaction between principle component and observed variable.

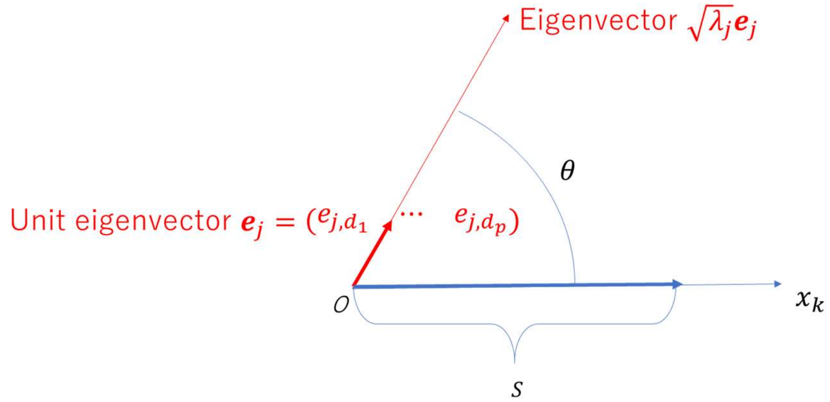

Fig75. Inner product between eigenvector and observed variable and correlation coefficient. $(k = 1, \cdots p)$

We take an eigenvector and vector of a variable as shown in figure 75. We fix the length of the vector as root of eigenvalue and S. Actually, we do not need to fix the length of eigenvector to be $\sqrt{\lambda_j}$. It can be any arbitrary real number, because the relation between vector correlation coefficient and correlation efficient is cosine of angle of two vectors. Here, the author selects $\sqrt{\lambda_j}$ for the connection to principal component loading.

There two definition is inner product of vector.

$$\boldsymbol{V_1} \cdot \boldsymbol{V}_2 = [\boldsymbol{V}_1][\boldsymbol{V_2}] \cos\theta = (v_{1,1} \quad \cdots \quad v_{1,n})(v_{2,1} \quad \cdots \quad v_{2,n})^T$$

$$\text{Here, } \boldsymbol{V_1} = (v_{1,1} \quad \cdots \quad v_{1,n}),\ \boldsymbol{V}_2 = (v_{2,1} \quad \cdots \quad v_{2,n})$$

And, correlation coefficient of $\boldsymbol{V_1}$ and $\boldsymbol{V}_2$ is $\cos\theta$

In the case of eigenvector and vector of a variance shown in figure 75,

$$\sqrt{\lambda_j} S \cos\theta_{j,k} = \left(\sqrt{\lambda_j} e_{j,d_1} \quad \cdots \quad \sqrt{\lambda_j} e_{j,d_p}\right)\begin{pmatrix} 0 \\ \vdots \\ S \\ \vdots \\ 0 \end{pmatrix} = \mathrm{S}\sqrt{\lambda_j} e_{j,d_k}$$

$$\cos\theta = e_{j,d_k}$$

Correlation coefficient is e_{j,d_j},

Definition of principal component loading (PCL) is magnitude of correlation of an observed variable to eigenvector. When we express the magnitude as portion of correlated area in the standard length of principal component.

$$\mathrm{PCL}_{j,k} = \sqrt{\lambda_j} r_{j,k} = \sqrt{\lambda_j} e_{j,d_k}$$

We can make following table

<u>Principle component loading</u>

PC1 PC2 $\cdots$, PCp

	S.D	$\sqrt{\lambda_1}$	$\sqrt{\lambda_2}$	$\cdots$	$\sqrt{\lambda_p}$
Variable					
Variable 1		$\sqrt{\lambda_1}e_{1,d_1}$	$\sqrt{\lambda_2}e_{2,d_1}$	$\cdots$	$\sqrt{\lambda_p}e_{p,d_1}$
Variable 2		$\sqrt{\lambda_1}e_{1,d_2}$	$\sqrt{\lambda_2}e_{2,d_2}$	$\cdots$	$\sqrt{\lambda_p}e_{p,d_2}$
$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$
Variable p		$\sqrt{\lambda_1}e_{1,d_p}$	$\sqrt{\lambda_2}e_{2,d_p}$	$\cdots$	$\sqrt{\lambda_p}e_{p,d_p}$

We can understand relation between principle component and variables from this table, though we have to consider meaning of comparison among $\sqrt{\lambda_i}$ among principle components. If our data include data measured by different minimum units, $\sqrt{\lambda_i}$ is larger in the data measured by smaller minimum unit. More theoretically, variances in a variable are different among variable. It may be no meaning to compare the importance of principle component among principle components which have different variance. There two methods in PCA. One is method to use eigenvalue of variance and covariance matrix and the other is method to use eigenvalue of correlation matrix. Two methods provide different result, because former method uses unstandardized data, and second method uses standardized data.

The purpose of analysis is different between the two methods. If we cannot neglect difference of variance among principle components. It is better to make following table.

<u>Correlation matrix of PCs and variables</u>

	PC1	PC2	$\cdots$,	PCp
Variable				
Variable 1	e_{1,d_1}	e_{2,d_1}	$\cdots$	e_{p,d_1}
Variable 2	e_{1,d_2}	e_{2,d_2}	$\cdots$	e_{p,d_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Variable p	e_{1,d_p}	e_{2,d_p}	$\cdots$	e_{p,d_p}

As in the explanation in calculation procedure, e_{j,x_k} is correlation coefficient.

$$e_{j,d_k} = r_{jk}$$

One of the visualization methods of the table is to make following figure.

The circle is a section of hypersphere at $\mathrm{PC}a - \mathrm{PC}b$ plane. The radius of the sphere is 1. $\mathrm{V}c$ is projection of vector of variable c of which length is standard deviation. The coordinate of $\mathrm{V}c$ is correlations with $\mathrm{PC}a$ and $\mathrm{PC}b$

$$\mathrm{V}c = (r_{ac} \quad r_{bc})$$

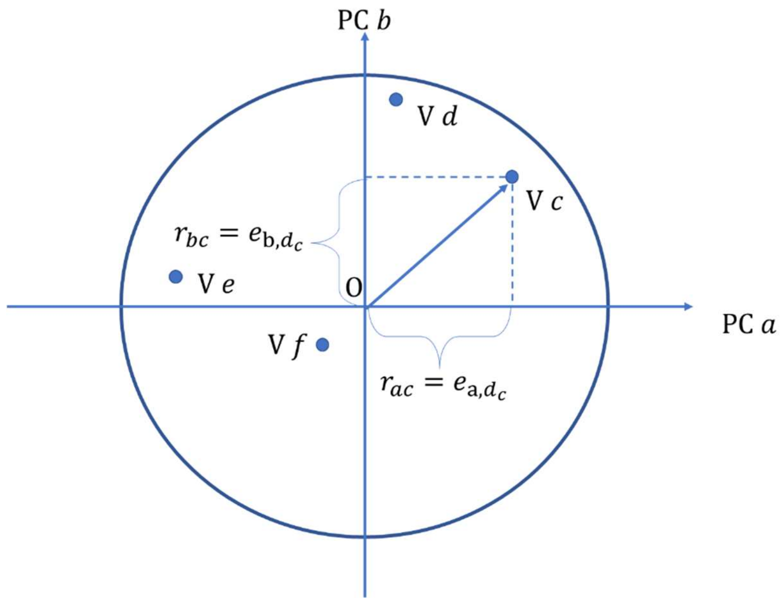

Fig. 76. Visualization of contribution ratio of principle component to variables

$$\left|\overrightarrow{\mathrm{OV}c}\right|^2 = {r_{ac}}^2 + {r_{bc}}^2$$

$$\left|\overrightarrow{\mathrm{OV}c}\right| = \sqrt{{r_{ac}}^2 + {r_{bc}}^2}$$

$\left|\overrightarrow{\mathrm{OV}c}\right|^2$ is contribution ratio of $\mathrm{PC}a$ and $\mathrm{PC}b$. When $\mathrm{V}c$ is completely explained only by the relation with $\mathrm{PC}a$ and $\mathrm{PC}b$,

$$\left|\overrightarrow{\mathrm{OV}c}\right|^2 = {r_{ac}}^2 + {r_{bc}}^2 = 1$$

and

$$\left|\overrightarrow{\mathrm{OV}c}\right| = \sqrt{{r_{ac}}^2 + {r_{bc}}^2} = 1$$

The length of $\left|\overrightarrow{\mathrm{OV}c}\right|$ in the figure is around 0.8. The contribution ratio of $\mathrm{PC}a$ and $\mathrm{PC}b$ is $0.8^2 = 0.64$. this means that more than half of variance of $\mathrm{V}c$ can be explained in the relation with $\mathrm{PC}a$ and $\mathrm{PC}b$. We can conclude that $\mathrm{PC}a$ and $\mathrm{PC}b$ are major component of $\mathrm{V}c$, and the contributions of $\mathrm{PC}a$ and $\mathrm{PC}b$ are nearly the same. $\mathrm{V}d$ is closer to the circle, though the value of r_{ad} is small. From this we can estimate that $\mathrm{V}d$ is mainly related with $\mathrm{PC}b$. $\mathrm{V}e$ is also exist near the circle and the value of r_{ae} is negative and r_{be} is small. From this, we can estimate that $\mathrm{V}e$ is inversely related with $\mathrm{PC}a$. The distance from the origin to $\mathrm{V}f$ is short. That means $\mathrm{V}f$ has little relation with $\mathrm{PC}a$ and $\mathrm{PC}b$. Possibly, vector $\mathrm{V}f$ has slope in the orthogonal direction to the $\mathrm{PC}a - \mathrm{PC}b$ plane. We have to make another section of hypersphere by selecting other combination of PCs.

Simplest method of interpretation of result is to draw projection of vector of original variables in the scatter diagram on the plane of two PCs. For comparison we fix the length of vector of original variables as standard deviation.

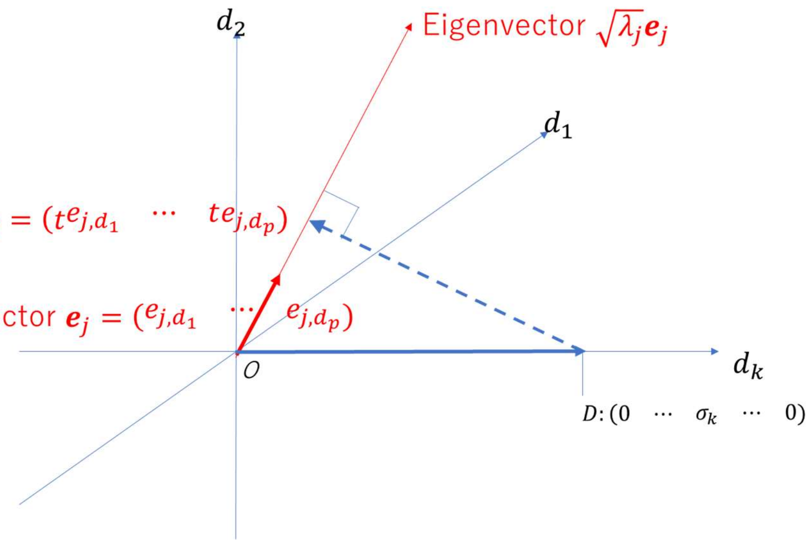

Fig 77. Projection of vector on the axis of x_k to vector of PCs.

Calculation procedure of the projection (Figure 77) is the same as coordinate transfer from observed variables to eigenvectors. The length of the vector is σ_k. This means that the coordinate of D is $(0 \quad \cdots \quad \sigma_k \quad \cdots \quad 0)$.

$$\overrightarrow{\mathrm{OH}} \perp \overrightarrow{\mathrm{DH}}$$

$$\overrightarrow{\mathrm{DH}}=(te_{j,d_1} \quad \cdots \quad te_{j,d_p}) - (0 \quad \cdots \quad \sigma_k \quad \cdots \quad 0) = \left(te_{j,d_1} \quad \cdots \quad te_{j,d_k} - \sigma_k \quad \cdots \quad te_{j,d_p}\right)$$

$$\overrightarrow{\mathrm{OH}} \cdot \overrightarrow{\mathrm{DH}}=0$$

$$\boldsymbol{e}_j \cdot \overrightarrow{\mathrm{DH}}=0$$

$$(e_{j,d_1} \quad \cdots \quad e_{j,d_p}) \begin{pmatrix} te_{j,d_1} \\ \vdots \\ te_{j,d_k} - \sigma_k \\ \vdots \\ te_{j,d_p} \end{pmatrix} = 0$$

$$t\left(e_{j,d_1}{}^2 + \cdots + e_{j,d_p}{}^2\right) - \sigma_k e_{j,d_k} = 0$$

$$t = \sigma_k e_{j,d_k}$$

$$\because e_{j,x_1}{}^2 + \cdots + e_{j,x_p}{}^2 = 1$$

Consequently, we can make following table.

	PC					
Variable	PC1	PC2	$\cdots$	PCk	$\cdots$	PCp
Variable 1	$\sigma_1 e_{1,d_1}$	$\sigma_1 e_{2,d_1}$	$\cdots$	$\sigma_1 e_{k,d_1}$	$\cdots$	$\sigma_1 e_{p,d_1}$
Variable 2	$\sigma_2 e_{1,d_2}$	$\sigma_2 e_{2,d_2}$	$\cdots$	$\sigma_2 e_{k,d_2}$	$\cdots$	$\sigma_2 e_{p,d_2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$

Variable k	$\sigma_k e_{1,d_k}$	$\sigma_k e_{2,d_k}$	$\cdots$	$\sigma_k e_{k,d_k}$	$\cdots$	$\sigma_k e_{p,d_k}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Variable p	$\sigma_p e_{1,d_p}$	$\sigma_p e_{2,d_p}$	$\cdots$	$\sigma_p e_{k,d_p}$	$\cdots$	$\sigma_p e_{p,d_p}$

We can project vectors of the variables of which length is standard deviation in the scatter diagram on the plane of selected two PCs. When we select PC1-PC2 plane for the projection of data, the coordinate of arrowhead of the vectors are as follow.

Variable 1: $(\sigma_1 e_{1,d_1} \quad \sigma_1 e_{2,d_1})$

Variable 2: $(\sigma_2 e_{1,d_2} \quad \sigma_2 e_{2,d_2})$

$\vdots \qquad \vdots$

Variable k: $(\sigma_k e_{1,d_k} \quad \sigma_k e_{2,d_k})$

$\vdots \qquad \vdots$

Variable p: $(\sigma_p e_{1,d_p} \quad \sigma_p e_{2,d_p})$

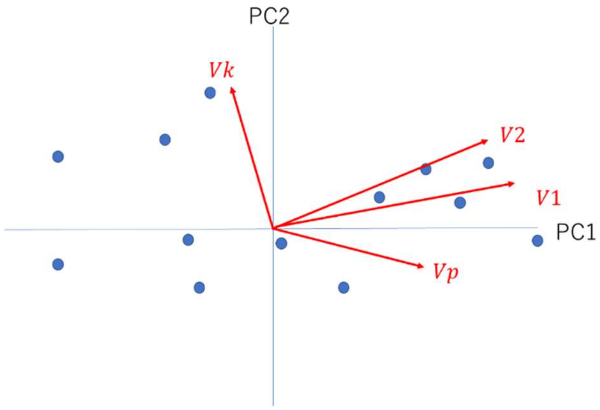

Fig 78. Example of scatter diagram with projection of vector of original variables.

Figure 78 shows an example of scatter diagram with projection of vector of original variables on coordinate of PC1-PC2. This procedure is similar as correspondence analysis. Putting this aside, direction of arrow of variable 1 and variable 2 are approximately same with direction of PC1. And the lengths are long. This means that the variables have high correlation with PC1. It is not always, though PC1 generally expresses something relating to magnitude of phenomenon or event, and qualitative character such as gap, ratio, instability, ambiguity and so on comes after PC2. When PC1 is expressing quantitative characters, variable 1 and variable 2 are indicating similar quantitative characters. These similarities are named internal consistency. Internal consistency is not so important concept in PCA. However, it is important concept in factor analysis (FA) particularly in questionnaire survey in psychological research. When answers to plural number of questions have high correlation, the questions have higher internal consistency. This means that the questions have asked same content in different expression. This is important in psychology. However, internal consistency is negatively evaluated in other research fields, because replication of similar questions decreases the value of information. Internal consistency has been evaluated by Cronbach's α. When we obtain high Cronbach's α, we can consider to reduce of the number of variables by selection of a typical variable in next questionnaire survey for downsizing of questionnaire or inversely, making new synthetic variable by combining the variables to make more sensitive indicator using internal consistency.

VI-2-1-5. Interpretation of results of PCA.

The readers can easily understand mathematical meaning of PCA. However, interpretation of results of PCA is sometimes difficult, and there are various explanations of what is PCA. It depends on purpose of analysis. Most general explanation of PCA is a method for aggregation of components composing dataset to several main components. However, generally, we can select representative components from all components using information and knowledges in our own discipline without PCA. Method for seeking latent factors exist in phenomena is another explanation of PCA. We sometimes can find mechanisms existing in the background of phenomena by PCA. However, such happy case is rare. Extracted components by PCA are generally too abstract to explain by verbalized causal laws. Result interpretation is easier in factor analysis (FA), particularly when we use oblique rotation. However, in early days of factor analysis, PCA was mainly used for extraction of factors, and PCA is used as default for extraction of factors for FA in SPSS. The reason why PCA was used for extraction of factor in early days is generally considered that extraction of factors by PCA is easier than other methods of factor extraction (calculation process is far shorter than other method) and PCA robustly gives result. Other methods sometimes give improper solution or do not converge. The author does not confirm his guess, though he is thinking that there is historic background. When we learn linear algebra, we can naturally notice that we can use diagonalization and spectral decomposition to explain the phenomena by rectangular component (independent factor). PCA gives its result by axes of components. The axes are rectangular and independent. However, meanings of the axes are often not understandable by terminology in our daily life. Sometimes it means something likes general mass or gaps among elements and so on. It quite natural to consider rotation of axes to fit more understandable vectors in our perception of each scientific discipline or experiences in our daily life neglecting independency of components. Essence of FA is separation of sharing parts among variances from observed variances. When it is possible to extract sharing parts by any method, we do not need to keep diagonality among components. Machine power of computer enables successive approaches to reach solutions of separation and we do not need to keep diagonality in extraction of factors. Now a days, there are no direct relation between PCA and FA. However, PCA is still a portion of FA as a method of factor extraction. Moreover, PCA still has important roll as a method to understand structure of phenomena. The author is thinking that PCA is a method to relocate data in a multidimensional space to other space removing correlation among variables. This is spectral decomposition.

Distribution of organisms are determined by environment, and environmental physical factors are originally independent such as relation between temperature and salinity. However, when we measure salinity and temperature in coastal area, there is correlation in reality. When we measure the salinity and temperature from river mouth to offshore in summer season, temperature decrease with distance from the river mouth and salinity increase with the distance, as a result, temperature and salinity have clear correlation. Similarly, chemical environment such as phosphate and nitrate concentration have clear relation with salinity. Living organisms are adopted such combination of environmental factors. We cannot understand distribution of living organisms only by philosophy of simple reductionism. Element reduction philosophy may effectively work for understanding of physiological reaction of living organisms, though we cannot understand complex system such as behavior and evolution of living organisms by simple factor. In such case, we can use PCA for summary and understanding of phenomena. A common strategy for understanding of complex system is exhaustive accumulation of data including variables which probably has no relation to the phenomenon. As an example, when we want know the mechanism of distribution of fish species in a coastal area without prior information, we catch the fish in various place in the coastal area and record the biological data such as species name, number of individuals in the species, body weights, body length as well as chemical and physical environment data, water depth, salinity, water temperature, transparency, bottom condition, time of the sampling in the day and so on. Some variables have correlation and some have inverse correlation and the others have no relation. We cannot say this research method is well designed, though we need to start such visionless approach, when we lack prior information. This is not hypothesis verification. Some may say science should be hypothesis verification. This is misunderstanding. We cannot verify hypothesis without hypothesis. Science should make hypothesis from randomly piled up dataset at first. For this purpose, analyzers make histograms of variables to confirm distribution of data, and some analyzers make X-Y plots for confirmation of relation among observed variables, and some analyzers implement correlation analysis. Most commonly, we make variance and covariance matrix or correlation matrix to see the entire picture of latent relations. In the days before popularization of computer, the author analyzed major factors of production of algae in coastal area of Japan. The author collected data of production of unit area and price of the algae for more than 50 years from all regions of coastal area of Japan and made round robin correlation matrix of production and price among regions. He needed more than 2 months to make the matrix. Finally, he color-coded correlating region in the map of Japan, and he found an important factor relating with the production of the algae. He found interlocking fluctuation among areas remoted each other and consider the background mechanism. This approach is, by now, unsophisticated and childish, though he could reach conclusion. He did not know PCA at that time. He could not implement PCA, even if he knew PCA, because he did not have computer. The author is thinking that he may recommend application of PCA of all data including environment data to himself in young age. Most important function of PCA is visualization of structure of relation including background latent mechanism. The author in his young age had vague expectation for finding of background mechanism of phenomena by structuring whole relationship by round robin correlation analysis of variables. His success was only by blind luck. Because correlation does not always express causal connection. For instance, production of algae among distant areas cannot be causes and results. We can structuralize phenomena by finding latent factors which make the fluctuation in each distant area. This means that we need information in each discipline and we cannot reach any conclusion only by PCA. In the case of experience of the author in his young age, he had knowledges of costal environment of Japan and physiology of algae. However, we can say that his expectation is a motive of PCA. PCA is very easy when we can use computer. We can draw out overall picture of phenomena by PCA, though we do not always understand the meaning of the picture. The author recommend to implement PCA at first.