

## Annex: Weighting – More Details for the Technically Minded

This section describes in more detail the principal component weighting process that we use. Let the variables that make up the index be denoted  $i=1,..n$ , and the years for which data is available be denoted  $t=1,..T$ , and let the countries for which data is available be denoted  $j=1,..m$ . Then let  $\mathbf{X}$  be the *data matrix*. It has  $n$  columns, one for each of the variables, and in the  $i$ th column, there are  $mT$  elements, so  $\mathbf{X}$  has  $mT$  rows. The first  $T$  elements are the observations on variable  $i$  in country 1 in years  $t=1,2,..T$ , the next  $T$  elements are the observations on variable  $i$  in country 2 in years  $t=1,2,..T$ , and so on. Let  $\mathbf{x}_i$  be the  $i$ th column of  $\mathbf{X}$ .

Suppose that the globalisation index for any country and year is constructed by attaching a weight to the  $i$ th variable of  $\lambda_i$  and summing the weighted variables (as we do). Then the value of the index for each country and year is given by a  $mT \times 1$  vector  $\mathbf{z}$ , where

$$z = \lambda_1 x_1 + \lambda_2 x_2 + \dots \lambda_n x_n \quad (1)$$

Now – and here is the clever part - the index  $\mathbf{z}$  can be regarded as an approximation to each of the  $i=1,..n$  variables that are used to construct it. In particular, we can approximate  $\mathbf{x}_1$  by some scalar multiple  $\alpha_1$  of  $\mathbf{z}$ ,  $\mathbf{x}_2$  by some scalar multiple  $\alpha_2$  of  $\mathbf{z}$ , and so on. So, if we denote the approximation to  $\mathbf{x}_i$  as  $\hat{x}_i$ , we have

$$\hat{x}_1 = \alpha_1 z, \quad \hat{x}_2 = \alpha_2 z, \dots \hat{x}_n = \alpha_n z \quad (2)$$

We want these approximations to be as good as possible. The goodness of fit measure used in the principal component approach is the sum of squared deviations

$$S = \sum_{l=1}^{mT} \sum_{i=1}^n (x_{il} - \hat{x}_{il})^2 \quad (3)$$

where  $l$  is an index that runs over all  $mT$  observations for a given variable (such as the trade-GDP ratio).

So, the “optimal weighting” problem is to choose weights  $\lambda_1, \dots, \lambda_n$  and  $\alpha_1, \dots, \alpha_n$  to minimise  $S$  in (3) subject to (1) holding - which defines  $\mathbf{z}$  - and (2) holding - which defines the approximations. The solution to the problem therefore chooses weights which make  $\mathbf{z}$  a best approximation to all of the variables  $\mathbf{x}_1, \dots, \mathbf{x}_n$  simultaneously, subject to  $\mathbf{z}$  being a linear combination of the same variables.

The solution to this problem is well-known. It is that both  $\lambda_1, \dots, \lambda_n$  and  $\alpha_1, \dots, \alpha_n$  are equal to the *principal eigenvector* of the  $n \times n$  matrix  $\mathbf{X}'\mathbf{X}$ . This principal eigenvector is easily computed in any statistical package - we used Stata. That gives us the weights  $\lambda_1, \dots, \lambda_n$  in the Table above. And I think we will stop there!

(if you know enough matrix algebra to have got this far, but are unfamiliar with eigenvalues and eigenvectors, you should consult any textbook on linear algebra, such as *Mathematical Methods for Economists*, by Steven Glaister).