We carry out eigenvalue analysis on the following:

(1) Matrix $\boldsymbol{\rho}$.

(2) Matrix **A**, as described in Section 3.8.3.3.

(3) A transformed form of $\boldsymbol{\rho}$ in which $\boldsymbol{\rho}$ is standardized to give $\boldsymbol{\rho}_s$ in which the rows and columns have zero mean and unit variance. The matrix $\boldsymbol{\rho}_s\boldsymbol{\rho}_s^T$ is then computed and subjected to eigenanalysis. It tends to give a lower estimate of cluster numbers than (1).

The most detailed study on cluster counting is that of Milligan & Cooper (1985), and is summarized by Gordon (1999). From this we have selected three tests that seem to operate effectively with powder data:

(4) The Caliński & Harabasz (1974) (CH) test:

$$\mathrm{CH}(c) = [B/(c-1)]/[W/(n-c)]. \quad (3.8.12)$$

A centroid is defined for each cluster. W denotes the total within-cluster sum of squared distances about the cluster centroids, and B is the total between-cluster sum of squared distances. Parameter c is the number of clusters chosen to maximize CH.

(5) A variant of Goodman & Kruskal's (1954) γ test, as described by Gordon (1999). The dissimilarity matrix is used. A comparison is made between all the within-cluster dissimilarities and all the between-cluster dissimilarities. Such a comparison is marked as concordant if the within-cluster dissimilarity is less than the between-cluster dissimilarity, and discrepant otherwise. Equalities, which are unusual, are disregarded. If S_+ is the number of concordant and S_- the number of discrepant comparisons, then

$$\gamma(c) = (S_+ - S_-)/(S_+ + S_-). \quad (3.8.13)$$

A maximum in γ is sought by an appropriate choice of cluster numbers.

(6) The C test (Milligan & Cooper, 1985). This chooses the value of c that minimizes

$$\mathrm{C}(c) = [W(c) - W_{\min}]/(W_{\max} - W_{\min}). \quad (3.8.14)$$

$W(c)$ is the sum of all the within-cluster dissimilarities. If the partition has a total of r such dissimilarities, then $W_{\min}$ is the sum of the r smallest dissimilarities and $W_{\max}$ is the sum of the r largest.

The results of tests (4)–(6) depend on the clustering method being used. To reduce the bias towards a given dendrogram method, these tests are carried out on four different clustering methods: the single-link, the group-average, the sum-of-squares and the complete-link methods. Thus there are 12 semi-independent estimates of the number of clusters from clustering methods, and three from eigenanalysis, making 15 in all.

A composite algorithm is used to combine these estimates. The maximum and minimum values of the number of clusters ($c_{\max}$ and $c_{\min}$, respectively) given by the eigenanalysis results [(1)–(3) above] define the primary search range; tests (4)–(6) are then used in the range $\min(c_{\max}+3, n) \le c \le \max(c_{\min}-3, 0)$ to find local maxima or minima as appropriate. The results are averaged, any outliers are removed, and a weighted mean value is taken of the remaining indicators, then this is used as the final estimate of the number of clusters. Confidence levels for c are also defined by the estimates of the maximum and minimum cluster numbers after any outliers have been removed.

A typical set of results for the PXRD data from 23 powder patterns for doxazosin (an anti-hypertension drug) in which five polymorphs are present, as well as two mixtures of polymorphs, is shown in Fig. 3.8.2(*a*) and (*b*) (see also Table 3.8.2). The scree plot arising from the eigenanalysis of the correlation matrix indicates that 95% of the variability can be accounted for by five components, and this is shown in Fig. 3.8.2(*a*). Eigenvalues from other matrices indicate that four clusters are appropriate. A search for local optima in the CH, γ and C tests is then initiated in the range 2–8 possible clusters. Four different clustering methods are tried, and the results indicate a range of 4–7 clusters. There are no outliers, and the final weighted mean value of 5 is calculated. As Fig. 3.8.2(*b*) shows, the optimum points for the C and γ tests are often quite weakly defined (Barr *et al.*, 2004*b*).

Table 3.8.2

Estimate of the number of clusters for the 23 sample data set for doxazosin

There are five polymorphs present, plus two mixtures of these polymorphs. The maximum estimate is 7; the minimum estimate is 4; the combined weighted estimate of the number of clusters is 6, and the median value is 5. The dendrogram cut level is set to give 5 clusters, and the lower and upper confidence limits are 4 and 7, respectively.

Method	No. of clusters
Principal-component analysis (non-transformed matrix)	5
Principal-component analysis (transformed matrix)	4
Multidimensional metric scaling	4
γ statistic using single linkage	7
CH statistic using single linkage	7
C statistic using single linkage	—
γ statistic using group averages	7
CH statistic using group averages	5
C statistic using group averages	—
γ statistic using sum of squares	—
CH statistic using sum of squares	5
C statistic using sum of squares	—
γ statistic using complete linkage	—
CH statistic using complete linkage	5
C statistic using complete linkage	—

3.8.3.3. *Metric multidimensional scaling*

This is, in its essentials, the particle-in-a-box problem. Each powder pattern is represented as a single sphere, and these spheres are placed in a cubic box of unit dimensions such that the positions of the spheres reproduce as closely as possible the distance matrix, **d**, generated from correlating the patterns. The spheres have an arbitrary orientation in the box.

To do this, the $(n \times n)$ distance matrix **d** is used in conjunction with metric multidimensional scaling (MMDS) to define a set of p underlying dimensions that yield a Euclidean distance matrix, $\mathbf{d}^{\rm calc}$, whose elements are equivalent to or closely approximate the elements of **d**.

The method works as follows (Cox & Cox, 2000; Gower, 1966; Gower & Dijksterhuis, 2004).

The matrix **d** has zero diagonal elements, and so is not positive semidefinite. A positive definite matrix, $\mathbf{A}(n \times n)$ can be constructed, however, by computing

$$\mathbf{A} = -\frac{1}{2}\left[\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right]\mathbf{D}\left[\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right], \quad (3.8.15)$$

where $\mathbf{I}_n$ is an $(n \times n)$ identity matrix, $\mathbf{i}_n$ is an $(n \times 1)$ vector of unities and **D** is defined in equation (3.8.8). The matrix $[\mathbf{I}_n - (1/n)\mathbf{i}_n\mathbf{i}_n']$ is called a centring matrix, since **A** has been derived from **D** by centring the rows and columns.

The eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and the corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are then obtained. A total of p eigenvalues of **A** are positive and the remaining $(n - p)$ will be zero. For the p