Clustering by Local Skewering*

David W. Scott

Department of Statistics
Rice University, Houston, Texas 77005
http://www.stat.rice.edu/∼scottdw

March 31, 2005

Abstract

Clustering p-dimensional data by fitting a mixture of K normals has enjoyed renewed interest (for example, see Splus function "mclust"). However, the number of parameters for the model grows rapidly with dimension p. For example, even if all the covariance matrices are assumed to be equal, the number of parameters is $(K-1)+K*p+p(p+1)/2$ for the weights, means and covariance matrix. At ACAS in 2001, Scott introduced the partial mixture component algorithm which fits only one component of the mixture model at a time. This algorithm requires only $1+p+p*(p+1)/2$ parameters for the weight, mean vector, and covariance matrix. In this talk, we introduce a new algorithm which attempts to find the "best" line through individual clusters. This model requires only $2*p-1$ parameters. That is, the new algorithm is linear rather than quadratic in p. By repeatedly reinitializing the search algorithm, all clusters may be identified. Intuitively, the line found is approximately the largest eigenvector of the local covariance matrix. The GGobi visualization program will be used to illustrate the success of this algorithm on real and simulated data.

1 Introduction

Exploratory data analysis and its development owe much to problems and support of the Army scientists and the Army Research Office. One of the mainstays of exploratory analysis of multivariate data is the principal components technique for dimension reduction. For data $\mathbf{x}_k \in \Re^p$, the sample covariance matrix, S, is estimated and its eigenvalues and eigenvectors computed. The eigenvalues are examined in order to determine the number of dimensions, $p' << p$, to retain (through a scree plot, for example). Finally, the data vectors are projected onto the corresponding p' eigenvectors. If the data follow a multivariate normal distribution, even approximately, then investigation of the principal components (rather than the raw data) is extremely useful as a first step.

Even for large dimensions, p, estimation of S is no problem. However, even with today's computing power, finding all of the eigenstructure often leads to software failure. As an extreme example, the data vector could represent a 1000×1000 gray scale image, so that the covariance matrix is a million by a million. Even a numerically stable approach of avoiding the formation of the covariance matrix

*Research supported in part by NSF grant DMS 02-04723 (non-parametric methodology) and NSF contract EIA-9983459 (digital government). Presented October 29, 2003, Army Conference on Applied Statistics, Napa Valley, California

by computing the singular value decomposition of the data matrix, $\mathbf{X}$, is not computationally feasible. (Recall that the SVD of $\mathbf{X} = \mathbf{UDV}^T$ and that the eigenvectors of the covariance matrix are contained in the matrix, $\mathbf{V}$.) Fortunately, there exists specialized software that computes the first p' singular vectors (ARPACT; see Maschhoff and Sorensen, 1996). Since we only require the first p' singular vectors (or eigenvectors), the remaining $p - p'$ eigenvectors need not be computed.

However, this happy situation does not address a number of important problems in data reduction. In particular, multivariate data are much more likely in Army and applied settings to come from a mixture of normal densities, rather than just a single normal density. Of course, statisticians can use the EM algorithm (Dempster et al, 1977) to fit a mixture of normals. But with large problems, estimation of the covariance matrices cannot be avoided. Furthermore, we seem to need to simultaneously estimate the covariance matrices. The SVD approach will not help us here. Whereas ARPACK can find the p' singular vectors, each of length n, the EM approach requires the estimation of K covariance matrices, each of size $n \times n$.

In the following sections, we think about the unthinkable. Can we estimate individual components in a normal mixture without simultaneously having to estimate the other $K - 1$ components? Of course, we are still stuck estimating an $n \times n$ matrix. The second question we consider is the possibility of estimating a few singular vectors without the estimation of S at all. Affirmative answers are shown for both. Computational challenges still remain, but the framework for the optimization problem is provided.

2 Partial Mixture Estimation

Mixture estimation by EM is well-studied; see Titterington et al. (1985). General alternatives to likelihood criteria exist, for example, minimum distance estimation (Beran, 1977). The use of integrated squared error as an estimation criterion has also been considered by Terrell (1990), Basu et al. (1998), and Scott (2001). Given a model, $f_\theta(x)$, and data from the true but unknown density, $g(x)$, we seek to find θ which minimizes

$$\int_{-\infty}^{\infty} [f_\theta(x) - g(x)]^2 \, dx$$

or

$$\int_{-\infty}^{\infty} f_\theta(x)^2 dx - 2 \int_{-\infty}^{\infty} f_\theta(x) \, g(x) \, dx + \int_{-\infty}^{\infty} g(x)^2 dx.$$

An unbiased risk estimate is given by

$$\int_{-\infty}^{\infty} f_\theta(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f_\theta(x_i),$$

where the final term is an unbiased estimate of $2 \int f_\theta(x) g(x) dx$. The integral, $\int g(x)^2 dx$, does not depend upon the unknown parameter, θ, and so may be ignored. If the L_2 norm of the model, $f_\theta(x)$, exists in closed form, then the criterion may easily be minimized numerically. Scott (2001) called the estimator the L_2E estimator, since integrated squared error is in fact the L_2 norm.

Recently, the estimation of normal mixture densities by L_2E was described by Scott (1999, 2004). For example, if the model is the 5-parameter mixture,

$$f_\theta(x) = wN(\mu_1, \sigma_1^2) + (1 - w)N(\mu_2, \sigma_2^2),$$

then the L_2E criterion is easily seen to be

$$\frac{w^2}{2\sqrt{\pi}\sigma_1} + \frac{(1-w)^2}{2\sqrt{\pi}\sigma_2} + 2w(1-w)\phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) - \frac{2}{n} \sum_{i=1}^{n} f_\theta(x_i).$$

Similar expressions exist in the multivariate normal case.

The L_2E technique has a number of interesting (and unique) features. First, it shares the robustness property of all minimum distance techniques. For example, in Figure 1, a single normal density is fitted to a 2-component mixture by L_2E. Rather than compromising over the two components as in MLE, the L_2E estimator focuses on the larger component, ignoring the smaller component. This practical behavior is our solution to the problem of finding individual components. Wojciechowski and Scott (1999) report a number of simulations comparing L_2E and other robust estimators of location.

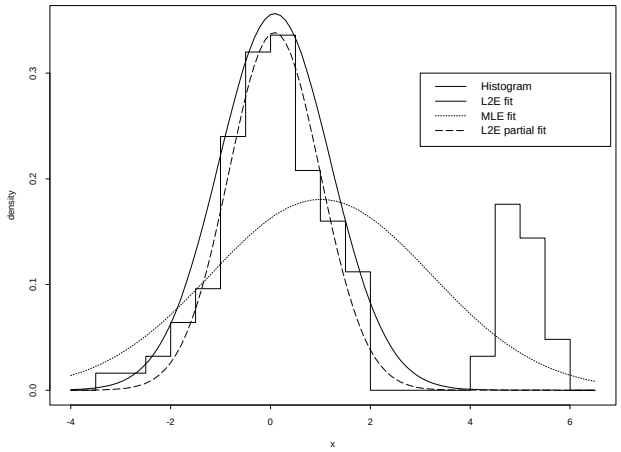

Figure 1: Histogram of 125 points from the mixture $0.8\,N(0,1) + 0.2\,N(5,1)$. Also shown are the maximum likelihood and L2E fits using the incorrect model $N(\mu, \sigma^2)$. Finally, the L2E fit of the 3-parameter model $w \cdot N(\mu, \sigma^2)$ is shown.

However, a second and unique feature of L_2E is also displayed in this figure. In the derivation of the L_2E criterion, the fact that $g(x)$ is a true (if unknown) density was of critical importance in order to estimate the integral $\int f_\theta(x)\, g(x)\, dx$ in the L_2E expression. However, the fact that the estimator, $f_\theta(x)$, is a true density is not used. Thus, we propose to use estimators that are not (complete) densities. For example, the second L_2E estimator in Figure 1 uses the 3-parameter normal model, $w\,N(\mu, \sigma^2)$. This equation is called a partial density component (PDC) model. Notice that the area of this PDC L_2E estimate is in fact less than 1.0, and very close to the true value of 0.8 for the left component. (Of special interest is the fact that L_2E can estimate the right component just as well. Which component L_2E converges to is a function of the initial guess for the parameter vector, θ. Since the value of $\hat{w}$ is about 0.20, the usual robust theory about breakdown points never less than 0.50 must be relaxed.)

A similar example, but in two dimensions, is shown in Figure 2, together with the estimated value of $\hat{w}$. The 6 parameters of the MLE fit (with $w = 1$) were used as initial guesses in the L_2E iterations.

The PDC model can have more than one component. The L_2E estimate found depends entirely upon the initial guess for θ. In practice, a large number of guesses for θ are found by sampling, and the most commonly occurring solutions examined carefully. In Figure 3, we show eight such solutions for the Old Faithful Geyser dataset, which has been lagged and blurred to avoid rounding errors. Clearly these data have three components. Depending upon the choice for θ, the fits may find individual components or combine pairs. Thus we have provided a solution to the vexing problem of mixture estimation when the number of components is unknown. Useful estimates can be found when the number of components is

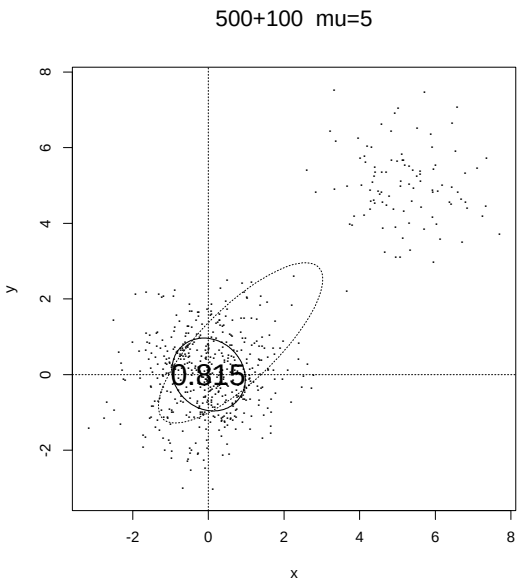

Figure 2: MLE and PDC L_2E contours.

underestimated, even severely. (Of course, if the number of components is overestimated, L_2E will suffer the same fate as MLE and overfitting will result.)

3 Skewers and Principal Components

As is well-known, the first principal component gives the univariate projection of the data with largest variance. In Figure 4, we look at an example of principal components for two variables of the Fisher/Anderson Iris data. Notice that the axis for the principal components goes through the origin (of course).

Principal components also solves a related problem, which is not often used for motivation. Consider finding a set of points, constrained to lie on a line in $\Re^p$, that are closest to the original data. The solution is provided by the points on the first principal component, where the line is shifted away from the origin to go through the sample mean, $\bar{\mathbf{x}}$. In Figure 5, we show the "skewered" version of the data shown in Figure 4.

Of course, there are 3 species of flowers in the Iris data, so that 3 skewers may be computed. The first principal component for each species, but centered at the mean for each species, is shown in Figure 6. Of course, it is instructive to visualize all four "skewers" for each of the 3 species; see Figure 7.

As instructive as these figures (and animated versions in *ggobi*) are, we are estimating the covariances matrices separately and then computing the eigenvectors of each. Can we find a criterion that is attracted to a skewer without going through the covariance calculation or estimation? Let us look closely at the line segments shown in Figure 5. Clearly, for the Iris Setosa species (for which the skewer was estiamted), the distances between the raw data and their projections onto the skewer are quite small, compared to the projections of the Iris Versicolor and Iris Virginica species onto the Setosa skewer. A histogram of these 150 distances is shown in Figure 8. If we compute the Iris Setosa skewer using all 4 variables, we obtain the distance-to-skewer histogram shown in Figure 9.

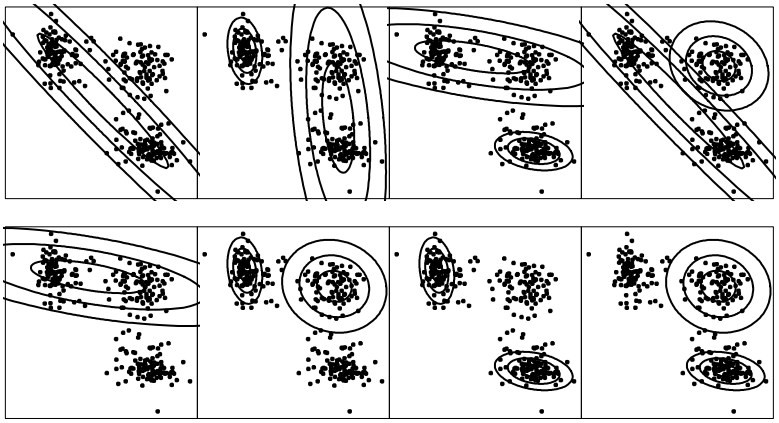

Figure 3: Examples of one- and two-component PDC L_2E fits. (top) The weights in each frame are (.78), (.25, .69), (.68, .28), and (.75, .30). (bottom) The weights in each frame are (.68), (.25, .32), (.25, .28), and (.32, .28).

What are the essentials of a skewer? Like the first principal component, a skewer has a direction, $\mathbf{v}$. While the principal component goes through the origin, the skewer goes through a general point, P. (If the data are labelled, we know that we can take P to be the sample mean of any single group.) For data in $\Re^p$, the dimension of the point P is p, while the dimension of the direction vector, $\mathbf{v}$, is $p-1$. Thus the dimension of the skewer in terms of unknown parameters is $2p-1$. Thus, the dimension of the search for a skewer grows linearly with dimension, p, rather than quadratically as for the covariance matrix. Thus we have traded a computationally infeasible search for high-dimensional covariance matrices and associated eigenvectors to a linear search for a skewer. However, many random starts will be required in order to have a reasonable chance at finding some number of skewers.

We have not yet specifically stated what the criterion is for finding the skewer, only how we propose to parametrize the search for it. The answer lies in the bimodal structure of the histogram in Figure 9, which should be compared to the bimodal structure in Figure 1. To make our problem easier, imagine that we are more specific about the point, P, on the skewer. Suppose P is the point on the skewer closest to the origin. (Note, we do not advocate using this choice numerically, as instability may arise if the skewer happens to go through the origin, or nearly so.) We now have a vector, $\mathbf{u}$, which goes from the origin to the new point, P, on the skewer. We can use this vector, $\mathbf{u}$, in order to create an artificial "sign" on the distance from a data point, $\mathbf{x}_k$, to its projection onto the skewer, call it $\mathbf{y}_k$. We do so by taking the inner product of the vector from $\mathbf{y}_k$ to $\mathbf{x}_k$ with the vector $\mathbf{u}$. Thus the distance histogram as shown in Figure 9 will not have only positive values, but the signed distances of the points corresponding to the skewer will be almost exactly symmetric around the origin. The data points not coming from the skewer group (i.e. the Versicolor and Virginica data in our example) will still be farther away from 0, possibly all on one side, but not necessarily in general. By the robustness property of L_2E, we propose to model the distribution of points "in" the cluster by the PDC model, $w\,N(0,\sigma^2)$. Note that by fixing the mean at zero, we are asking the skewer to pierce a cluster of the data. The resulting value of w will indicate the rough size of the cluster the skewer has been attracted to. Note, however, that if the data contains 6 clusters, then depending upon the orientation of the first eigenvector of each cluster and the direction to other clusters, it may or may not be possible to isolate each cluster individually. Also, in high dimensions, the use of the normal model needs to be replaced by something closer to a chi-squared

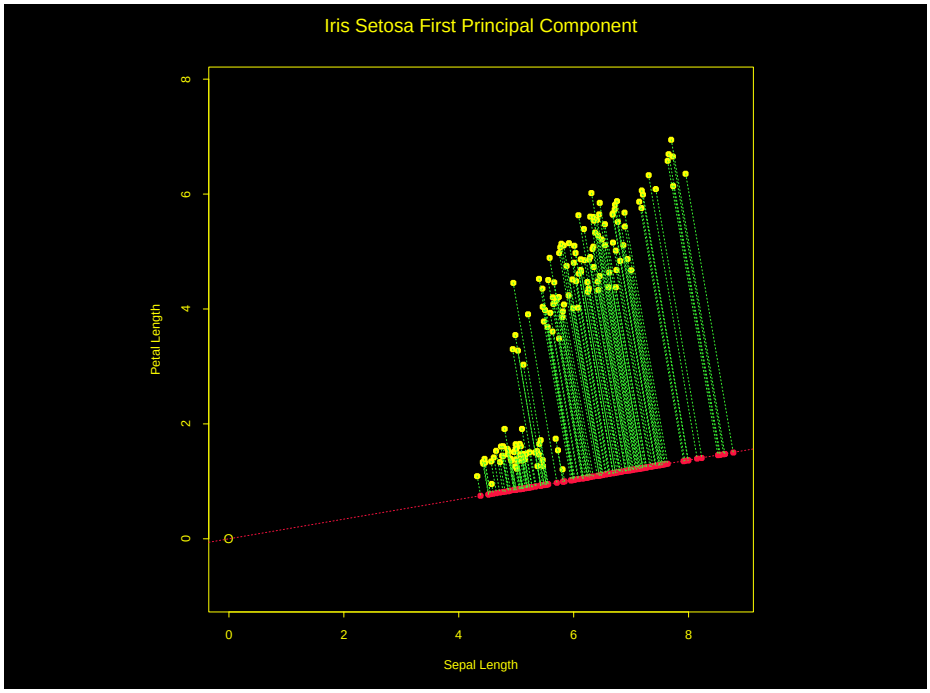

Figure 4: Example of principal components for two variables of the Iris data.

distribution. But if there is a gap in the histogram at the true skewer, then the robustness properties of the L_2E PDC algorithm suggest that a precise model for the PDC is not necessary. (However, the less precise the model, the less precise the estimated values of w and σ will be. Marking points as belonging to the skewer or not relies strongly on at least reasonable value for w and σ) Finally, note that the PDC density model only has 2 parameters, w and σ, no matter the dimension of the data. Of course, the skewer is also part of the estimation, so that the total number of parameters estimated simultaneously by the L_2E PDC skewer algorithm is $(2p-1)+2$ or $2p+1$.

We implemented this algorithm in Splus. For the Iris data (in all 4 dimensions), we found only two skewers. They are shown in Figures 10 and 11 together with the first principal component of the Iris Setosa species. Clearly one estimated skewer is very close to that eigenvector. The other skewer is very close to the first eigenvector for the covariance matrix of all 150 data points (i.e., the overall covariance matrix with no group labels). While a skewer representing just the Versicolor and Virginica species (combined) might be expected, our algorithm always moved away from such an initial orientation to the skewer shown in Figure 11.

4 Extensions

We have limited our discussion to search for one-dimensional skewers. The search for a skewer "plane" or "hyperplane" is a straightforward extension of the algorithm described here. The only difference is that the skewer points, $\mathbf{y}_k$, lie on a hyperplane rather than a line. The criterion is still the distance from the raw data point, $\mathbf{x}_k$ to the point $\mathbf{y}_k$ on the skewer. Note that our "trick" of constructing a signed

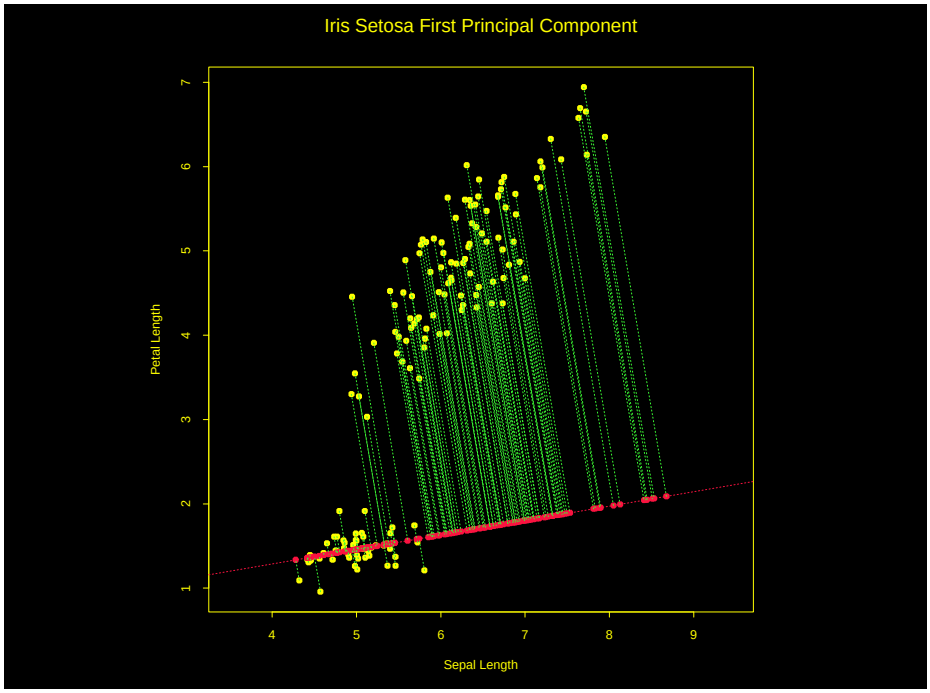

Figure 5: Skewer of the Iris data shown in Figure 4

distance so that the PDC model would be symmetric around the value 0 is now immediately obvious, as the hyperplane divides the space into two parts. The vector $\mathbf{u}$ can be taken as any vector orthogonal to the skewer plane and used to take inner products to assign a sign to each distance computed.

5 Discussion

The ARPACK software allows principal components to be applied to enormously large datasets by avoiding computation of the covariance matrix in order to estimate its eigenvectors. However, if the data set is in fact a mixture of normals, a new attack is required.

In this paper, we have shown how individual mixtures may be estimated without having to estimate or identify all clusters using the L_2E criterion and PDC model. However, such an approach is still quadratic in the number of dimensions, p. But by utilizing a very simple 2-parameter PDC model on the distances from points to their projection onto the skewer, we have demonstrated the existence of a criterion that is linear in the number of dimensions, p. Many random initializations are suggested in order to obtain a reasonable coverage of interesting skewer solutions. But such a task is happily easily accomplished with a farm of parallel computers and requires almost no sophisticated programming tools.

Finally, the Army has a long history of supporting advanced statistical tools and visualization support, beginning with the PRIM9 work of John Tukey and colleagues. The high-dimensional data faced by researchers and workers today requires a whole array of new tools and out-of-the-box thinking. I have tried to illustrate how the use of minimum-distance criteria can free one from the usual set of behaviors into a new realm where seemingly impossible tasks may in fact be successfully addressed.

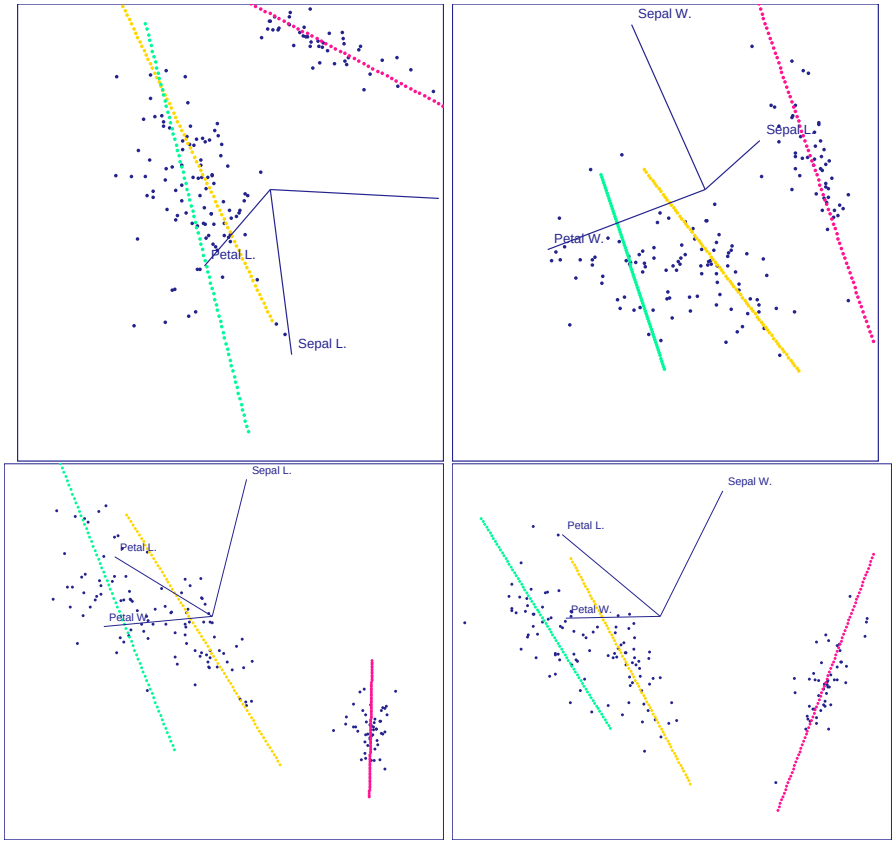

Figure 6: "4-D Skewers" of 3 Iris species in $\Re^3$ (top left: variables-123; top right: variables-124; bottom left: variables 134; bottom right: variables-234).

6 References

Basu, A. and Harris, I.R., Hjort, N.L. and Jones, M.C. (1998), "Robust and Efficient Estimation by Minimising a Density Power Divergence," *Biometrika*, 85, 549–560.

Beran, R. (1977), "Robust Location Estimates," it The Annals of Statistics, 5, 431–444.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–37.

Maschhoff, K.J. and Sorensen, D.C. (1996), "P-ARPACK: An Efficient Portable Large Scale Eigenvalue Package for Distributed Memory Parallel Architectures," in *PARA*, pp. 478–486.

Scott, D.W. (1999), "Remarks on Fitting and Interpreting Mixture Models," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, Eds., 31, 104–109.

Scott, D.W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics*, 43, 274–285.

Scott, D.W. (2004), "Partial Mixture Estimation and Outlier Detection in Data and Regression," in *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S.

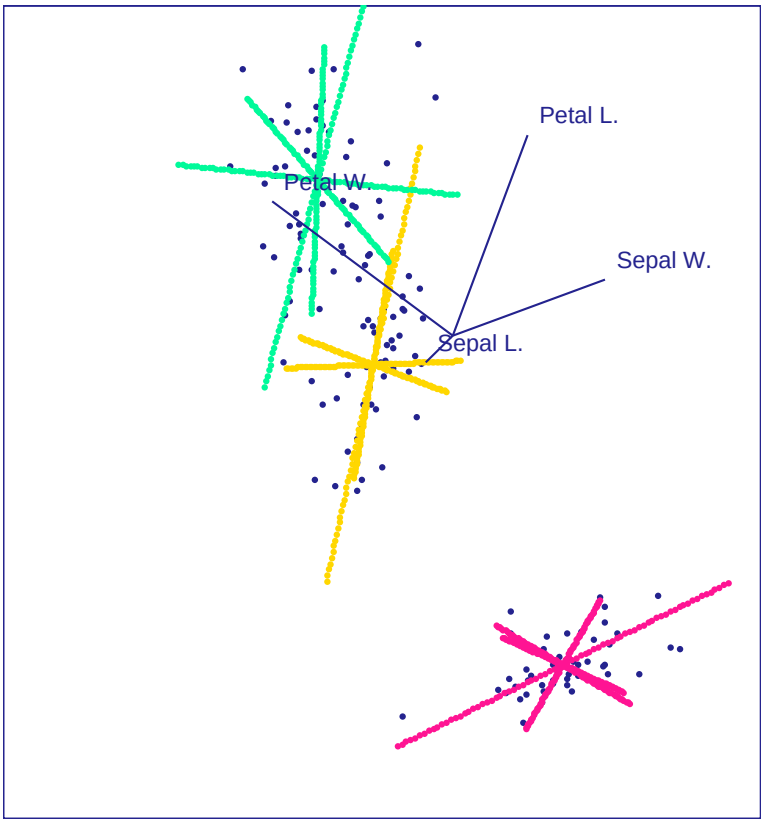

Figure 7: All four eigenvectors for each of the three Iris species in $\Re^4$ (grand tour view in the *ggobi* package).

Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, pp. 297–306.

Terrell, G.R. (1990), "Linear Density Estimates," *Proceedings of the Statistical Computing Section*, American Statistical Association, 297–302.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.

Wojciechowski, W.C. and Scott, D.W. (1999), "Robust Location Estimation with L2 Distance," *Computing Science and Statistics*, K. Berk and M. Pourahmadi, eds, 31, 292–295.

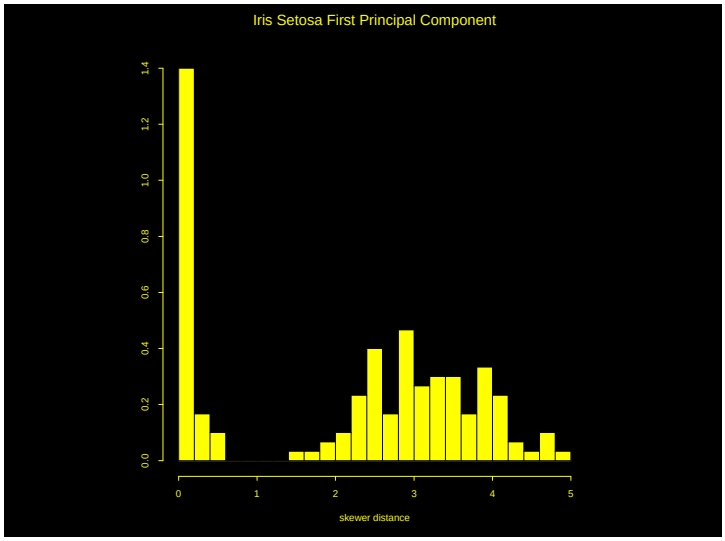

Figure 8: Histogram of distances from the 2-D Iris data to the Iris Setosa Skewer shown in Figure 7.

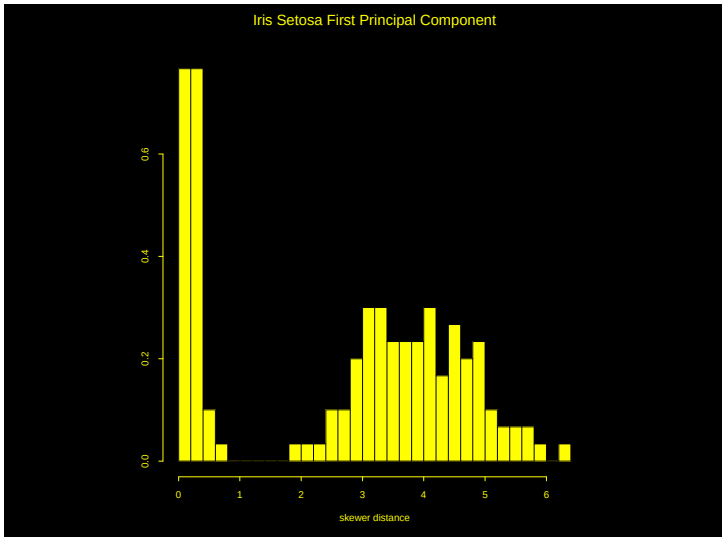

Figure 9: Histogram of distances from the full 4-D Iris data to the 4-D Iris Setosa Skewer.

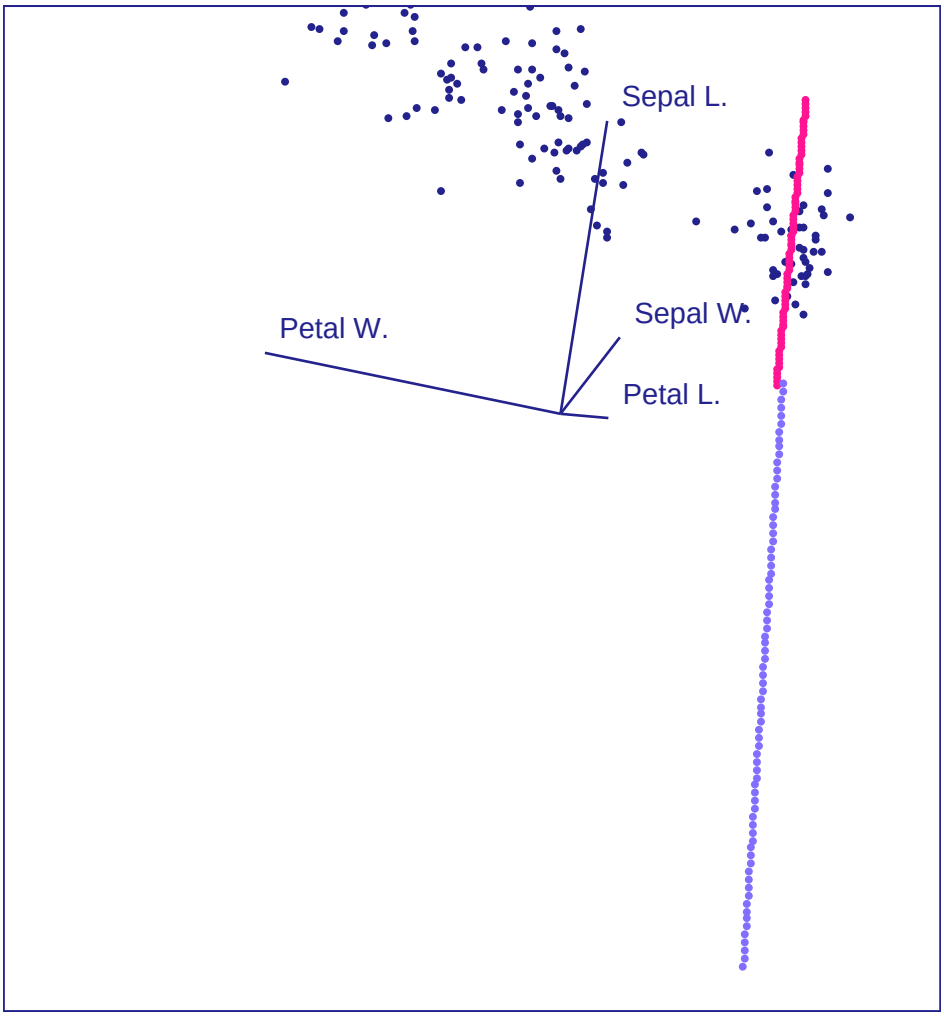

Figure 10: 4-D Skewer of Iris Data. The skewer is blue, while the principal component is red.

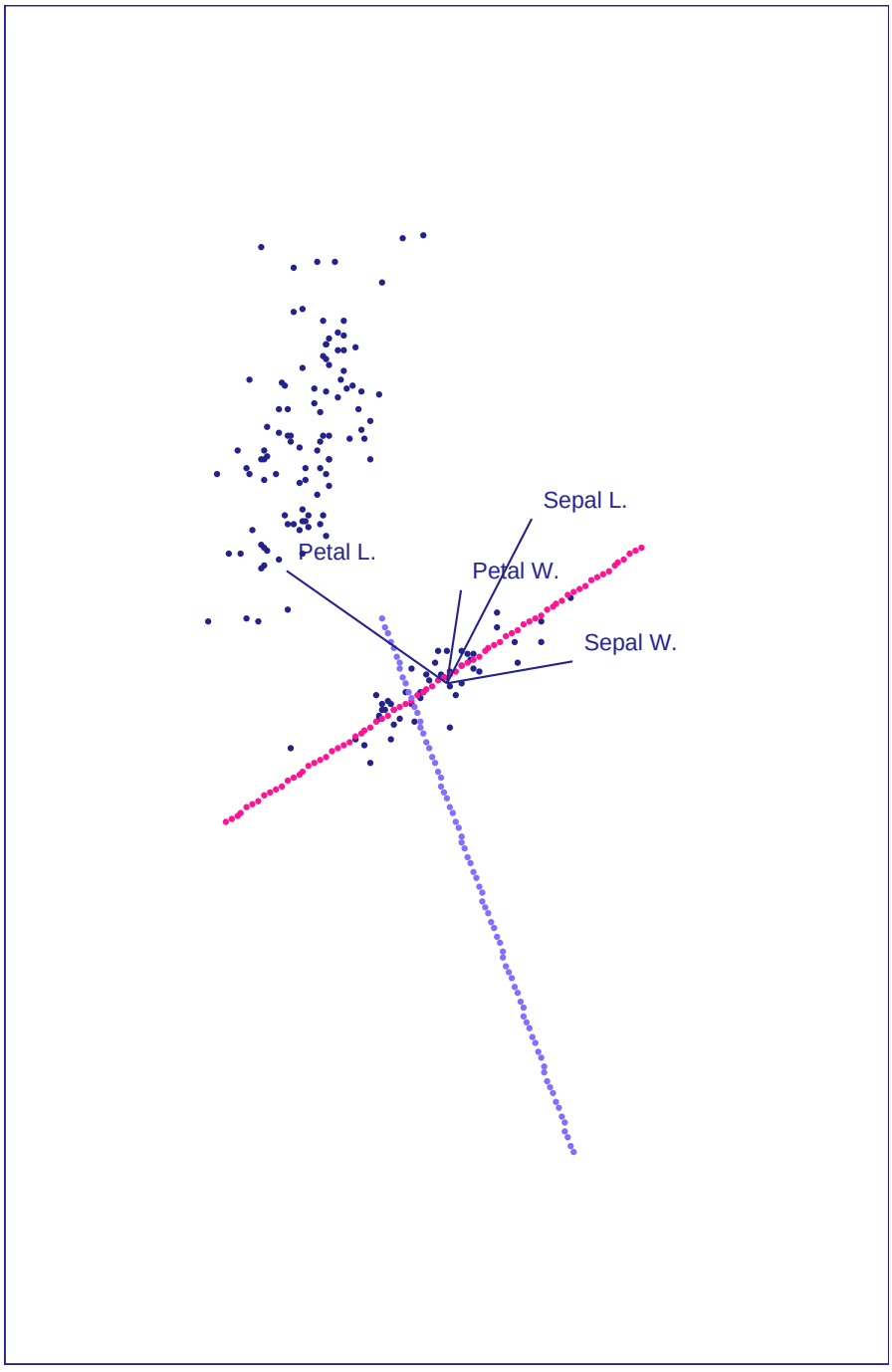

Figure 11: A second 4-D Skewer of Iris Data.