Appendix F: Principal Components (Factor) Analysis

Principal Components Analysis is one of a family of techniques that go under the collective name of *factor analysis*. Unlike most of the statistical techniques we have encountered so far in this book, it has nothing to do with causal modelling of data or with hypothesis testing.

Principal Components Analysis, is an *exploratory* technique. It is used, *not* to trace possible causal connections between independent and dependent variables, but rather to look for underlying patterns (or *latent structures*) in our data. More specifically, we use it to create new 'latent' variables, called *factors*, out of existing 'observed' ones.

The way this is done is by looking at the co-variances between variables to see if we can find an underlying pattern, a 'hidden component' that they seem to share in common and which distinguishes them from other groups of variables. Factor analysis is the technique we use for identifying these hidden components.

What is a factor?

A factor is an artificial variable that correlates highly with several real ones and that is believed to explain some characteristic that they share in common.

Why should we want to create artificial variables? The reason is often that we want to reduce a lot of different but related variables to a smaller and more manageable set of factors that can more readily be analyzed and comprehended. We shall give two examples where this can be useful.

The uses of factor analysis (1) overcoming collinearity

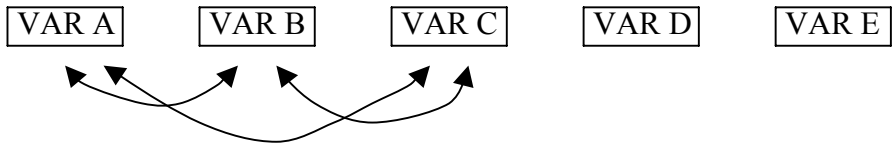
One way in which factor analysis can be helpful to us is in *overcoming collinearity problems* in multiple regression.

We saw in chapter 10 that one of the conditions required by multiple regression modelling is that none of the independent variables should correlate too closely with any of the others. So what are we supposed to do if several of our independent variables are highly inter-correlated?

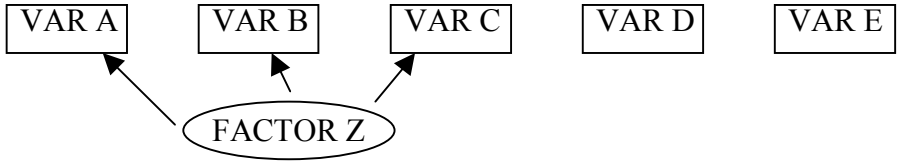
One answer is simply to drop them from the model, but this is not very helpful. A better solution might be to combine them into a new, single variable. After all, if they are all highly correlated with each other, it stands to reason that they might all be measuring much the same thing, but in slightly different ways.

Figure F.1: Creating latent variables to overcome problems of multicollinearity in multiple regression models

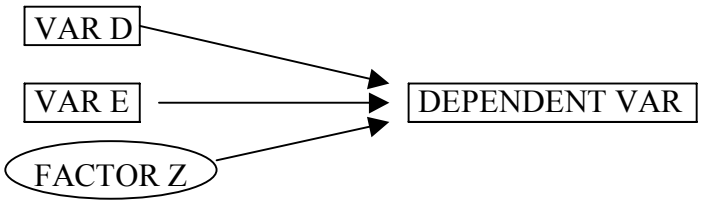
Step 1: We encounter multicollinearity among some of the independent variables we wish to use in a multiple regression model:



Step 2: We look for a common factor:



Step 3: We use this new latent variable instead of the inter-correlated variables in our multiple regression model:



Factor analysis allows us to see whether a number of different observed variables appear to be linked through a common association with one or more underlying factors. If they are, it allows us to measure the relative strength of association between each of our observed variables and the newly-discovered common factors, and we can then use these coefficients to weight the contribution which each of our observed variables should make to the measurement of the new, latent variables.

How all this works will become clearer in a moment, when we look at the results of a factor analysis based on the Smoking Survey data.

The uses of factor analysis (2) testing the validity of the items in an attitude scale

We have already encountered a second use for factor analysis, back in Chapter 5, where we discussed the construction of *attitude scales*.

In Chapter 5 we saw why we might try to measure a phenomenon like 'tolerance of smoking' by asking people lots of different questions – do they think smoking should be banned in restaurants?; do they believe the government is doing enough to discourage smoking?; and so on. We ask questions like these, not because we are particularly interested in the specific answers people give to each individual item, but because we hope that their answers *taken together* will provide us with a way of assessing their *underlying* values and beliefs. Each question thus provides us with a visible indication of people's tolerance or intolerance – it is because they are tolerant or intolerant that they answer in the way that they do.

In the Smoking Survey, we ended up asking eight different attitude questions, and we then added all eight sets of answers together to give us an overall tolerance score. The assumption we made when we did this was that people's answers to these questions all indicate the *same thing* – i.e. that every attitude question is a valid indicator of tolerance or intolerance about smoking. Put in more technical terms, we assumed that each of our observable indicator variables was telling us something about the same unobservable, or *latent*, factor.

But how do we know that all of our observable indicators really are measuring the same underlying factor? Perhaps some of our attitude questions are successful in pinpointing people's tolerance of smoking, while others reflect some other type of concern altogether.

This is where factor analysis can be really helpful. Preferably at the pilot stage, before going on to do a full survey, we can use factor analysis to help us identify which items are strongly measuring a common underlying factor, and which are only weakly associated with it. By looking at the *factor loadings* for each variable, we can select the items which really do seem to be measuring what we are after, and discard the others.

How does it work?

People's responses to our questions are patterned. Those who oppose smoking at work also tend to oppose smoking in restaurants, for example; those who think that

too much fuss is being made about smoking tend to oppose moves to ban it in public places. And so on.

One way in which we can inspect these patterns in different sets of answers is by looking at a correlation matrix. We saw in Chapter 8, that most of the eight attitude items used in the Smoking Survey correlate quite highly with each other (although one or two of them seemed to stand out a bit from the others). If a set of variables like this did not correlate significantly with each other, there would be no point in doing a factor analysis, for if there is no pattern in people's answers across a given set of items, it clearly cannot be the case that there are any underlying factors that can be said to have generated their responses.

If we do find that a set of variables is significantly inter-correlated, however, then we can use Principal Components Analysis to help us identify the underlying factors that might have produced this pattern of association.

- We look first for the largest amount of variance that is shared in common by our eight different items. The assumption is that evidence of shared variance indicates the existence of a common causal component – if people's answers on one item seem to vary in the same way as they do on another, then something must presumably be creating the pattern that these answers share in common. We therefore 'extract' this common component – it is our first 'principal component' or factor. We measure the amount of shared variance that it accounts for by a statistic known as its *eigenvalue*.

- The next step is to see whether we can find a second principal component that can account for a significant amount of the shared variance that still remains in our data (this is rather like trying to fit a second straight-line on a multidimensional scatterplot, the condition being that it must run through the origin at right-angles to the first – i.e. it is *orthogonal*). This second component is then also 'extracted', and the total amount of shared variance that it accounts for (its eigenvalue) is calculated.

- We continue in this way, finding and extracting further components, each of which accounts for a diminishing proportion of the shared variance, and each of which is orthogonal to each of the others. The process ends when we have extracted as many components as there are variables, at which point, all the shared variance has obviously been accounted for. We end up with (in this case) eight components, or factors, all of them quite distinct from each other.

This all sounds very abstract! Let us look at an example.

Extracting the principal components

In the Smoking Survey, we added people's scores on all eight attitude items together to give every respondent a total score between 8 (indicating extreme tolerance) and 40 (indicating extreme intolerance).

By adding all these scores together in this way, we assumed that all eight items are equally valid indicators of tolerance. However, in Chapter 8 we inspected the correlation matrix for these attitude variables which demonstrated that some items are more highly inter-correlated than others. Six of the eight correlated with at least two other items in the matrix at 0.5 or higher, but the other two – var00024 (smoking should be left to personal choice) and var00021 (the government is not doing enough to discourage smoking) – seemed to 'fit in with' the others rather more weakly.

Should one or both of these two items be dropped from the combined scale? By doing a Principal Components Analysis of these eight attitude variables, we should be able to judge whether all eight really are 'loading' on the same common factor, or whether some appear to be measuring one thing, and some another.

Table F.1A: Principal Components Analysis of 8 attitude items: Percent of variance explained by each component

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.847	48.090	48.090	3.847	48.090	48.090	2.543	31.788	31.788
2	1.041	13.011	61.101	1.041	13.011	61.101	2.345	29.314	61.101
3	.833	10.418	71.519						
4	.676	8.453	79.972						
5	.459	5.734	85.706						
6	.442	5.530	91.236						
7	.379	4.736	95.972						
8	.322	4.028	100.000						

Extraction Method: Principal Component Analysis.

Given that we have eight original variables, the analysis begins by identifying eight different components or factors (see Table F.1A, first column). However, inspection of the eigenvalues tells us that most of these are only accounting for a very small proportion of the total variance shared in common by our eight variables, and can therefore be discounted.

We can see from the table that component 1 has an eigenvalue of 3.8. With a total of eight components, this means that it can explain 3.8/8 = 48% of the total variance in people's answers to our eight attitude items. Similarly, factor 2 has an eigenvalue of just over 1, so it can soak up just over $1/8^{th}$ (13%) of the total shared variance remaining. Between them, these two components account for 61% of the total variance in attitude scores.

There is no firm rule about when to stop extracting factors, but there is a common convention that any component with an eigenvalue less than 1 is not worth bothering with. This is because each individual variable has a total variance set at 1, so a factor with an eigenvalue below 1 is actually accounting for *less* variance than each of the individual variables which it is meant to be replacing. In Table F.1A, for example, we can see that the third component has an eigenvalue of just 0.83, which means it can account for 10% of the shared variance – but this is hardly impressive given that each of our original attitude variables would, if treated as a separate component, account for 12.5% (one-eighth).

Identifying the composition of the factors

We now 'know', therefore, that our eight attitude items are indicating the existence of two distinct factors. But what are they?

Table F.1B: Principal Components Analysis of 8 attitude items: Factor loadings before rotation

Component Matrix[a]

	Component	
	1	2
VAR20REV	.614	.432
VAR23REV	.699	.446
VAR24REV	.572	.390
VAR26REV	.830	6.549E-02
Attitude: ban in restaurants	.752	-.415
Attitude: govt not doing enough to discourage	.567	5.160E-02
Attitude: ban in workplaces	.693	-.476
Attitude: should be made illegal	.774	-.312

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Let us leave Table F.1A for a moment, and turn to the next part of the output, Table F.1B. This does not tell us what our two factors mean, but it does tell us how each of our eight original variables 'loads' on each of them.

The numbers in the table are correlation coefficients expressing the strength of association between each of our eight variables and each of the two components. For example, people's answers to the question about banning smoking in restaurants are strongly and positively correlated with component 1 ($r = 0.75$), but are rather more weakly, and negatively, correlated with component 2 ($r = -0.42$).

The two peculiar notations in the component 2 column indicate very small correlations – 6.549E-02 for var26rev, for example, means the decimal point has to be shifted 2 places to the left, giving a correlation of just 0.065.

It is difficult to make much sense of this table, however, for every one of our eight items loads more heavily on component 1 than on component 2. This is not particularly surprising – the first component extracted in a factor analysis will always mop up the lion's share of the common variance, and this may well mean that most or even all of the variables will correlate more highly with it than with subsequent components.

What we need is a technique that brings out the distinction between these two components more clearly. There are a number of ways of doing this, the most common being *orthogonal rotation*.

Rotating the components

What is essentially involved in rotation is that we shift the axes of our graph around so that they line up with the direction of the principal components that we have identified in our data.

Figure F.2: The rotation of axes in principal components analysis

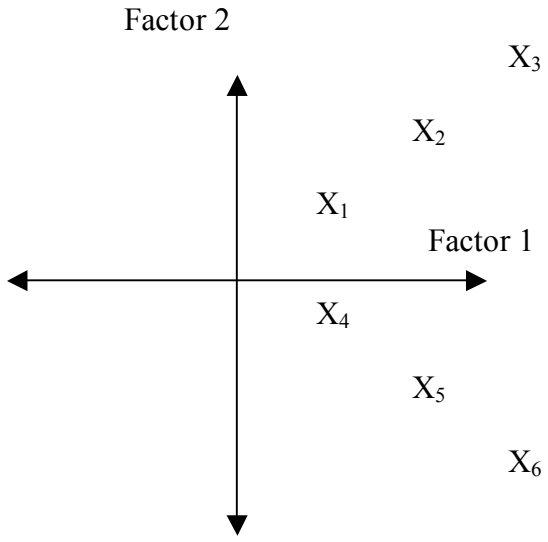


Figure F.2 shows an imaginary graph in which we plot each of six variables against two factors identified in a Principal Components Analysis. As things stand, all six variables seem to load equally on each of the two factors. But if we now turn the axes

of the graph clockwise through 45 degrees, we can align three of the variables (X_1, X_2 and X_3) much more closely with Factor 2 while the other three now align much more closely with Factor 1.

When we rotate our components in this way, the structure of the data remains exactly the same, but we sharpen the patterns of alignment between particular variables and particular components. If we look back to the final columns of Table F.1A, for example, we can see that after rotation, component 1 accounts for 31.8% of the total shared variance (as compared with 48% before rotation), and component 2 accounts for 29.3% (as compared with just 13% before), but that the total variance explained by these two factors is still the same as it was before. Rotation simply aligns the components differently to bring out the differences between them more clearly. The result can be seen in Table F.1C.

Table F.1C: Principal Components Analysis of 8 attitude items: factor loadings after rotation

Rotated Component Matrix[a]

	Component	
	1	2
VAR20REV	.154	.735
VAR23REV	.208	.803
VAR24REV	.152	.675
VAR26REV	.563	.614
Attitude: ban in restaurants	.833	.209
Attitude: govt not doing enough to discourage	.379	.424
Attitude: ban in workplaces	.831	.124
Attitude: should be made illegal	.779	.299

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

This Table can be interpreted much more clearly than Table F.1B, for we can now see that, although all eight of our variables continue to load on both factors, they do so with very different weights:

- Respondents' scores on factor 1 are most heavily influenced by what they said about banning smoking in restaurants (a loading of 0.83), banning smoking at work (0.81) and making it illegal to smoke in public places (0.78).

- Respondents' scores on factor 2 are most heavily influenced by whether they believed that too much fuss was being made about smoking (0.80 on var23rev), whether they thought that society was becoming too intolerant towards smoking (0.74 on var20rev) and whether smokers should be persuaded to quit (0.68 on var24rev).

The remaining two variables seem not to line up so clearly with either factor, however. Conventionally, variables that correlate below 0.3 are dropped as constituent elements of a factor, but both var26rev (on the number of restrictions on smoking) and variable 21 (on the government's role in discouraging smoking) exceed this level on both components. We might therefore decide to include them in both, but if we want to produce 'pure' measures of whatever it is that these two factors are measuring, we might decide to drop these two variables altogether.

The dangers of factor analysis

We started out on this exercise hoping to find that all eight of our attitude items are measuring the same underlying component, but we have ended up identifying two different components in our data, each arguably constituted by just three of our original variables. Every one of our respondents now has a score on each of these factors – but what do these scores mean, and what can we do with these two new latent variables?

We need to think carefully at this point. Factor analysis is a powerful technique, but it has its critics. In particular, critics have argued:

- We can end up with totally meaningless clusters of variables. There is a danger when using factor analysis that we forget that even exploratory research like this needs to be theoretically informed. The danger is that we are tempted to engage in

atheoretical *post hoc* rationalization to try to come up with some plausible reason for treating newly 'discovered' latent variables as measures of some sociologically meaningful characteristic, when in fact they have no sociological significance at all.

- The decisions that we take in factor analysis – particularly on the method of rotation and on how many factors are extracted – are ultimately quite arbitrary, yet they govern the findings that we come up with. We therefore need to use a certain amount of discretion and common sense when interpreting these findings

These are important warnings. Reflecting on the results of our Principal Components Analysis, we clearly need to consider whether the components which we have identified actually *mean* anything, and we need to consider whether there is a compelling case for accepting the existence of two distinct components.

What do the factors signify?

Let us reflect for a moment on what Table F.1C tells us about the composition of our two new factors. Presumably, they cannot *both* be measuring people's tolerance of smoking, for the analysis suggests that we have tapped two distinct dimensions of people's beliefs and values, not one.

What might these two dimensions be? The way to start thinking about this is to look at the final factor loadings (Table F.1C). Reflecting on the nature of the variables which load most heavily on each factor, we might decide, for example:

- The variables that load most strongly on component 1 seem to be all about *prohibition*. The major items shaping scores on this factor are those to do with banning smoking in restaurants, at work, and in public places.

- The variables that load most strongly on component 2, by contrast, seem to have much more to do with a 'live-and-let-live' attitude to life – issues about whether

we are now too intolerant of smokers, whether we should try to get them to stop for their own good, and whether this is an issue worth making a fuss about.

Perhaps, therefore, we should treat our new factor 1 (which has been saved to the final data file under the SPSS default name of FAC1_1) as a measure of people's willingness to use *repression* against other people, while factor 2 (FAC2_1) might come closer to what it was we were after in the first place by providing us with a measure of their willingness to *tolerate* other people's behaviour.

Perhaps! But there are three good reasons for exercising extreme caution at this point!

- The first is that 'tolerance' and 'repression' look suspiciously like opposite sides of the same coin, rather than two distinct underlying factors. Just because our factor analysis can identify more than one common component does not mean that we should override our own common sense! In the end, we might decide that it makes no substantive sense to see the two components identified by our exploratory analysis as designating two different dimensions in the way our respondents think about smoking.

Independent factors or a single dimension?

In psychology, researchers have found distinct, orthogonal factors for things like 'positive affect' and 'negative affect', 'high self-esteem' and 'low self-esteem', 'conservatism' and 'liberalism', and so on, and these different factors are then used as measures of distinct components of people's personalities. Critics, however, claim that it makes little sense to see these pairs as different phenomena, for they look suspiciously like different extremes of the same phenomenon.

W. van Schuur and H. Kiers, 'Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead' *Applied Psychological Measurement*, vol.18, 1994, pages 97-110.

- The second reason for caution is that our two factors may have more to do with the way we asked our questions than with different dimensions of our respondents' beliefs and values. It is noticeable that all of the four items which were originally phrased in a way that was supportive of smoking (variables 20, 23,

24 and 26) load heavily on factor 2; and that three of the four items which were critical of smoking (variables 19, 22 and 25) load heavily on factor 1. This should set alarm bells ringing! Are our two factors really measuring different aspects of people's values about smoking, or are they simply the product of the way we asked our questions?

Are the factors a product of the way we asked our questions?

Back in chapter 3, when we were compiling the questionnaire, we suggested that it is often advisable to avoid implying a 'correct opinion' when asking questions, and that one way to do this is to make some items positive in tone, and some negative. We now see that this may not be such a good idea after all, for in analyzing the results of the Smoking Survey, it seems that the way we phrased our attitude questions may have influenced the pattern of answers that we got.

We suspect that, had we phrased all 8 attitude items 'the same way around', we may well have found much more consistency in the way people answered them – in which case, we may have discovered only one principal component when we analyzed the latent structure of their responses, rather than two. Of course, we might also then have influenced our respondents by appearing to favour either a 'tolerant' or 'intolerant' set of statements – it seems we are caught to some extent between the devil of asking leading questions and the deep blue sea of finding spurious latent components in our answers.

- The third reason for caution is that factor 2 fares rather badly on the criterion of *predictive validity* – i.e. it does not correlate very well with other variables which we would expect to be strongly associated with attitudes of tolerance or intolerance regarding smoking (see chapter 3 on the different criteria of validity). For example, if we look at the association between whether or not people smoke (var00003), and their 'tolerance' as measured by this new factor, we get a correlation which, though significant ($p<0.001$), is quite low ($r=0.27$). This compares with $r=0.45$ for factor 1 – which suggests that factor 1 may be a more valid measure of people's thinking about smoking.

Interestingly, though, neither of these two new latent variables correlates as strongly with people's smoking behaviour as our original – and much simpler – additive scale. When we use a measure of people's overall attitude which is derived by simply adding up their scores on the eight individual items, we

achieve a correlation with their smoking behaviour of r=0.52 – higher than that achieved on either of our two factors.

How many factors should there be?

The fact that our eight-item scale seems to perform better than either of our two new latent variables might suggest that we should retain it as our summary measure of people's tolerance (perhaps modifying it slightly to remove the one or two items which do not inter-correlate as strongly as the others). Perhaps we should simply forget about the two factors produced by the Principal Components Analysis? After all:

- The inclusion of the second component in the Principal Components Analysis was a very marginal decision in the first place – its eigenvalue only just exceeded our threshold point of 1 (explaining 1/8th of total variance);

- Also, the fact that two of our eight items cannot easily be allocated to either of the components suggests that the two-factor solution is not fitting our data very satisfactorily, even after rotation.

The important point to remember about factor analysis is that it will always give you a solution, and when this solution involves two or more components, they will always look radically different because orthogonal rotation ensures maximal differentiation of factor loadings. In the end, however, it is you, and not the computer, who must decide whether this output makes sense and how, if at all, you are going to use it.