Statistical methods aimed to explain variation in correlated data by few latent variables

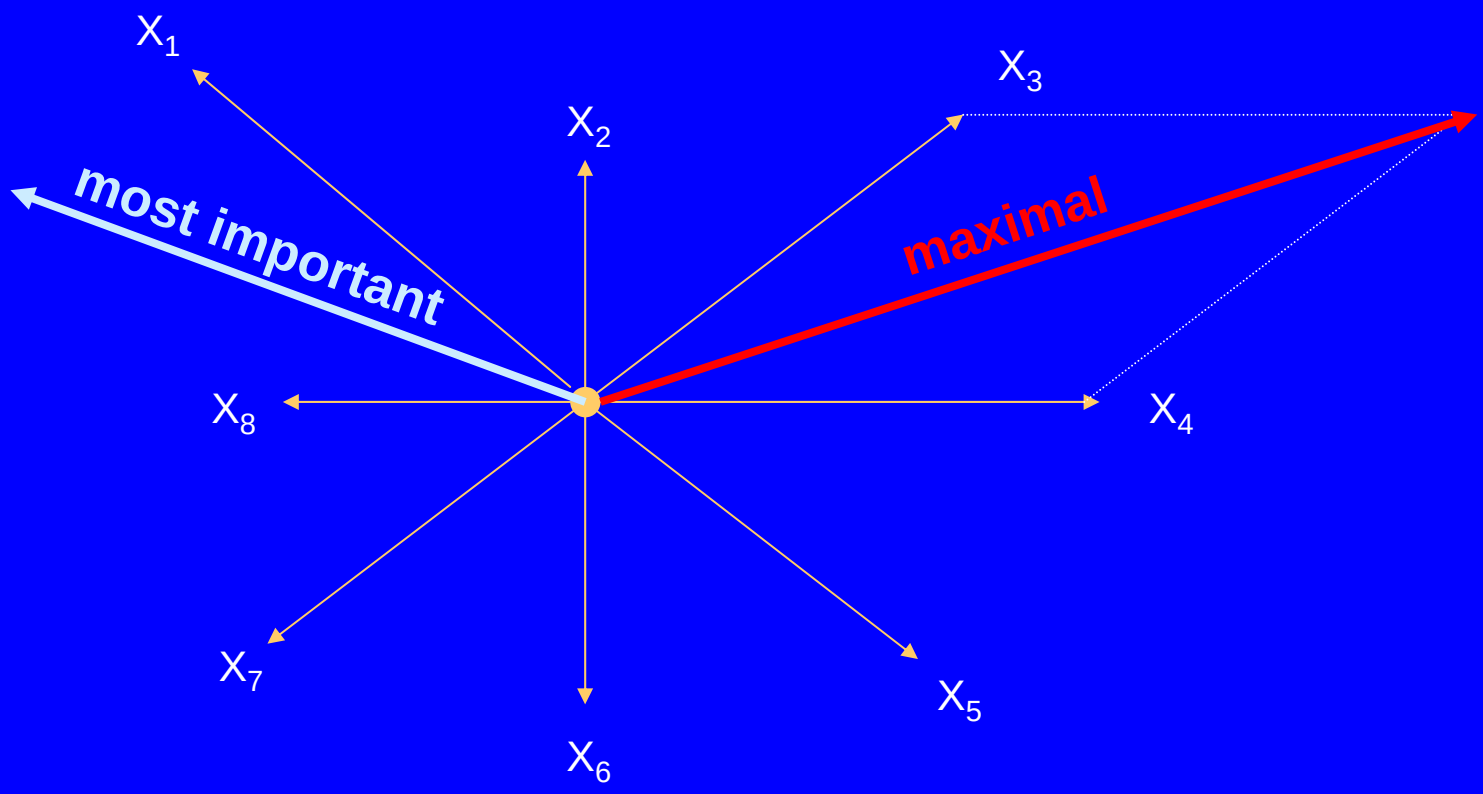
Kurt Hoffmann and Heiner Boeing
Department of Epidemiology
German Institute of Human Nutrition
Potsdam-Rehbrücke

Objective

To evaluate and compare different statistical methods that aim to explain maximal variation in selected correlated variables.

The comparison refers to theoretical assumptions, methodological aspects and applications to real data.

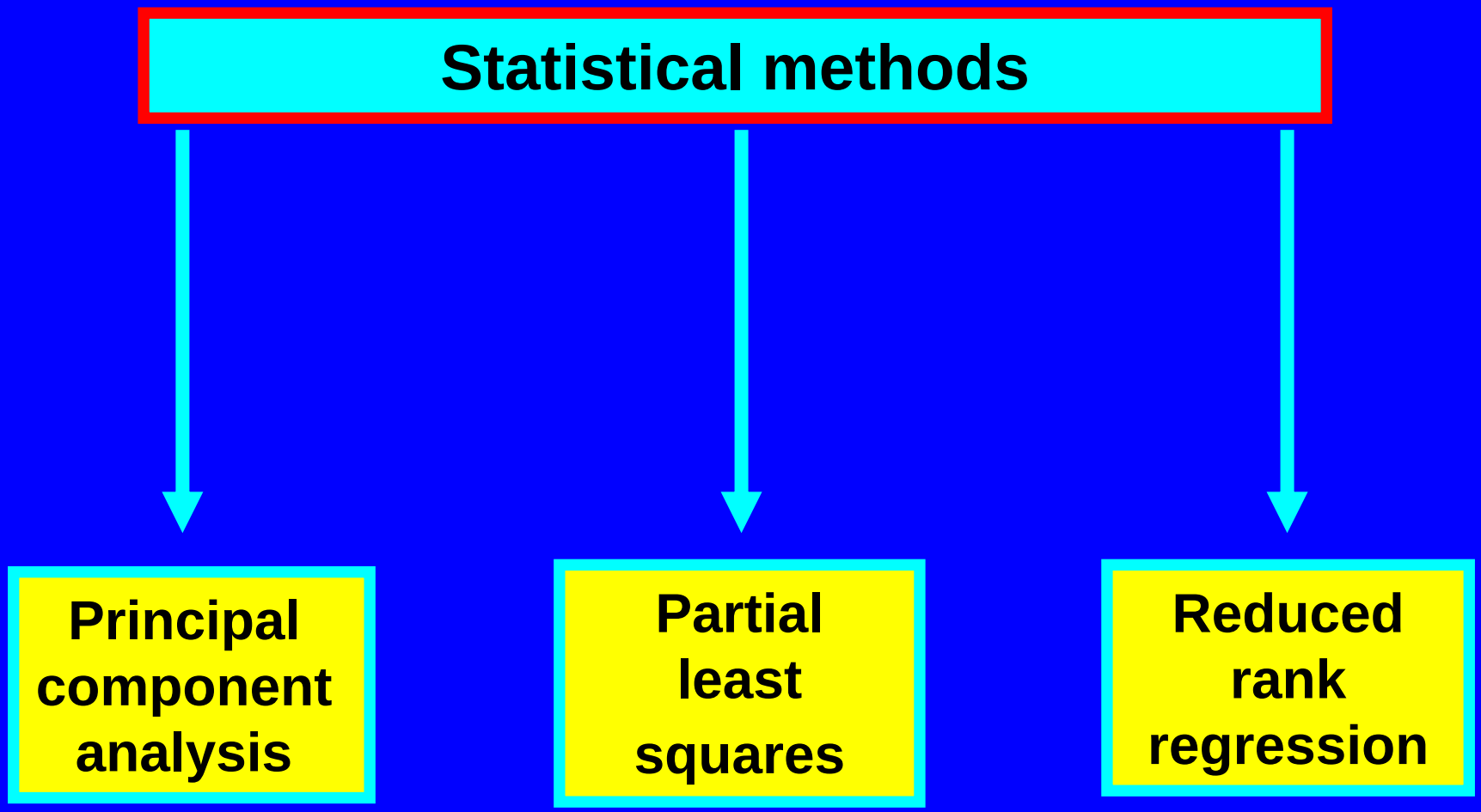
Directions of variation



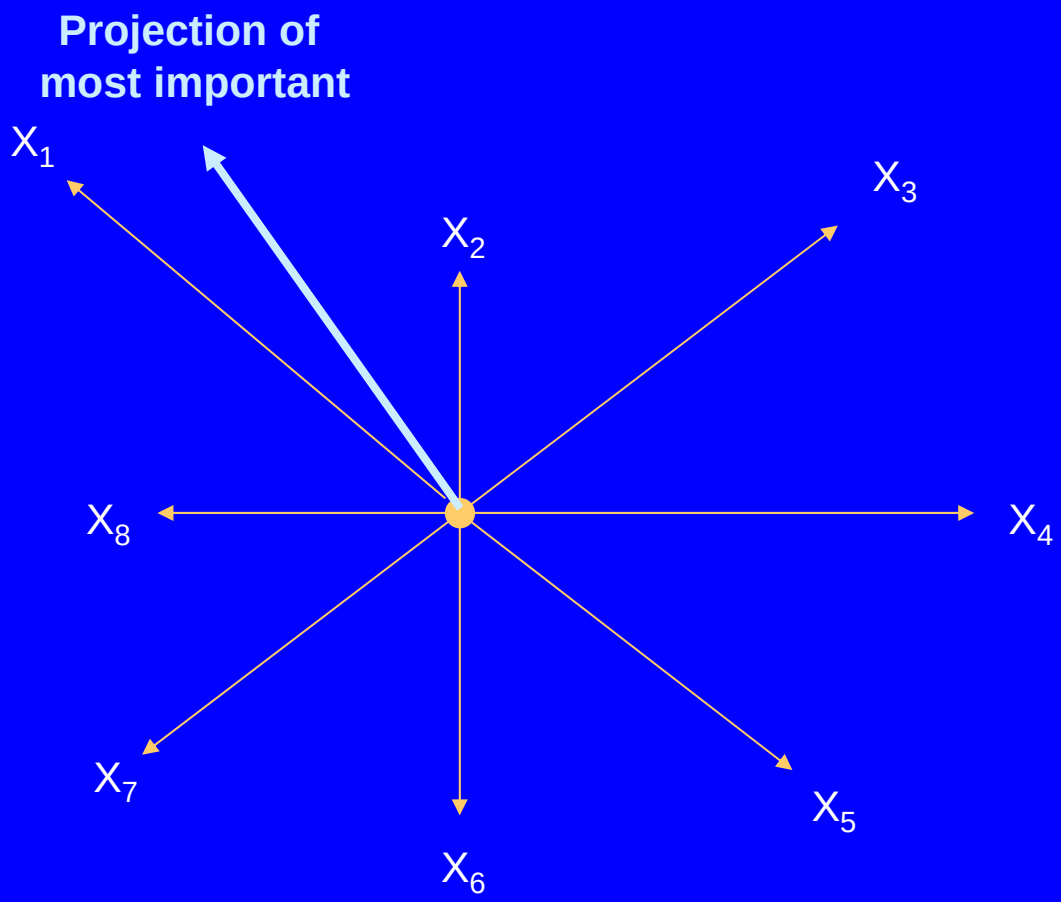
What is the direction of maximal variation ?

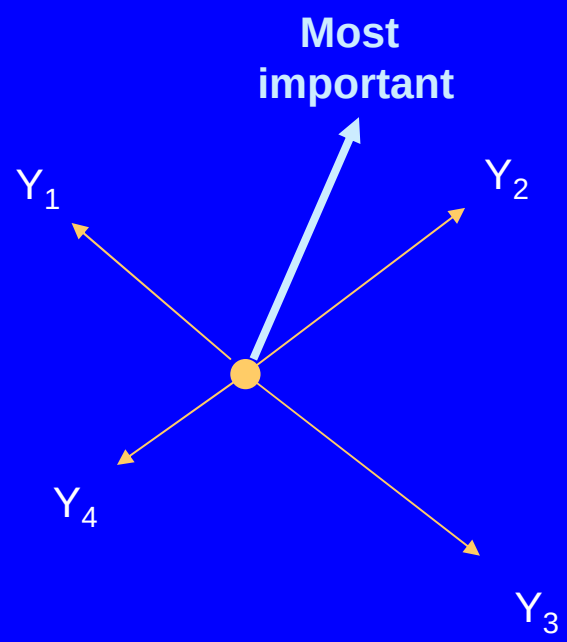
What is the most important direction of variation ?

Overview


Statistical methods
Principal component analysis
Partial least squares
Reduced rank regression

Variation in two sets of variables





Predictors (original variables)

Responses (ancillary variables)

Comparison of objectives

Method		Objective
Principal component analysis	*PCA*	Explaining as much predictor variation as possible
Reduced Rank regression	*RRR*	Explaining as much response variation as possible
Partial least squares	*PLS*	Explaining much predictor and response variation

Method Description

Principal component analysis

It is a dimension-reduction technique.
Starting point are the eigenvalues
of the covariance matrix of predictors.

$$\sum_X = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Var(X_n) \end{pmatrix}$$

Principal component analysis

$\lambda_1, \lambda_2, \ \ldots \ , \lambda_n$

Eigenvalues of Σ_X (decreasing)

$e_1, e_2, \ \ldots \ , e_n$

Corresponding eigenvectors

$X = (X_1, X_2, \ \ldots \ , X_n)$

Vector of predictors

$$F_1 = e_1^T X,$$
$$F_2 = e_2^T X,$$
$$\cdot \quad \cdot \quad \cdot$$
$$\cdot \quad \cdot \quad \cdot$$
$$F_n = e_n^T X$$

Principal components (factors)

Principal component analysis

The first factor is the linear function of predictors that maximises the explained variation of predictors.

The k^{th} factor is the linear function of predictors that maximises the explained variation of predictors within the class of linear functions that are orthogonal to the first k-1 factors.

There are so many eigenvalues as predictors.

An eigenvalue describes the fraction of predictor variation explained by the corresponding factor.

The factors are uncorrelated.

Method Description

Reduced rank regression

It is a dimension-reduction technique.
Starting point are the eigenvalues
of the covariance matrix of responses.

$$\sum_Y = \begin{pmatrix} Var(Y_1) & Cov(Y_1,Y_2) & . & . & . & . & Cov(Y_1,Y_m) \\ Cov(Y_2,Y_1) & Var(Y_2) & . & . & . & . & Cov(Y_2,Y_m) \\ . & . & . & . & . & & \\ . & . & . & . & . & & \\ . & . & . & . & . & & \\ Cov(Y_m,Y_1) & Cov(Y_m,Y_2) & . & . & . & . & Var(Y_m) \end{pmatrix}$$

Reduced rank regression

$\lambda_1, \lambda_2, \quad \ldots \quad, \lambda_m$

Eigenvalues of Σ_Y (decreasing)

$e_1, e_2, \quad \ldots \quad, e_m$

Corresponding eigenvectors

$Y = (Y_1, Y_2, \quad \ldots \quad, Y_m)$

Vector of responses

$F_1 = P_X(e_1^T Y), F_2 = P_X(e_2^T Y), \ldots, F_n = P_X(e_n^T Y),$

RRR factors

Projection onto the space of predictors

Reduced rank regression

The first factor is the linear function of predictors that maximises the explained variation of responses.

The k^{th} factor is the linear function of predictors that maximises the explained variation of responses under a certain orthogonality restraint of dimension k-1.

There are so many eigenvalues as the minimum of the number of responses and the number of predictors.

An eigenvalue describes the fraction of response variation explained by the corresponding factor.

The factors are nearly uncorrelated.

Method Description

Partial least squares

It is a dimension-reduction technique. Starting point are the eigenvalues of the matrix of covariance between predictors and responses.

$$\sum_{X,Y} = \begin{pmatrix} Cov(X_1,Y_1) & Cov(X_1,Y_2) & \ldots & Cov(X_1,Y_m) \\ Cov(X_2,Y_1) & Cov(X_2,Y_2) & \ldots & Cov(X_2,Y_m) \\ \ldots & & & \\ \ldots & & & \\ \ldots & & & \\ Cov(X_n,Y_1) & Cov(X_n,Y_2) & \ldots & Cov(X_n,Y_m) \end{pmatrix}$$

Partial least squares

$\lambda_1, \lambda_2, \quad \ldots \quad , \lambda_m$

Eigenvalues of Σ_{XY} (decreasing)

$e_1, e_2, \quad \ldots \quad , e_m$

Corresponding eigenvectors

The eigenvectors will be projected onto the space of predictors and onto the space of responses resulting in a factor score and a response score.

There are so many eigenvalues as the minimum of the number of responses and the number of predictors.

The response and factor scores possess no optimality property. The factors are nearly uncorrelated.

THE SAS PROCEDURE FOR *PCA, PLS* AND *RRR*

```
proc pls data=..... method=...;
   model y1 .....ym = x1 .....xn ;
run;
```

$y_1 \ldots\ldots y_m$ = response variables

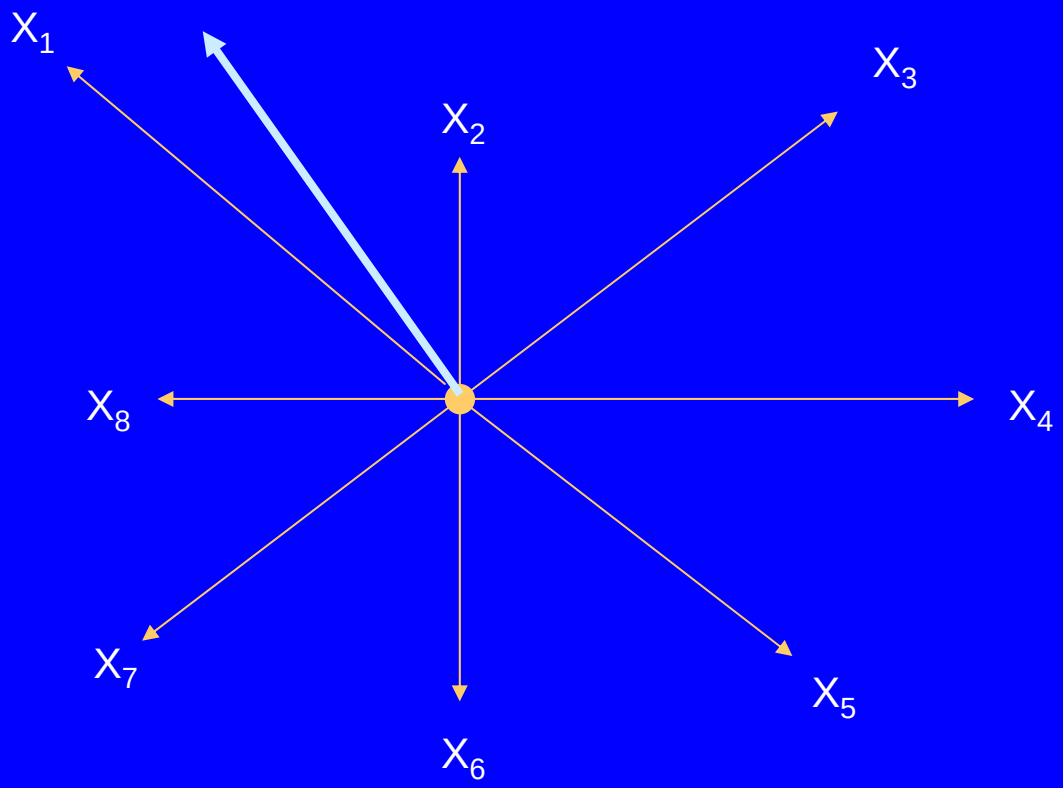
$x_1 \ldots\ldots x_n$ = predictor variables

method = *PCR*, *PLS* or *RRR*

APPLICATIONS TO NUTRITIONAL EPIDEMIOLOGY

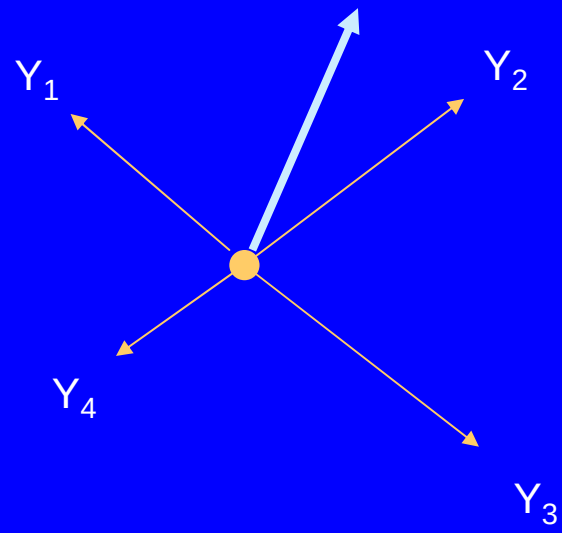
Typical situation

Observed variation



food group intake

Variation of interest



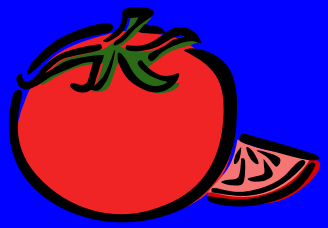
nutrient intake

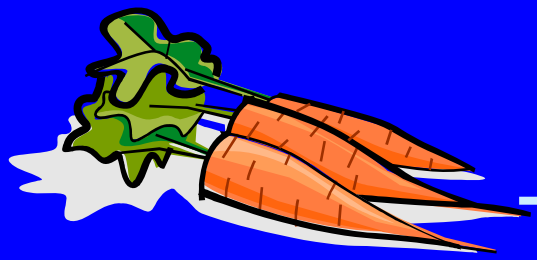
Data basis

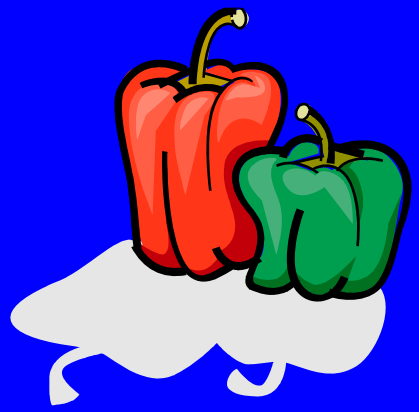
The EPIC-Potsdam study

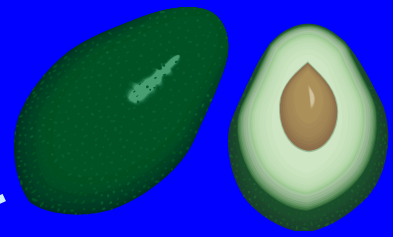
Data assessment in:	**1994-98**
Number of participants:	**27 548**
Women/ Men:	**16 644 / 10 904**
Mean follow-up time:	**7 years**
Items in food frequency questionnaire:	**148**
Number of food groups:	**39**
Nutrients of interest (e.g.):	**vitamins**

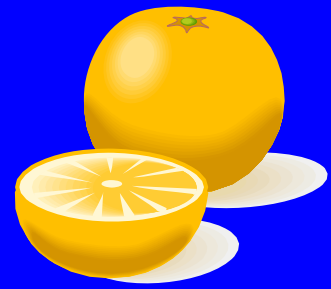
Vitamins of interest

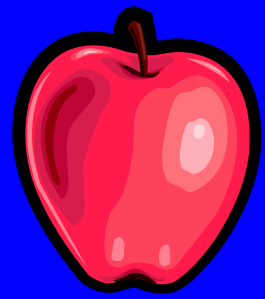












- Vitamin A
- Vitamin B1
- Vitamin B2
- Vitamin B6
- Vitamin B9
- Vitamin C

Pretreatment of responses

Logarithmic transformation

High correlation of logarithmically transformed nutrient intakes reflect proportionality of concentrations in foods

Energy adjustment

Regression on (logarithmically transformed) energy intake and using the residuals removes the quantitative component of intake

Correlation matrix

	A	B1	B2	B6	B9	C	
$\sum_Y =$	1	0.21	0.34	0.24	0.25	0.12	A
		1	0.31	0.67	0.18	0.25	B1
			1	0.41	0.60	0.14	B2
				1	0.54	0.37	B6
					1	0.44	B9
						1	C

Correlation of energy adjusted logarithmically transformed vitamin intakes

Explained variation of food groups

	PCA	*PLS*	*RRR*
1. Factor	9.8	7.0	4.3
2. Factor	7.2	6.2	3.3
3. Factor	5.8	5.5	3.5
4. Factor	4.1	4.2	2.5
5. Factor	4.0	4.3	2.8
6. Factor	3.5	3.2	3.3
Total	34.4	30.4	19.7

Explained variation of vitamins

	PCA	*PLS*	*RRR*
1. Factor	2.2	17.6	23.5
2. Factor	5.8	7.4	9.3
3. Factor	6.0	6.3	9.0
4. Factor	1.1	6.1	3.4
5. Factor	1.5	2.9	2.3
6. Factor	1.5	2.4	1.4
Total	**18.1**	**42.7**	**48.9**

Variation of single vitamins explained by the first three RRR factors

	1. Factor	2. Factor	3. Factor
Vitamin A	5.2	0.0	0.4
Vitamin B1	26.0	2.6	22.4
Vitamin B2	26.2	14.5	12.5
Vitamin B6	34.9	0.3	6.9
Vitamin B9	28.6	0.1	12.4
Vitamin B9	20.0	39.2	0.7

Response score of the first RRR factor

$$
\begin{aligned}
&0.19 \times f\,(\mathrm{A}) \\
+\ &0.43 \times f\,(\mathrm{B1}) \\
+\ &0.43 \times f\,(\mathrm{B2}) \\
+\ &0.50 \times f\,(\mathrm{B6}) \\
+\ &0.45 \times f\,(\mathrm{B9}) \\
+\ &0.38 \times f\,(\mathrm{C})
\end{aligned}
$$

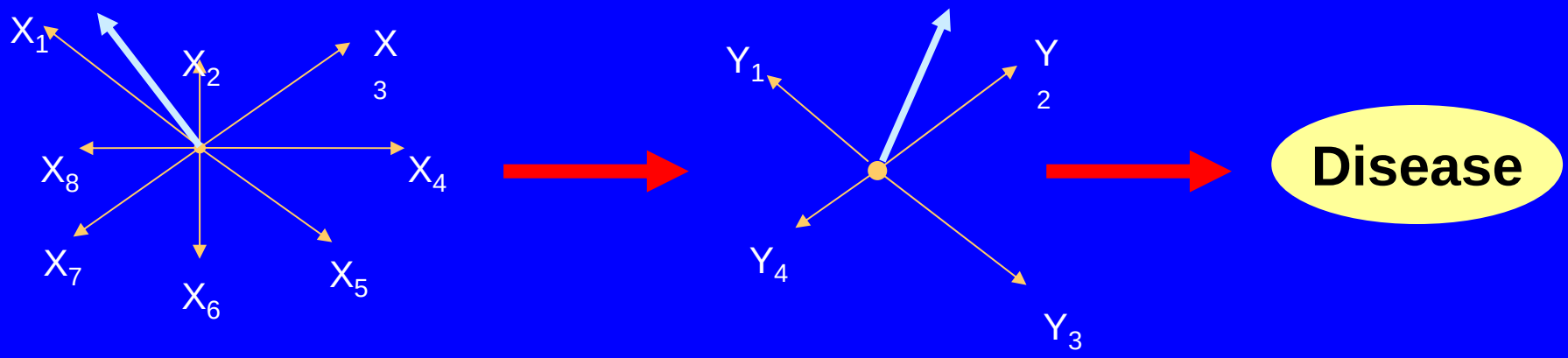
f **= standardized variable of the logarithmic transformed energy adjusted intake**

Most contributing food groups to the first RRR factor

Food group	Loading
Fruiting and root vegetables	0.38
Fresh fruits	0.31
Milk and milk products	0.25
Other vegetables	0.24
Leafy vegetables	0.22

Another kind of applications

Pathway from exposure to disease



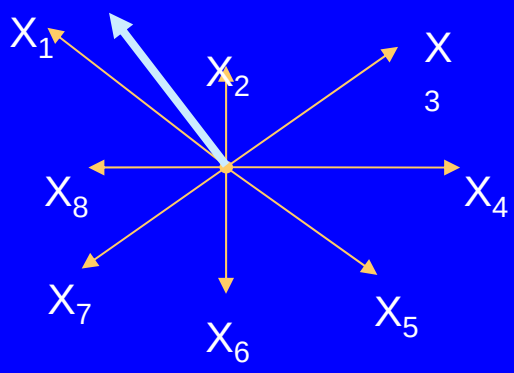
food group intake

biomarker levels

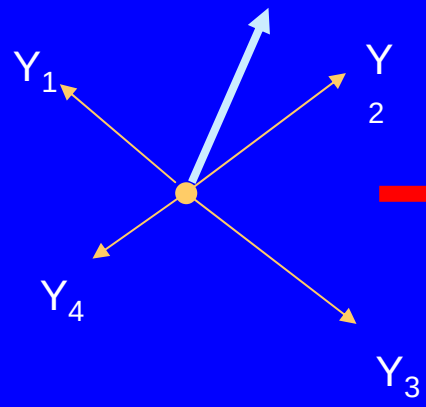
Predictors

Responses

Published results (1)



food group intake

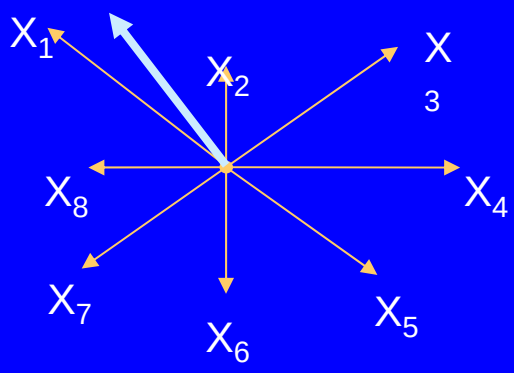


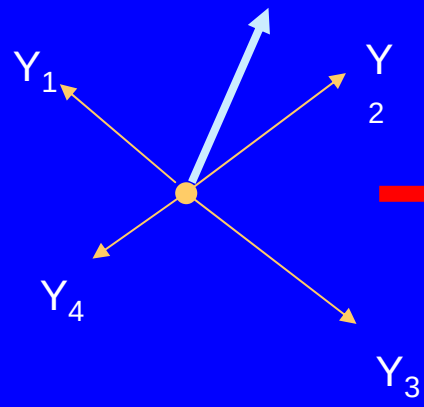
HDL cholesterol
LDL cholesterol
Lipoprotein(a)
C-peptide
C-reactive protein

CHD

Hoffmann et al. Am J Clin Nutr 2004;80:633-40.

Published results (2)





Diabetes

food group intake

C-reactive protein
E-selectin
TNF-alpha Receptor 2
IL_6
VCAM_1
ICAM-1

Schulze et al. Am J Clin Nutr 2005;82:675-84.

CONCLUSIONS

1. PCA, PLS and RRR are similar methods, all starting with eigenvalues and eigenvectors of a covariance matrix and ending with latent variables that are linear functions of the original variables.
2. They are all aimed to explain much variation, but differ in the considered set of variation directions.
3. The outstanding feature of RRR is that it can maximise the explained variation of variables different from the original ones.
4. In applications, RRR should be used if ancillary variables exist which variation is more important than the variation of the original variables.