Visualising multivariate analysis - an intuitive approach to high dimensional statistical extractions

Stephen Lewis

Abstract

The numerical output of multivariate statistical analyses may extend to a greater number of dimensions than can be comprehended and so may appear abstract and divorced from the original data. A need arises, therefore, for the provision of a more intuitive understanding of the results of such techniques - perhaps of a graphical nature.

A simple method is to plot, what have come to be known as, Andrews' curves. A tabular procedure, using a standard computer spreadsheet, is described whereby the coefficients produced by various multivariate statistical techniques can be substituted into a simple equation to produce a smooth, wave-like curve characterising the source data. Importantly, this technique also provides a means whereby groups of curves may be compared visually to identify clusters and curves of similar or dissimilar overall shape. Similarly, "outliers" may also be spotted.

Introduction

The numerical output of multivariate statistical analyses may extend to a greater number of dimensions than can be comprehended and so may appear abstract and divorced from the original data. A need arises, therefore, for the provision of a more intuitive understanding of the results of such techniques. One of the most useful tools in any data analysis is graph plotting. Most commonly, this involves simple two-dimensional plotting of bivariate data. A method whereby the results of multivariate statistical analysis may be plotted two-dimensionally is to construct what have become known as Andrews' curves — curvilinear representations of multidimensional factors (Andrews 1972).

Procedure

Given the output of a multivariate analysis in the form x_1, x_2, x_3, ..., x_n, each term is substituted into the function:

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots$$

… Eq. 1

The value of $\boldsymbol{f_x(t)}$, at conveniently chosen values of $\boldsymbol{t}$, is plotted within the range -π to +π (-3.142 to +3.142) radians (≡ -180° to +180°). This produces a smooth, two-dimensional curve which may be considered as a linear characterisation of the multivariate data upon which the analysis was originally performed. Figure 1 shows an example of a curve produced in this way. Ideally, the spacing between $\boldsymbol{t}$ values should be as small as practicable - 0.2 radian (or about 5°) or less has been found to give useful curves.

Table 1 gives an example of how a spreadsheet may be used to generate the points to be plotted. Multivariate output values, x_1, x_2, x_3, ..., x_n, are entered in column C and the individual elements of Eq. 1 are calculated vertically within the body of the spreadsheet for a range of ***t*** values which run horizontally. The plot points for those t values are calculated at the base of their respective columns. Table 2 gives the relevant formulae associated with the spreadsheet cells - as the same formulae recur horizontally and vertically, only a selection is shown.

Importantly, this technique provides a way of visually comparing groups of curves, derived from different multivariate data sets, to see if they have similar or dissimilar overall shapes and to see whether they form clusters (Fig. 2). Similarly, "outliers" may also be spotted. One of the important properties of Andrews' curves is that points which are close together in terms of the multivariate analysis remain close for *all* values of ***t*** whilst more distant points will diverge from each other - at least for some values of ***t***. Lines showing the same overall pattern represent similar objects whilst curves which differ represent dissimilar objects.

The variables x_1, x_2, x_3, ..., x_n in Eq. 1 are not equally weighted. Those associated with the low frequency components appearing earlier in the equation tend to be more obvious when plotted than those associated with higher frequency components. It is, therefore, necessary to order the variables used in the plot such that x_1 is the most important variable, x_2 the next important etc. If an order of importance is not known, performing a principal components analysis in conjunction with plotting Andrews' curves has been suggested (Everitt 1978). (Alternatively, a non-rotated factor analysis with the same number of factors to be extracted as variables also provides this information (Manley 1986).)

Andrews' original method was to plot these curves two-dimensionally on paper. This tended to be cumbersome and time-consuming and could lead quickly to cluttered graphs. With the advent of computerised graph-plotting packages and spreadsheets, it is now possible to display Andrews' curves rapidly and with greater control, further enhancing their intuitive rôle.

References

Andrews, D. F. 1972 'Plots of high dimensional data', *Biometrics* 28, 125-36.
Everitt, B. S. 1978 *Graphical techniques for multivariate data*. Heinemann Educational Books, London, pp. 81-7.
Manley, B. F. J. 1986 *Multivariate statistical methods - a primer*. Chapman & Hall, London, p.78.

Figure 1 – A single curve plotted using Andrews' procedures

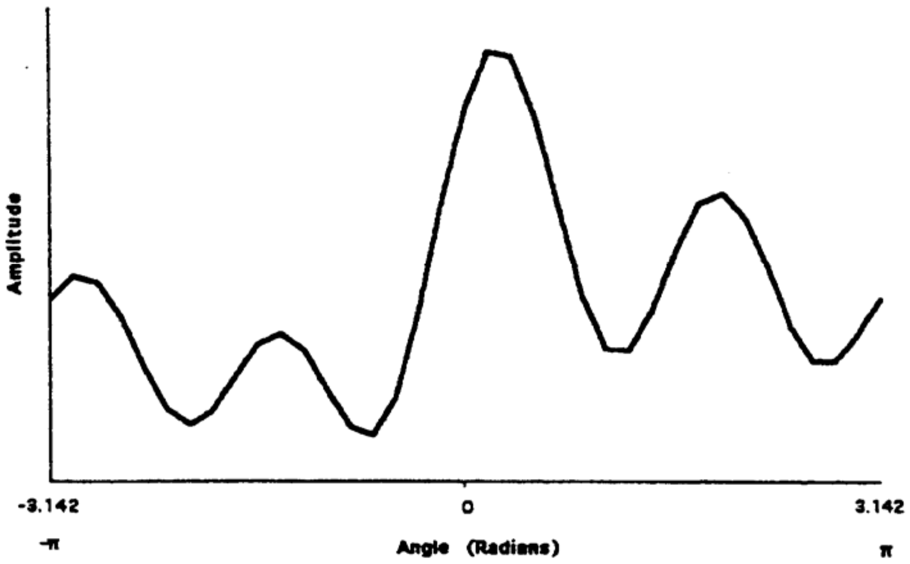


Figure 2

a) Andrews' curves showing similar patterns suggestive of clustering;
b) Andrews' curves showing dissimilar patterns suggestive of unrelated data

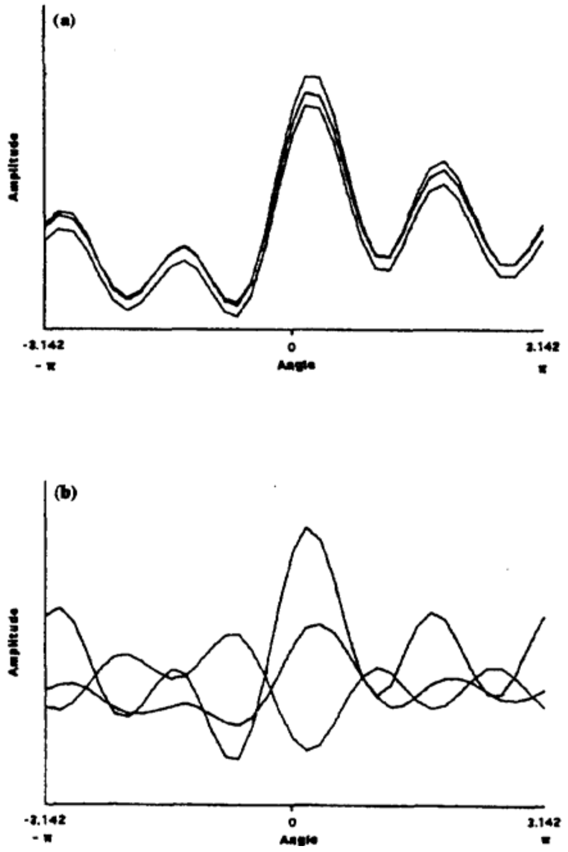
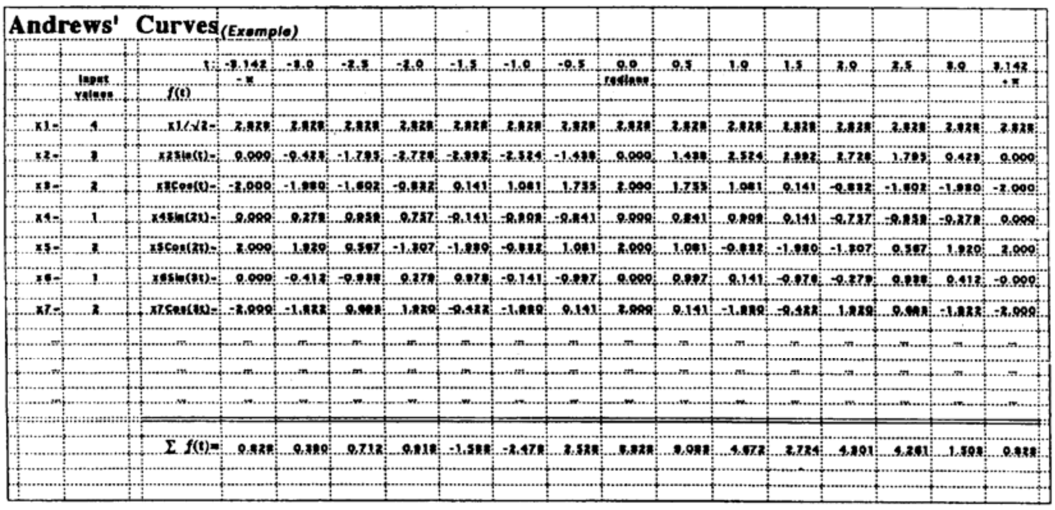


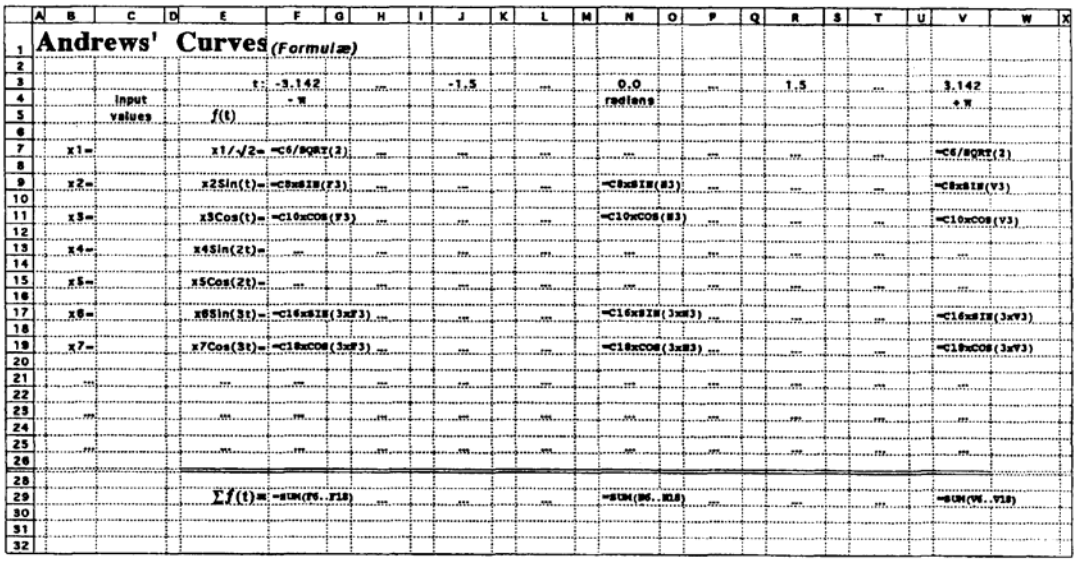
Table 1 – Example of the use of spreadsheet to generate the points to be plotted



Andrews' Curves *(Example)*

		t :	-3.142	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.142
	input		- π							radians							+ π
	values	$f(t)$															
x1 =	4	x1/√2 =	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828	2.828
x2 =	3	x2Sin(t) =	0.000	-0.423	-1.795	-2.728	-2.992	-2.524	-1.438	0.000	1.438	2.524	2.992	2.728	1.795	0.423	0.000
x3 =	2	x3Cos(t) =	-2.000	-1.980	-1.602	-0.832	0.141	1.081	1.755	2.000	1.755	1.081	0.141	-0.832	-1.602	-1.980	-2.000
x4 =	1	x4Sin(2t) =	0.000	0.279	0.959	0.757	-0.141	-0.909	-0.841	0.000	0.841	0.909	0.141	-0.757	-0.959	-0.279	0.000
x5 =	2	x5Cos(2t) =	2.000	1.920	0.567	-1.307	-1.980	-0.832	1.081	2.000	1.081	-0.832	-1.980	-1.307	0.567	1.920	2.000
x6 =	1	x6Sin(3t) =	0.000	-0.412	-0.938	0.279	0.978	-0.141	-0.997	0.000	0.997	0.141	-0.978	-0.279	0.938	0.412	-0.000
x7 =	2	x7Cos(3t) =	-2.000	-1.822	0.693	1.920	-0.422	-1.980	0.141	2.000	0.141	-1.980	-0.422	1.920	0.693	-1.822	-2.000
...		...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...		...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...		...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
		$\Sigma f(t)$ =	0.828	0.390	0.712	0.919	-1.588	-2.478	2.528	8.828	9.082	4.672	2.724	4.301	4.261	1.508	0.828

Table 2 – Formulae associated with spreadsheet cells



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Andrews' Curves					*(Formulæ)*																		
2																								
3					t:	-3.142		...		-1.5		...		0.0		...		1.5		...		3.142		
4			input			- π								radians								+ π		
5			values		*f*(t)																			
6																								
7		x1=			x1/√2=	=C6/SQRT(2)		...		...		...		...		...		...		...		=C6/SQRT(2)		
8																								
9		x2=			x2Sin(t)=	=C8xSIN(F3)		...		...		...		=C8xSIN(N3)		...		...		...		=C8xSIN(V3)		
10																								
11		x3=			x3Cos(t)=	=C10xCOS(F3)		...		...		...		=C10xCOS(N3)		...		...		...		=C10xCOS(V3)		
12																								
13		x4=			x4Sin(2t)=	...		...		...		...		...		...		...		...		...		
14																								
15		x5=			x5Cos(2t)=	...		...		...		...		...		...		...		...		...		
16																								
17		x6=			x6Sin(3t)=	=C16xSIN(3xF3)		...		...		...		=C16xSIN(3xN3)		...		...		...		=C16xSIN(3xV3)		
18																								
19		x7=			x7Cos(3t)=	=C18xCOS(3xF3)		...		...		...		=C18xCOS(3xN3)		...		...		...		=C18xCOS(3xV3)		
20																								
21		...			...	...		...		...		...		...		...		...		...		...		
22																								
23		...			...	...		...		...		...		...		...		...		...		...		
24																								
25		...			...	...		...		...		...		...		...		...		...		...		
26																								
28																								
29					Σ*f*(t)=	=SUM(F6..F18)		...		...		...		=SUM(N6..N18)		...		...		...		=SUM(V6..V18)		
30																								
31																								
32																								