Principal Components Analysis

Today . . .

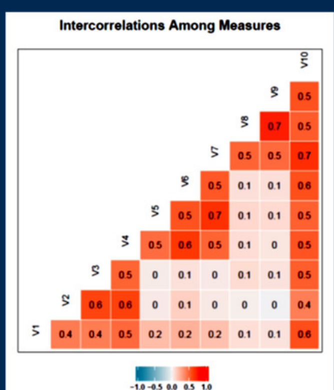
- Simplified composites
- Group contamination
- Reducing multicollinearity
- Some other PCA-like methods
- What to do with ordinal data

Sometimes researchers will use principal components analysis to determine how composite scores should be created, but then will create these scores as simple sums rather than optimally weighted principal component scores.

Why would it matter?

To explore this issue, we will generate a random sample of 500 cases for 9 standard normal variables from a population having moderate correlations (.45 to .70) among items in partially overlapping sets (variables 1-4, variables 4-7, variables, 7-9). A 10th variable is included to serve as an outcome variable to explore the impact of different composites on regression estimates.

$$R = \begin{bmatrix} 1.00 & 0.50 & 0.45 & 0.45 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.55 \\ 0.50 & 1.00 & 0.60 & 0.60 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 \\ 0.45 & 0.60 & 1.00 & 0.55 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.55 \\ 0.45 & 0.60 & 0.55 & 1.00 & 0.50 & 0.60 & 0.50 & 0.00 & 0.00 & 0.55 \\ 0.00 & 0.00 & 0.00 & 0.50 & 1.00 & 0.50 & 0.65 & 0.50 & 0.00 & 0.45 \\ 0.00 & 0.00 & 0.00 & 0.60 & 0.50 & 1.00 & 0.50 & 0.00 & 0.00 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.50 & 0.65 & 0.50 & 1.00 & 0.50 & 0.50 & 0.65 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 1.00 & 0.70 & 0.55 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.70 & 1.00 & 0.60 \\ 0.55 & 0.50 & 0.55 & 0.55 & 0.45 & 0.50 & 0.65 & 0.55 & 0.60 & 1.00 \end{bmatrix}$$



The underlying correlations reflect three overlapping components, all of which are related to a 10th variable.

```
PCA <- principal(Data[, c(1:9)], nfactors = 3, rotate = "none", residuals = TRUE,
     scores = TRUE)
```

```
    PC1   PC2   PC3   h2   u2 com
V1 0.56 -0.38  0.16 0.49 0.51 2.0
V2 0.51 -0.61  0.33 0.75 0.25 2.5
V3 0.48 -0.55  0.43 0.72 0.28 2.9
V4 0.86 -0.32 -0.13 0.86 0.14 1.3
V5 0.65  0.24 -0.52 0.74 0.26 2.2
V6 0.69  0.09 -0.50 0.73 0.27 1.9
V7 0.75  0.52 -0.14 0.85 0.15 1.9
V8 0.39  0.61  0.57 0.85 0.15 2.7
V9 0.37  0.63  0.56 0.84 0.16 2.6

                       PC1  PC2  PC3
SS loadings           3.28 2.02 1.52
Proportion Var        0.36 0.22 0.17
Cumulative Var        0.36 0.59 0.76
Proportion Explained  0.48 0.30 0.22
Cumulative Proportion 0.48 0.78 1.00
```

Three components account for over three-fourths of the variance.

	PC1	PC2	PC3
V1	0.56	-0.38	0.16
V2	0.51	-0.61	0.33
V3	0.48	-0.55	0.43
V4	0.86	-0.32	-0.13
V5	0.65	0.24	-0.52
V6	0.69	0.09	-0.50
V7	0.75	0.52	-0.14
V8	0.39	0.61	0.57
V9	0.37	0.63	0.56

Three non-optimal rules were used:

(a) Use all items but add or subtract depending on the sign of the loading on a component.

Unit_1 = V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9

Unit_2 = -V1 - V2 - V3 - V4 + V5 + V6 + V7 + V8 + V9

Unit_3 = V1 + V2 + V3 - V4 - V5 - V6 - V7 + V8 + V9

	PC1	PC2	PC3
V1	0.56	-0.38	0.16
V2	0.51	-0.61	0.33
V3	0.48	-0.55	0.43
V4	0.86	-0.32	-0.13
V5	0.65	0.24	-0.52
V6	0.69	0.09	-0.50
V7	0.75	0.52	-0.14
V8	0.39	0.61	0.57
V9	0.37	0.63	0.56

(b) Use only those items that load at least .30 in absolute value.

L30_1 = V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9

L30_2 = -V1 - V2 - V3 - V4 + V7 + V8 + V9

L30_3 = V2 + V3 - V5 - V6 + V8 + V9

	PC1	PC2	PC3
V1	0.56	-0.38	0.16
V2	0.51	-0.61	0.33
V3	0.48	-0.55	0.43
V4	0.86	-0.32	-0.13
V5	0.65	0.24	-0.52
V6	0.69	0.09	-0.50
V7	0.75	0.52	-0.14
V8	0.39	0.61	0.57
V9	0.37	0.63	0.56

(c) Use only those items that load at least .50 in absolute value.

L50_1 = V1 + V2 + V4 + V5 + V6 + V7

L50_2 = -V2 - V3 + V7 + V8 + V9

L50_3 = - V5 - V6 + V8 + V9

	PC1	PC2	PC3
PC1	1	0	0
PC2	0	1	0
PC3	0	0	1

	Unit_1	Unit_2	Unit_3
Unit_1	1.000	0.157	-0.057
Unit_2	0.157	1.000	-0.421
Unit_3	-0.057	-0.421	1.000

	L30_1	L30_2	L30_3
L30_1	1.000	-0.150	0.275
L30_2	-0.150	1.000	-0.033
L30_3	0.275	-0.033	1.000

	L50_1	L50_2	L50_3
L50_1	1.000	0.080	-0.431
L50_2	0.080	1.000	0.415
L50_3	-0.431	0.415	1.000

Only the PC scores are independent; the other composites are moderately correlated and not always in a consistent direction.

	PC1	PC2	PC3	Unit_1	Unit_2	Unit_3	L30_1	L30_2
PC1	1.000	0.000	0.000	0.987	0.129	-0.187	0.987	-0.216
PC2	0.000	1.000	0.000	0.057	0.967	-0.226	0.057	0.972
PC3	0.000	0.000	1.000	0.145	-0.194	0.940	0.145	0.061
Unit_1	0.987	0.057	0.145	1.000	0.157	-0.057	1.000	-0.150
Unit_2	0.129	0.967	-0.194	0.157	1.000	-0.421	0.157	0.903
Unit_3	-0.187	-0.226	0.940	-0.057	-0.421	1.000	-0.057	-0.128
L30_1	0.987	0.057	0.145	1.000	0.157	-0.057	1.000	-0.150
L30_2	-0.216	0.972	0.061	-0.150	0.903	-0.128	-0.150	1.000
L30_3	0.141	-0.071	0.978	0.275	-0.237	0.888	0.275	-0.033
L50_1	0.968	-0.097	-0.201	0.921	0.067	-0.341	0.921	-0.315
L50_2	0.186	0.964	0.083	0.247	0.926	-0.169	0.247	0.899
L50_3	-0.228	0.381	0.886	-0.080	0.156	0.779	-0.080	0.474

	L30_3	L50_1	L50_2	L50_3
PC1	0.141	0.968	0.186	-0.228
PC2	-0.071	-0.097	0.964	0.381
PC3	0.978	-0.201	0.083	0.886
Unit_1	0.275	0.921	0.247	-0.080
Unit_2	-0.237	0.067	0.926	0.156
Unit_3	0.888	-0.341	-0.169	0.779
L30_1	0.275	0.921	0.247	-0.080
L30_2	-0.033	-0.315	0.899	0.474
L30_3	1.000	-0.057	0.026	0.808
L50_1	-0.057	1.000	0.080	-0.431
L50_2	0.026	0.080	1.000	0.415
L50_3	0.808	-0.431	0.415	1.000

Smaller components are not preserved as well in the simpler composites.

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0154	0.0000	0.0188	-0.82	0.41
PC1	0.8427	0.8645	0.0188	44.73	< 2e-16
PC2	0.1229	0.1261	0.0188	6.53	1.7e-10
PC3	0.2213	0.2270	0.0188	11.74	< 2e-16

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01538	0.00000	0.01690	0.91	0.36
Unit_1	0.16546	0.89207	0.00326	50.82	< 2e-16
Unit_2	0.03568	0.14802	0.00466	7.66	9.8e-14
Unit_3	0.04978	0.18061	0.00527	9.45	< 2e-16

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01775	0.00000	0.01801	0.99	0.32484
L30_1	0.16678	0.89918	0.00360	46.29	< 2e-16
L30_2	0.02266	0.08580	0.00493	4.59	0.0000056
L30_3	0.02242	0.06815	0.00632	3.55	0.00043

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01194	0.00000	0.02167	0.55	0.58
L50_1	0.22737	0.94863	0.00623	36.51	<2e-16
L50_2	0.01336	0.04150	0.00830	1.61	0.11
L50_3	0.16220	0.40400	0.01143	14.19	<2e-16

Predicting a 10th variable, the magnitude of prediction is not preserved for smaller components.

Principal components have desirable properties (variance maximizing, orthogonal). Short-cut procedures can lead to composite scores that are no longer orthogonal and that are missing important sources of information.

But, minor principal component scores may not replicate well.

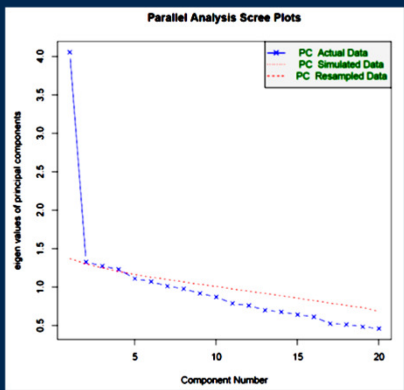
The use of principal components analysis has a hidden danger when used in experimental research. Numerous measures might be collected and principal components analysis might be a reasonable way to simplify the data prior to conducting major analyses.

In experimental data, however, treatment-induced mean differences can impose a structure on the data that my distort a principal components analysis attempting to uncover the underlying dimensionality of the outcome measures.

The sample data contain two experimental groups and 20 measures (N = 500).

A principal components analysis can be used to reduce the set of measures, avoiding redundancy in the significance tests, but it needs to be used correctly.

Let's first see what happens if we ignore the experimental nature of the data.



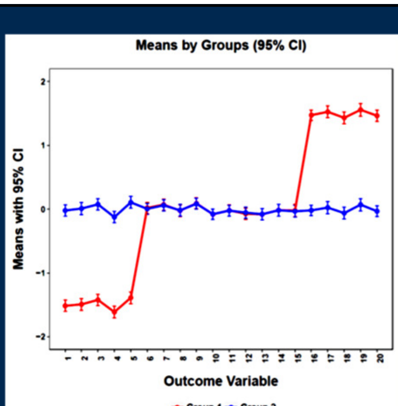
Appears that a single component can be used in place of the 20 original measures.

Standardized loadings (p

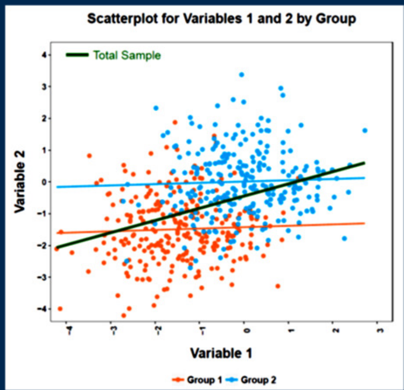
	PC1	h2	u2
v1	0.64	0.405669	0.59
v2	0.64	0.408551	0.59
v3	0.63	0.399740	0.60
v4	0.68	0.463356	0.54
v5	0.62	0.385216	0.61
v6	0.02	0.000483	1.00
v7	0.01	0.000218	1.00
v8	0.01	0.000061	1.00
v9	-0.02	0.000403	1.00
v10	0.03	0.000793	1.00
v11	0.02	0.000523	1.00
v12	0.03	0.000949	1.00
v13	0.02	0.000433	1.00
v14	0.02	0.000299	1.00
v15	0.00	0.000014	1.00
v16	-0.68	0.461628	0.54
v17	-0.62	0.387380	0.61
v18	-0.58	0.331961	0.67
v19	-0.62	0.382573	0.62
v20	-0.65	0.423010	0.58

How would this component be interpreted? Any potential problems in its derivation?

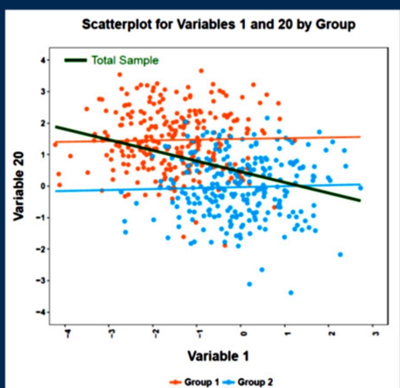
Because the data came from an experiment, there is probably variation in the scores that is due to the manipulation. That variation could be artificially inflating or deflating the correlations among the variables. It needs to be removed ***before*** a principal components analysis is conducted.



The group differences are substantial and could be contaminating the correlations on which the PCA is based.



Within each group, Variables 1 and 2 are unrelated. Ignoring group, the two variables are positively related.



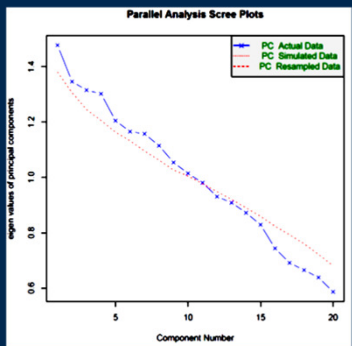
Within each group, Variables 1 and 20 are unrelated. Ignoring group, the two variables are negatively related.

The solution to this problem is to model the group (treatment) contribution and then remove it. Then analyze the residuals in a principal components analysis.

```
M_1 <- lm(v1 ~ group, data = PC_2)
M_2 <- lm(v2 ~ group, data = PC_2)
M_3 <- lm(v3 ~ group, data = PC_2)
M_4 <- lm(v4 ~ group, data = PC_2)
M_5 <- lm(v5 ~ group, data = PC_2)
```

```
PC_R <- cbind(M_1$residuals, M_2$residuals, M_3$residuals, M_4$residuals,
    M_5$residuals, M_6$residuals, M_7$residuals, M_8$residuals, M_9$residuals,
    M_10$residuals, M_11$residuals, M_12$residuals, M_13$residuals,
    M_14$residuals, M_15$residuals, M_16$residuals, M_17$residuals,
    M_18$residuals, M_19$residuals, M_20$residuals)
```

```
scree <- fa.parallel(PC_R[, c(1:20)], fa = "pc")
```



When the correct matrix is analyzed (residuals), there is no evidence for multidimensionality, indicating the need for 20 individual t-tests (and perhaps appropriate Type I error protection).

Principal components analysis can also be used to solve multicollinearity problems.

We will generate a multivariate standard normal data set (N=100) with the following variance-covariance matrix:

$$R = \begin{bmatrix} 1.00 & 0.90 & 0.90 & 0.90 & 0.90 & 0.90 & 0.60 \\ 0.90 & 1.00 & 0.90 & 0.90 & 0.90 & 0.90 & 0.60 \\ 0.90 & 0.90 & 1.00 & 0.90 & 0.90 & 0.90 & 0.60 \\ 0.90 & 0.90 & 0.90 & 1.00 & 0.90 & 0.90 & 0.60 \\ 0.90 & 0.90 & 0.90 & 0.90 & 1.00 & 0.90 & 0.60 \\ 0.90 & 0.90 & 0.90 & 0.90 & 0.90 & 1.00 & 0.60 \\ 0.60 & 0.60 & 0.60 & 0.60 & 0.60 & 0.60 & 1.00 \end{bmatrix}$$

The first 6 variables will be used as predictors of the last variable.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0558     0.0787    0.71     0.48
IV1          -0.0229     0.2607   -0.09     0.93
IV2           0.2914     0.2130    1.37     0.17
IV3          -0.1137     0.2196   -0.52     0.61
IV4           0.1202     0.2368    0.51     0.61
IV5           0.0129     0.2134    0.06     0.95
IV6           0.3738     0.2431    1.54     0.13

Residual standard error: 0.752 on 93 degrees of freedom
Multiple R-squared:  0.488,Adjusted R-squared:  0.455
F-statistic: 14.8 on 6 and 93 DF,  p-value: 8.37e-12
```

Despite the significant overall model, none of the individual predictors is significant.

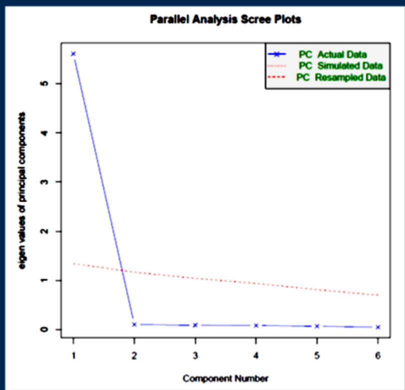
```
vif(lm_fit_1)

##    IV1    IV2    IV3    IV4    IV5    IV6
## 14.554  9.939 10.010 12.062  9.688 12.002

1/vif(lm_fit_1)

##     IV1     IV2     IV3     IV4     IV5     IV6
## 0.06871 0.10062 0.09990 0.08290 0.10322 0.08332
```

The predictors are highly correlated with little unique variance to contribute.



A scree test suggests a single dominant component.

```
Standardized loadings
     PC1   h2    u2 com
IV1 0.97 0.95 0.052   1
IV2 0.96 0.93 0.074   1
IV3 0.96 0.93 0.072   1
IV4 0.97 0.94 0.060   1
IV5 0.96 0.93 0.074   1
IV6 0.97 0.94 0.061   1

                PC1
SS loadings    5.61
Proportion Var 0.93
```

The single principal component accounts for most of the variance in the original set of predictors and can be used to replace those predictors in a multiple regression.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0597     0.0742    0.80     0.42
PC1           0.7028     0.0745    9.43  2.1e-15
```

Original analysis:

```
Multiple R-squared:  0.488,Adjusted R-squared:  0.455
F-statistic: 14.8 on 6 and 93 DF,  p-value: 8.37e-12
```

Analysis with PC score:

```
Multiple R-squared:  0.476,Adjusted R-squared:  0.47
F-statistic: 88.9 on 1 and 98 DF,  p-value: 2.11e-15
```

Of course, the structure of the predictors is quite pure in this example; it won’t always be so clean.

A number of other multivariate procedures resemble principal components analysis:

- Partial least squares regression
- Independent component analysis
- Multiple correspondence analysis

Our last example was a simple case of principal components regression (PCR). In PCR, we reduce the predictor set (X) via principal components analysis and then regress an outcome (Y) on that reduced set. The focus of PCR is entirely on the predictor variables.

Partial least squares regression (PLSR) attempts to identify linear combinations of the predictors that are independent (as in PCR) but that also maximize the covariance with the outcome. Its focus is on both predictors and outcomes. If more than one outcome is present, the method resembles canonical correlation analysis.

Independent component analysis (ICA) sounds like it must be similar to PCA, but is motivated by a different goal and set of assumptions. ICA is often used when the data represent complex signals assumed to be a mixture of multiple independent signal sources. ICA then has the goal of recovering those component signals and their relative contributions. The component signals are further assumed to be non-Gaussian.

ICA would be appropriate if the goal was to reduce the din of crown noise to the individual voice contributions.

Multiple correspondence analysis (MCA) represents a method with similar goals as PCA but is applied to qualitative data in multidimensional contingency tables. The goal is to facilitate a simple understanding of the relationships among the categories of the variables, ideally in a lower dimension space.

An MCA produces descriptive maps that identify the proximity of cases and variable response categories.

What about data that are strictly ordinal but often treated as though continuous? The Need for Cognition Scale used in our first PCA example had a 5-point rating scale:

1 = very characteristic of me
2 = somewhat characteristic of me
3 = neutral
4 = somewhat uncharacteristic of me
5 = very uncharacteristic of me

If the underlying construct is viewed as continuous, then crude categories such as this will attenuate correlations.

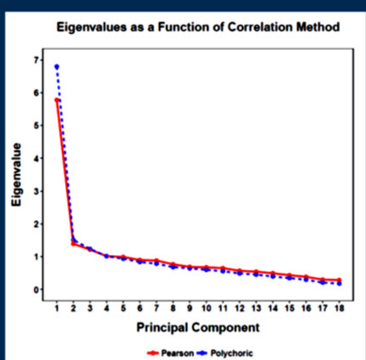
Does it matter that we are assuming the variables to be continuous when they are only approximately so? We can investigate this question by converting the empirical correlations to their expected values for the underlying (and truly) continuous and bivariate normally distributed latent variables. These are called polychoric correlations. Then we can repeat the principal components analysis on the polychoric correlations and compare the results to the original analyses.

The original items must be converted to ordered factors:

```
NC$I1 <- ordered(NC$item_1, levels = c(1, 2, 3, 4, 5))
NC$I2 <- ordered(NC$item_2, levels = c(1, 2, 3, 4, 5))
NC$I3 <- ordered(NC$item_3, levels = c(1, 2, 3, 4, 5))
NC$I4 <- ordered(NC$item_4, levels = c(1, 2, 3, 4, 5))
NC$I5 <- ordered(NC$item_5, levels = c(1, 2, 3, 4, 5))
```

The hetcor( ) function from the polycor package can estimate any of the categorical-to-continuous correlations: biserial, tetrachoric, polyserial, polychoric.

```
PR <- hetcor(NC[, 19:36], ML = TRUE, pd = TRUE)$correlations
```



The polychoric correlations are, on average, higher than the empirical correlations by .057. The higher polychoric correlations result in a clearer first principal component.

	PC1	h2	u2
item_1	0.60	0.364	0.64
item_2	0.74	0.553	0.45
item_3	0.61	0.370	0.63
item_4	0.66	0.436	0.56
item_5	0.70	0.483	0.52
item_6	0.44	0.194	0.81
item_7	0.55	0.297	0.70
item_8	0.49	0.245	0.76
item_9	0.56	0.317	0.68
item_10	0.62	0.387	0.61
item_11	0.68	0.459	0.54
item_12	0.62	0.388	0.61
item_13	0.56	0.312	0.69
item_14	0.55	0.305	0.69
item_15	0.41	0.164	0.84
item_16	0.39	0.156	0.84
item_17	0.55	0.304	0.70
item_18	0.21	0.044	0.96

	PC1
SS loadings	5.78
Proportion Var	0.32

Pearson

	PC1	h2	u2
I1	0.64	0.405	0.59
I2	0.78	0.610	0.39
I3	0.72	0.512	0.49
I4	0.71	0.506	0.49
I5	0.75	0.569	0.43
I6	0.47	0.224	0.78
I7	0.60	0.355	0.64
I8	0.51	0.257	0.74
I9	0.59	0.350	0.65
I10	0.68	0.466	0.53
I11	0.76	0.571	0.43
I12	0.70	0.490	0.51
I13	0.60	0.354	0.65
I14	0.58	0.341	0.66
I15	0.49	0.239	0.76
I16	0.43	0.182	0.82
I17	0.57	0.321	0.68
I18	0.22	0.048	0.95

	PC1
SS loadings	6.80
Proportion Var	0.38

Polychoric

The first principal component accounts for 6% more of variance with polychoric correlations.

Assumptions about underlying latent variables is a reminder that we often are less interested in the measures per se and more interested in what they represent at the construct level.

This emphasis on latent variables underlies factor analysis.

Next time . . .

Exploratory factor analysis