EXHIBIT 4
PART

Note that the total scatter matrix does not depend on how the set of samples is partitioned into clusters. It depends only on the total set of samples. The within-cluster and between-cluster scatter matrices do depend on the partitioning, however. Roughly speaking, there is an exchange between these two matrices, the between-cluster scatter going up as the within-cluster scatter goes down. This is fortunate, since by trying to minimize the within-cluster scatter we will also tend to maximize the between-cluster scatter.

To be more precise in talking about the amount of within-cluster or between-cluster scatter, we need a scalar measure of the "size" of a scatter matrix. The two measures that we shall consider are the *trace* and the *determinant*. In the univariate case, these two measures are equivalent, and we can define an optimal partition as one that minimizes S_W or maximizes S_B. In the multivariate case things are somewhat more complicated, and a number of related but distinct optimality criteria have been suggested.

6.8.3.2 THE TRACE CRITERION

Perhaps the simplest scalar measure of a scatter matrix is its trace, the sum of its diagonal elements. Roughly speaking, the trace measures the square of the scattering radius, since it is proportional to the sum of the variances in the coordinate directions. Thus, an obvious criterion function to minimize is the trace of S_W. In fact, this criterion is nothing more or less than the sum-of-squared-error criterion, since Eqs. (33) and (34) yield

$$\operatorname{tr} S_W = \sum_{i=1}^{c} \operatorname{tr} S_i = \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathscr{X}_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = J_e. \tag{38}$$

Since $\operatorname{tr} S_T = \operatorname{tr} S_W + \operatorname{tr} S_B$ and $\operatorname{tr} S_T$ is independent of how the samples are partitioned, we see that no new results are obtained by trying to maximize $\operatorname{tr} S_B$. However, it is comforting to know that in trying to minimize the within-cluster criterion $J_e = \operatorname{tr} S_W$ we are also maximizing the between-cluster criterion

$$\operatorname{tr} S_B = \sum_{i=1}^{c} n_i \|\mathbf{m}_i - \mathbf{m}\|^2. \tag{39}$$

6.8.3.3 THE DETERMINANT CRITERION

In Section 4.11 we used the determinant of the scatter matrix to obtain a scalar measure of scatter. Roughly speaking, this measures the square of the scattering volume, since it is proportional to the product of the variances in the directions of the principal axes. Since S_B will be singular if the number of clusters is less than or equal to the dimensionality, $|S_B|$ is obviously a poor choice for a criterion function. S_W can also become singular, and will certainly be so if $n - c$ is less than the dimensionality d.* However, if we assume that S_W is nonsingular, we are led to consider the criterion function

$$J_d = |S_W| = \left| \sum_{i=1}^{c} S_i \right|. \tag{40}$$

The partition that minimizes J_d is often similar to the one that minimizes J_e, but the two need not be the same. We observed before that the minimum-squared-error partition might change if the axes are scaled. This does not happen with J_d. To see why, let T be a nonsingular matrix and consider the change of variables $\mathbf{x}' = T\mathbf{x}$. Keeping the partitioning fixed, we obtain new mean vectors $\mathbf{m}_i' = T\mathbf{m}_i$ and new scatter matrices $S_i' = TS_iT^t$. Thus, J_d changes to

$$J_d' = |S_W'| = |TS_WT^t| = |T|^2 J_d.$$

Since the scale factor $|T|^2$ is the same for all partitions, it follows that J_d and J_d' rank the partitions in the same way, and hence that the optimal clustering based on J_d is invariant to nonsingular linear transformations of the data.

6.8.3.4 INVARIANT CRITERIA

It is not hard to show that the eigenvalues $\lambda_1, \ldots, \lambda_d$ of $S_W^{-1}S_B$ are invariant under nonsingular linear transformations of the data. Indeed, these eigenvalues are the basic linear invariants of the scatter matrices. Their numerical values measure the ratio of between-cluster to within-cluster scatter in the direction of the eigenvectors, and partitions that yield large values are usually desirable. Of course, as we pointed out in Section 4.11, the fact that the rank of S_B can not exceed $c - 1$ means that no more than $c - 1$ of these eigenvalues can be nonzero. Nevertheless, good partitions are ones for which the nonzero eigenvalues are large.

One can invent a great variety of invariant clustering criteria by composing appropriate functions of these eigenvalues. Some of these follow naturally from standard matrix operations. For example, since the trace of a matrix is the sum of its eigenvalues, one might elect to maximize the criterion function†

$$\operatorname{tr} S_W^{-1}S_B = \sum_{i=1}^{d} \lambda_i. \tag{41}$$

* This follows from the fact that the rank of S_i can not exceed $n_i - 1$, and thus the rank of S_W can not exceed $\Sigma(n_i - 1) = n - c$. Of course, if the samples are confined to a lower dimensional subspace it is possible to have S_W be singular even though $n - c \geq d$. In such cases, some kind of dimensionality-reduction procedure must be used before the determinant criterion can be applied (see Section 6.14).

† Another invariant criterion is

$$|S_W^{-1}S_B| = \prod_{i=1}^{d} \lambda_i.$$

However, since its value is usually zero it is not very useful.

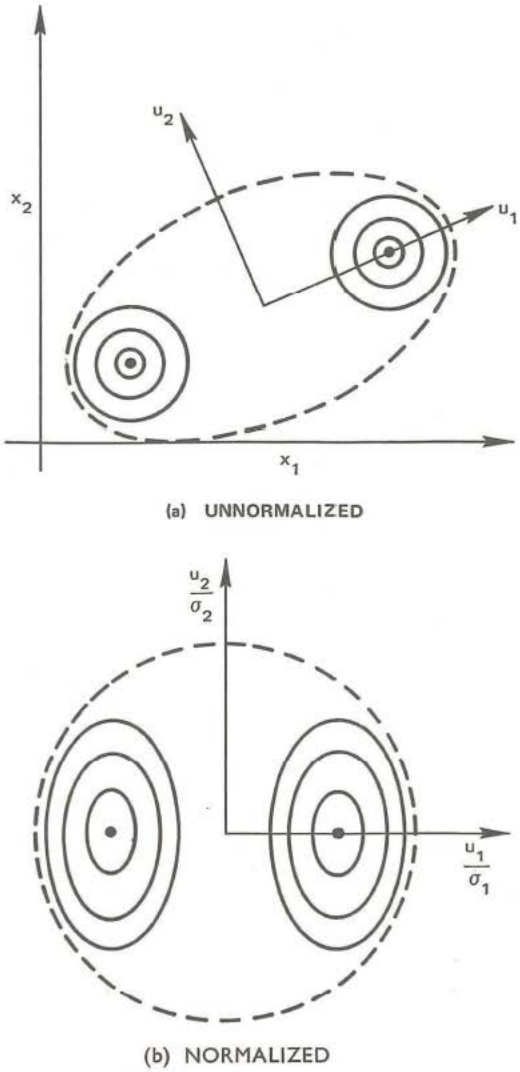

FIGURE 6.14. The effect of transforming to normalized principal components (Note: the partition that minimizes $S_T^{-1}S_W$ in (a) minimizes the sum of squared errors in (b).).

By using the relation $S_T = S_W + S_B$, one can derive the following invariant relatives of tr S_W and $|S_W|$:

$$\operatorname{tr} S_T^{-1} S_W = \sum_{i=1}^{d} \frac{1}{1 + \lambda_i} \tag{42}$$

$$\frac{|S_W|}{|S_T|} = \prod_{i=1}^{d} \frac{1}{1 + \lambda_i}. \tag{43}$$

Since all of these criterion functions are invariant to linear transformations, the same is true of the partitions that extremize them. In the special case of two clusters, only one eigenvalue is nonzero, and all of these criteria yield the same clustering. However, when the samples are partitioned into more than two clusters, the optimal partitions, though often similar, need not be the same.

With regard to the criterion functions involving S_T, note that S_T does not depend on how the samples are partitioned into clusters. Thus, the clusterings that minimize $|S_W|/|S_T|$ are exactly the same as the ones that minimize $|S_W|$. If we rotate and scale the axes so that S_T becomes the identity matrix, we see that minimizing tr $S_T^{-1}S_W$ is equivalent to minimizing the sum-of-squared-error criterion tr S_W after performing this normalization. Figure 6.14 illustrates the effects of this transformation graphically. Clearly, this criterion suffers from the very defects that we warned about in Section 6.7, and it is probably the least desirable of these criteria.

One final warning about invariant criteria is in order. If different apparent groupings can be obtained by scaling the axes or by applying any other linear transformation, then all of these groupings will be exposed by invariant procedures. Thus, invariant criterion functions are more likely to possess multiple local extrema, and are correspondingly more difficult to extremize.

The variety of the criterion functions we have discussed and the somewhat subtle differences between them should not be allowed to obscure their essential similarity. In every case the underlying model is that the samples form c fairly well separated clouds of points. The within-cluster scatter matrix S_W is used to measure the compactness of these clouds, and the basic goal is to find the most compact grouping. While this approach has proved useful for many problems, it is not universally applicable. For example, it will not extract a very dense cluster embedded in the center of a diffuse cluster, or separate intertwined line-like clusters. For such cases one must devise other criterion functions that are better matched to the structure present or being sought.

6.9 ITERATIVE OPTIMIZATION

Once a criterion function has been selected, clustering becomes a well-defined problem in discrete optimization: find those partitions of the set of samples

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase® Smarter legal research.