Singapore Med J 2004 Vol 45(12) : 558

CME Article

Biostatistics 302. Principal component and factor analysis

Y H Chan

Consider the situation where a researcher wants to determine the predictors for the fitness level (yes/no) to be assessed by treadmill by collecting the variables (Table I) of 50 subjects. Unfortunately the treadmill machine in the air-con room has broken down (the participants do not want to run in the hot sun!), and no assessment of fitness could be carried out. What could be done to analyse the data? A descriptive report would be of no value for an annual scientific meeting (ASM) presentation but there is still hope!

Table I. Variables in (hypothetical) fitness study.

X_1: Weight	X_4: Waist circumference	X_7: Diastolic BP
X_2: Height	X_5: Number of cigarettes smoked/day	X_8: Pulse rate
X_3: Age	X_6: Systolic BP	X_9: Respiratory rate

PRINCIPAL COMPONENTS ANALYSIS (PCA)

PCA describes the variation of a set of ***correlated*** multivariate data (X's) in terms of a set of ***uncorrelated*** variables (Y's), known as **principal components**. Each Y is a linear combination of the original variables X. For the example above we have,

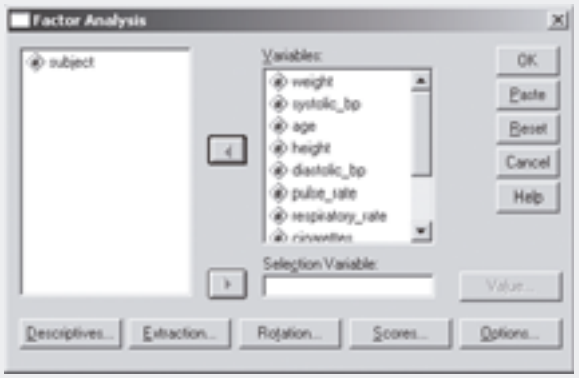
$$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{19}X_9$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{219}X_9$$
etc

In fact, 9 (= the number of X variables) principal components will be available. The aij's (between -1 to 1) are the weights of each X variable contributing to the new Y_i.

Each new Y_i variable is derived in decreasing order of importance, that is, the first principal component (Y_1) accounts for as much as possible of the variation in the original data and so on. The objective is to see whether a smaller set of variables (the first few principal components) could be used to summarise the data, with little loss of information.

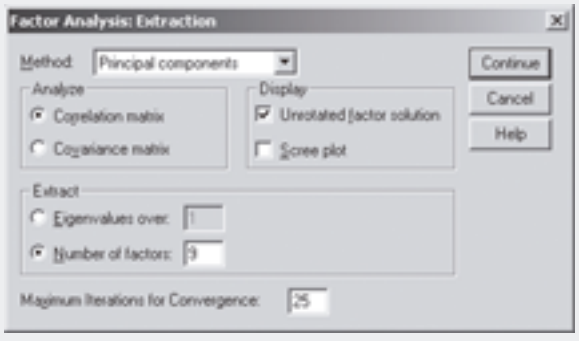
To perform PCA using SPSS, go to Analyse, Data Reduction, Factor to get Template I. Put the variables of interest into the Variables box.

Template I. Defining PCA.



Click on the Extraction folder, choose **Principal components** for the Method option and checked the **Unrotated factor solution** (see Template II). One should Analyse using the **Correlation matrix** (putting all the X variables on an equal footing). This is because the X variables with the largest variances (using the **Covariance matrix**) can dominate the results, since the X variables are of different units of measurements. Number of factors to be extracted = 9 (the total number of variables).

Template II. Extraction method.



Faculty of Medicine
National University of Singapore
Block MD11
Clinical Research Centre #02-02
10 Medical Drive
Singapore 117597

Y H Chan, PhD
Head
Biostatistics Unit

Correspondence to:
Dr Y H Chan
Tel: (65) 6874 3698
Fax: (65) 6778 5743
Email: medcyh@nus.edu.sg

Tables IIa - IIc show the PCA outputs. In PCA, all the variables are given the same weightage during the extraction process (Table IIa).

Table IIa. PCA communalities.

Communalities

	Initial	Extraction
weight	1.000	1.000
systolic_bp	1.000	1.000
age	1.000	1.000
height	1.000	1.000
diastolic_bp	1.000	1.000
pulse_rate	1.000	1.000
respiratory_rate	1.000	1.000
cigarettes	1.000	1.000
waist_circumference	1.000	1.000

Table IIb shows the amount of variance contributed by each component, with the first component explaining (the biggest), in this case, at least 53% of the data and the rest in decreasing order.

Table IIc shows the contribution of each variable to each component (components 7 to 9 have small loadings from the variables - ignored). The first component (PCA 1) has uniform loadings from all the variables and thus describes the unfitness-score (basing on the assumption that the above 9 variables were positively correlated with being unfit) of a subject, the higher the score, the more unfit the person is. The second component (PCA 2) has negative loadings on weight, height and waist circumference – a component to differentiate the physical characteristics. The interpretation of the principal components will be greatly dependent on the person analysing the data. Usually, the first principal component gives the weighted

Table IIb. PCA total variance explained.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.797	53.295	53.295	4.797	53.295	53.295
2	1.401	15.562	68.857	1.401	15.562	68.857
3	1.218	13.538	82.394	1.218	13.538	82.394
4	.604	6.715	89.109	.604	6.715	89.109
5	.552	6.139	95.248	.552	6.139	95.248
6	.172	1.915	97.163	.172	1.915	97.163
7	.156	1.728	98.891	.156	1.728	98.891
8	.077	.853	99.744	.077	.853	99.744
9	.023	.256	100.000	.023	.256	100.000

Extraction method: principal component analysis.

Table IIc. PCA loading of each variable.

Component Matrix[a]

	Component					
	1	2	3	4	5	6
weight	.838	-.362	-.079	-.170	.172	-.315
systolic_bp	.852	.417	.060	-.072	.096	-.028
age	.638	.259	.556	.409	-.171	-.017
height	.669	-.590	.308	.260	-.049	.034
diastolic_bp	.641	.269	-.418	.364	.435	.044
pulse_rate	.806	.036	-.305	-.026	-.484	-.046
respiratory_rate	.820	.275	-.423	-.144	-.158	.119
cigarettes	.547	.356	.598	-.396	.146	.058
waist_circumference	.695	-.636	-.018	-.157	.111	.222

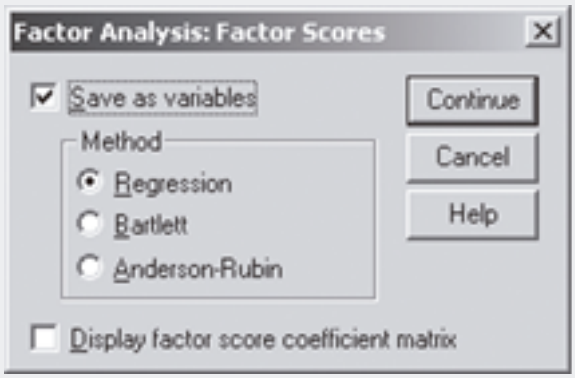
Extraction method: principal component analysis.

[a] 9 components extracted.

average of the data and can often satisfy the investigator's requirements. However, there are situations where the second or third components would be of more interest.

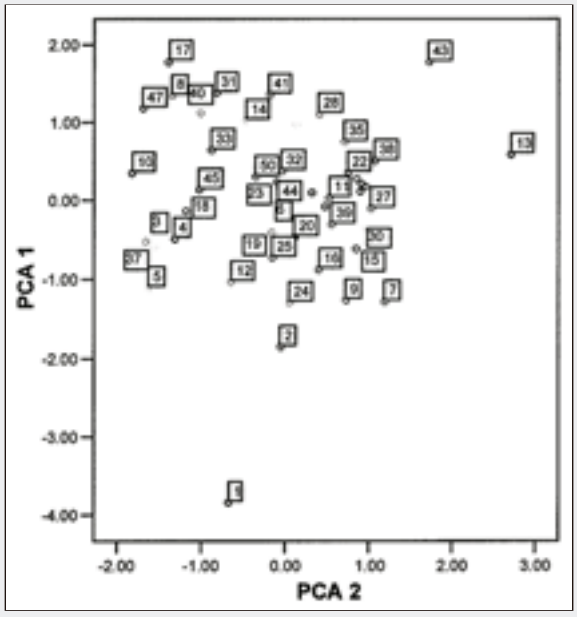
To obtain the calculated scores for the 9 components in Template I, click on the Scores folder to get Template III. Check the "Save as variables" box and choose Method = Regression. SPSS will generate 9 new variables (FAC1_1 to FAC9_1).

Template III. Saving the component scores.



A scatter plot using the first two components gives us an indication of the fitness level for each subject (Fig. 1). Subject 1 had an excellent fitness level and subject 2 displayed good fitness. Subjects 17 and 43 were unfit. PCA 1>0 signifies unfitness and PCA 2<0 shows that the person has a "small built". Further "analysis on fitness" could be performed using PCA 3 which differentiates (demographics: age, smoking and height with vital signs: diastolic, pulse and respiratory – the other variables had low loadings).

Fig. 1 PCA1 vs PCA2.



Number of components retained

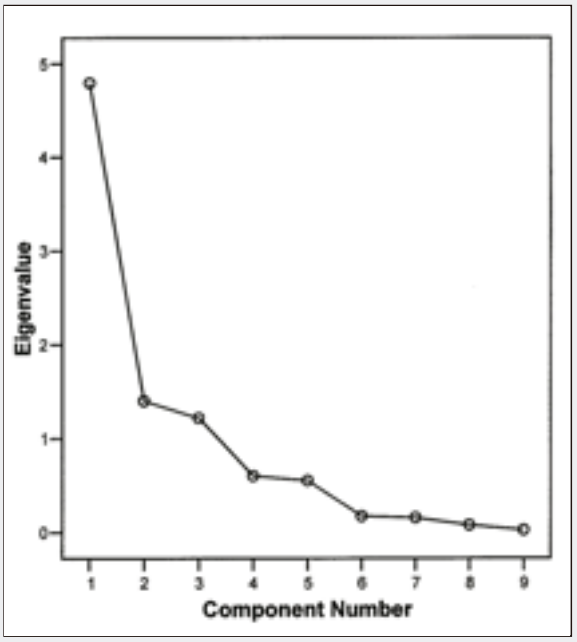
With n original variables, we will obtain n principal components - still have as many new components as original variables except uncorrelated. Often it is desirable to retain a smaller set of the principal components - for easier interpretation of the analysis or for using the components (which are uncorrelated) in a linear[1]/logistic[2] regression analysis to avoid multicolinearity problems. There are a number of approaches (generally used):

1. Retain all components with eigenvalues $\geq$1.0 (Components that have a substantial contribution to original data). In this case, three components will be retained explaining 82.39% of the total variance for the above example.

2. The 80% rule. Retain all components needed to explain at least 80% of the total variance; for this case, still three components retained.

3. Scree test (see Template II). This is performed by plotting the eigenvalues with their respective component number and retaining the number of components that come before a break in the plot. In this case, could be two or three components are retained (Fig. 2) – where there is a change of slope in the graph, a subjective judgment at times.

Fig. 2 Scree plot.



FACTOR ANALYSIS

Factor analysis, like PCA, is a variable reduction method and sometimes obtains very similar results; but there is an important difference between the two techniques. PCA is simply reducing the number of

Table III. Logistic regression to predict fitness.

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	weight	3.768	3762.103	.000	1	.999	43.289
	systolic_bp	-12.366	4315.025	.000	1	.998	.000
	age	-.909	7623.051	.000	1	1.000	.403
	height	7.159	3131.472	.000	1	.998	1285.228
	diastolic_bp	-3.654	16314.451	.000	1	1.000	.026
	pulse_rate	-12.870	13771.826	.000	1	.999	.000
	respiratory_rate	14.258	5709.714	.000	1	.998	1556137.9
	cigarettes	2.801	2271.705	.000	1	.999	16.468
	waist_circumference	.050	2688.225	.000	1	1.000	1.051
	Constant	-536.114	391711.595	.000	1	.999	.000

[a] Variable(s) entered on step 1: weight, systolic_bp, age, height, diastolic_bp, pulse_rate, respiratory_rate, cigarettes, waist_circumference.

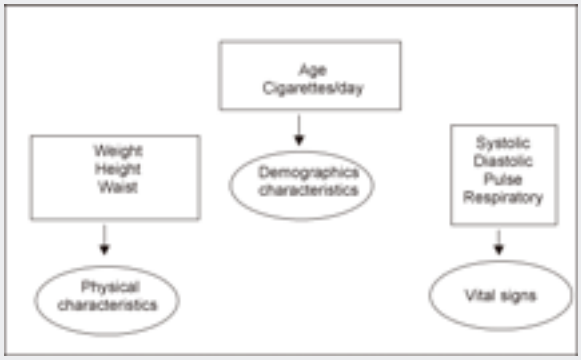
Table IV. Correlations.

	weight	systolic	age	height	diastolic	pulse	respiratory	cigarettes
systolic	0.589							
age	0.310	0.641						
height	0.679	0.314	0.534					
diastolic	0.472	0.612	0.314	0.222				
pulse	0.621	0.611	0.422	0.441	0.449			
respiratory	0.582	0.778	0.322	0.240	0.657	0.865		
cigarettes	0.356	0.641	0.574	0.247	0.144	0.230	0.333	
waist	0.795	0.359	0.202	0.767	0.273	0.478	0.426	0.215

observed variables to a relatively smaller number of components that account for most of the variances in the observed set and makes no assumption about an underlying causal model (which deals with the patterns of the correlations).

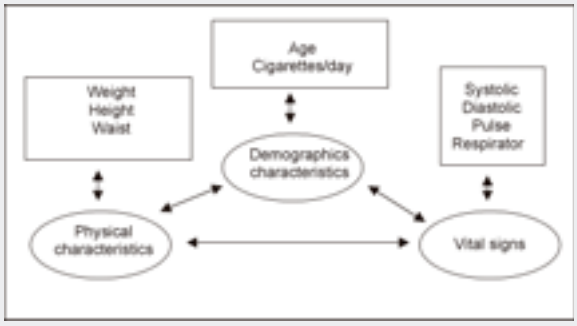
From the above hypothetical fitness example, using PCA allows us to use three (instead of nine, explaining about 82%) components to describe the data (Fig. 3)

Fig. 3 PCA reduction analysis.



Factor analysis assumes that the covariation in the observed variables is due to the presence of one or more factors that exert causal influence on these observed variables. The simple causal structure, using three factors (Fig. 4), is:

Fig. 4 Factor analysis model – causal structure.



Now, on a good clear day, the treadmill machine decides to start working and we are able to get the subjects to be assessed (fitness = yes/no) and using the 9 variables, a logistic regression was performed (Table III).

As expected, multi-colinearity[2] exists (the SE are large) among the variables. In order to get a statistically stable model, we have to sieve out the highly-correlated

variables. Table IV shows the correlations among the variables. The question is: which variable to discard? What could be done without discarding any of the variables and still have a meaning conclusion?

For this situation, PCA is not appropriate as no causal model is assumed. Using factor analysis, how many factors should be included? In PCA, regardless of the number of principal components chosen to be in the analysis, the values of these principal components will remain the same. In factor analysis, the values of the factors are dependent on the number of factors to be used in an analysis but the amount of variance contributed by each corresponding factor will not change. The number of factors to be included would depend on both the amount of variance explained (the higher the better) and the understanding of the causal model.

If we use the eigenvalue criterion (in Template II, click on Eigenvalues over 1) then 3 factors will be obtained. Table Va shows the communalities extraction. The initial extraction is 1.0 for all variables and the final communality shows the proportion of the variance of that variable that can be explained by the common factors.

Table Va. Factor analysis communalities.

Communalities

	Initial	Extraction
weight	1.000	.839
systolic_bp	1.000	.903
age	1.000	.784
height	1.000	.890
diastolic_bp	1.000	.658
pulse_rate	1.000	.743
respiratory_rate	1.000	.927
cigarettes	1.000	.783
waist_circumference	1.000	.889

Extraction method: principal component analysis.

The amount of variance explained by each factor is the same as that explained by the corresponding PCA (see Table IIb). Table Vb shows the unrotated-contribution of each variable for each factor.

Table Vb. Unrotated factor weights.

Component Matrix[a]

	1	2	3
weight	.838	-.362	-.079
systolic_bp	.852	.417	.060
age	.638	.259	.556
height	.669	-.590	.308
diastolic_bp	.641	.269	-.418
pulse_rate	.806	.036	-.305
respiratory_rate	.820	.275	-.423
cigarettes	.547	.356	.598
waist_circumference	.695	-.636	-.018

Extraction method: principal component analysis.
[a] 3 components extracted.

In factor analysis, one of our aims is to be able to determine the latent factors that best explain the data. To interpret the unrotated factors in Table Vb is difficult. We can simplify the loadings of each variable by **rotation** to get a simple structure which will help in the interpretation of the factors. In Template I, click on the Rotation folder (see Template IV), and check the **Varimax** option.

Template IV. Rotation option.

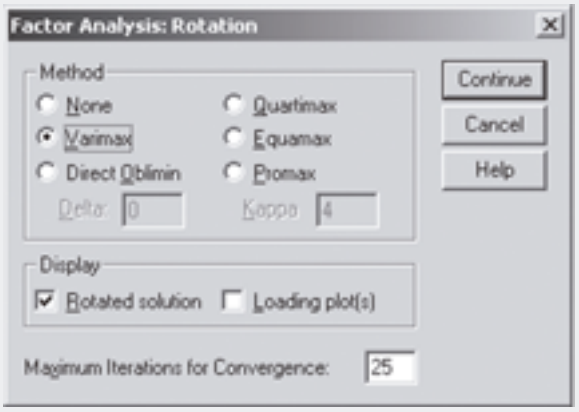


Table Vc shows the contribution of the variance for each rotated factor – the total variance explained (82.39%) and the communalities (Table Va) remained unchanged. The variance explained by the unrotated and rotated components are different.

Table Vc. Rotated total variance.

Total Variance Explained

	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.797	53.295	53.295	2.886	32.062	32.062
2	1.401	15.562	68.857	2.491	27.681	59.744
3	1.218	13.538	82.394	2.039	22.651	82.394

Extraction method: principal component analysis.

Table Vd shows the rotated loadings of each variable for the factors. Observe that the variable "weight" falls into factor 2 as the loading is the biggest in that row (0.761) compared to factor 1 (0.481) and factor 3(0.172). The variable "age" gets into factor 3 and so on.

We can simplify the output by suppressing loadings that are less than 0.5 (or any loading of your choice) for easier interpretation. In Template I, click on the Options folder (see Template V), click on "Suppress absolute values less than" and type "0.5". Table Ve shows the easier-to-interpret loadings. Most of the variables could be easily "assigned" to a factor except for systolic which had high loadings for both factor 1 and factor 3.

One "solution" is to take strictly the higher loading and keep systolic BP in factor 1. Thus the three factors are (Factor 1: systolic, diastolic, pulse, respiratory which we could term as vital signs; Factor 2: weight, height, waist circumference - physical characteristics; and factor 3 : age and cigarettes/day - demographics. It is always not that easy to name the factors!). Another possibility is to ignore the systolic BP and rerun the analysis (Table Vf)

Table Vd. Rotated loadings.

Rotated Component Matrix[a]

	Component		
	1	2	3
weight	.481	.761	.172
systolic_bp	.696	.151	.628
age	.178	.220	.839
height	.036	.898	.289
diastolic_bp	.799	.094	.099
pulse_rate	.752	.391	.160
respiratory_rate	.926	.190	.181
cigarettes	.127	.093	.871
waist_circumference	.242	.911	.031

Extraction method: principal component analysis.

Rotation method: varimax with Kaiser normalisation.

[a] Rotation converged in 5 iterations.

Template V. Options.

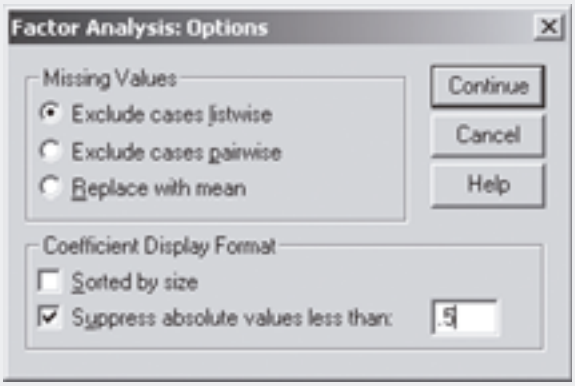


Table Ve. Easier-to-interpret loadings.

Rotated Component Matrix[a]

	Component		
	1	2	3
weight		.761	
systolic_bp	.696		.628
age			.839
height		.898	
diastolic_bp	.799		
pulse_rate	.752		
respiratory_rate	.926		
cigarettes			.871
waist_circumference		.911	

Extraction method: principal component analysis.

Rotation method: varimax with Kaiser normalisation.

[a] Rotation converged in 5 iterations.

Table Vf. Rerun of analysis.

Rotated Component Matrix[a]

	Component		
	1	2	3
weight	.770		
age			.857
height	.884		
diastolic_bp		.802	
pulse_rate		.776	
respiratory_rate		.927	
cigarettes			.858
waist_circumference	.919		

Extraction method: principal component analysis.

Rotation method: varimax with Kaiser normalisation.

[a] Rotation converged in 4 iterations.

The above example is hypothetical and the exercise is to explain the use of PCA and factor analysis. We shall use two real-life studies to show the application of factor analysis (factors that are assumed to actually exist in the person's belief systems, but cannot be measured directly but however will exert an influence on the person's responses to the questions).

The first case-study[3] looked into how healthcare workers were emotionally affected and traumatised during the SARS outbreak and the importance of the institutions to provide psychosocial support and interventions. One part of the study: a coping strategy questionnaire was designed and 6 factors explaining 91% of the variance were obtained (Table VI). Further analysis were then carried out using these factors to determine their "value" in helping the medical personnel's general health questionnaire (GHQ) and impact of events scale (IES) state.

The second study (not published) wanted to look into ways how a supervisor may recognise his/her staff's achievements/ job performance (Table VII).

Table VIIa shows 2, 3 & 4 factor results for the achievement study. The four areas that most people would appreciate to be acknowledged for their achievements were: public verbal, public written, private verbal and pay rise.

Table VI. Coping strategy questionnaire.

	Things that helped me to cope with the SARS situation:	Likert scale 1. Strongly disagree 2. Disagree 3. Not sure probably disagree 4. Not sure probably agree 5. Agree 6. Strongly agree	Factor
1	Clear communication of directives and disease information about SARS		1
2	Precautionary measures taken at work		1
3	Being able to give feedback to hospital management		2
4	Support from hospital administration		2
5	Support from my supervisor/manager/head of department		3
6	Support from my colleagues		3
7	Support from my family		4
8	Being able to talk to someone about my concerns		5
9	My religious convictions		6

Table VII. Staff achievement questionnaire.

	Recognition of achievements	Likert scale 1. Not at all 2. Very little 3. Moderate 4. Considerable 5. Great
Q1	My certification in an area of specialty is acknowledged by a pay rise.	
Q2	My school progress / formal education is acknowledged by a pay rise.	
Q3	My supervisor gives me private verbal feedback for my achievements to help me improves.	
Q4	My achievements are announced in the hospital newsletter.	
Q5	My supervisor sends a letter to the senior management regarding my achievements.	
Q6	My supervisor sent me a letter of congratulation for my achievements.	
Q7	My supervisor congratulates me in front of my colleagues for my achievements.	
Q8	My achievements are posted on the bulletin board.	
Q9	My supervisor brags about my achievements.	

Table VIIa. Factor analysis of the achievement study.

	2-factor (78.3%)		3-factor (83.7%)			4-factor (88.5%)			
	1	2	1	2	3	1	2	3	4
Q1		X		X			X		
Q2		X		X			X		
Q3		X			X				X
Q4	X		X			X			
Q5	X		X					X	
Q6	X		X					X	
Q7	X		X			X			
Q8	X		X			X			
Q9	X		X			X			
Factor Interpretation: Acknowledge by	Public Verbal & Written	Pay rise & Private verbal	Public Verbal & Written	Pay rise	Private verbal	Public Verbal	Pay rise	Public written	Private verbal

Oops, it is only a coincidence that all the three examples above had 9 variables! PCA and factor analyses could handle any number of variables as long as there are correlations amongst the variables. We have discussed on the differences between PCA and factor analysis by using the standard techniques of principal components and varimax for rotated loadings. There are other options for the extraction and rotation techniques (which are based on more restrictive assumptions) – interested readers are encouraged to read-up from any statistical text or seek the help of a biostatistician. Our next article will discuss on the discrimination and clustering of groups/cases – Biostatistics 303. Discriminant and cluster analysis.

REFERENCES

1. Chan YH. Biostatistics 201. Linear regression analysis. Singapore Med J 2004; 45:55-61.
2. Chan YH. Biostatistics 202. Logistic regression analysis. Singapore Med J 2004; 45:149-53.
3. Chan AOM, Chan YH. Psychological impact of the 2003 severe acute respiratory syndrome outbreak on healthcare workers in a medium size regional general hospital in Singapore. Occupational Med 2004; 54:190-6.