

The Geometry of LDA and PCA Classifiers

Illustrated with 3D Examples.

J. Ross Beveridge

Technical Report 01-101
Computer Science Department
Colorado State University
ross@cs.colostate.edu
Last Update, May 30, 2001

1 Overview

This report will help in developing a geometric interpretation of Fisher Linear Discriminants. The report builds upon an understanding of the connection between Principal Component Analysis and Gaussian Distributions. It contains a running example showing how Fisher Discriminants are computed and what they look like for an illustrative 3 class problem in 3 dimensional space.

Maple 6.0 was used to write this report, and consequently it is both a document and a program. It is available in three forms:

- **Maple 6.0 Worksheet:** This is the best format, since the reader may interact with the 3D plots. It is also simple to construct alternative examples by making minor modifications to the Maple source.
- **HTML:** This is the way most people will view the document. The 3D plots are animated so perceived 3D structure is evident. However, the user cannot alter the view or explore the data as is possible in Maple directly. The link to this version is:
<http://www.cs.colostate.edu/evalfacerec/papers/csuldareport/report01101.html>
- **PDF (from LaTeX):** This is a more standard version with only text, math and static figures. It is best if one wishes to print the document.

When viewing the Maple or HTML versions, the first "Section" is not intended to read. It is, instead, a collection of helper routines written to service the remainder of the document. The reader will also note that mathematics appears in three forms. In line math, input expressions, and the results of evaluating these expression. In line math appears is what one normally expects for type set mathematics. Inputs to expressions are set aside in blocks and are shown in red. Results of these expression appear below the expression in blue.

2 Introduction

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) play a critical role in many pattern classification tasks. It is helpful to develop a geometric intuition for how each transforms the coordinate reference frame in which data is classified. There are limits to how far this can go, since these techniques are typically applied to problems in higher dimensional spaces. However, conceding this weakness, with some small effort many people can usefully extend a natural 3D intuition to higher dimensions.

There are several motivations for this paper, but one in particular concerns whether LDA basis vectors are orthogonal. It has been reported in the literature that the LDA basis vectors are orthogonal: unfortunately this is false. It is easy show that LDA basis vectors are not orthogonal. As this paper will show, one may conceive of an intermediate space in which LDA basis vectors are indeed orthogonal. Moreover, understanding how this intermediate space is subsequently transformed provides insight into how LDA vectors are configured. The general mathematics of these transformations are reviewed here and each step is illustrated with a 3D running example.

We will start with some very basic properties of Gaussian point clouds in 3D and draw heavily upon the connection between these point clouds and PCA. For our purposes, a Gaussian point cloud is formed by sampling a finite number of observations from a multivariate Gaussian random variable in 3D. Our choice of a Gaussian point cloud establishes a well known and tight connection between PCA and Gaussian distributions. However, PCA classifiers are often successfully applied to non-Gaussian data: the choice should not be seen as limiting. The advantage of fixing our attention on Gaussian points clouds is that it allows us to develop an intuition for the geometry underlying both PCA and LDA spaces.

In particular, there is a tight coupling between scatter matrices and covariance matrices. In moving from PCA to LDA, emphasis shifts from the properties of a single scatter matrix to the properties of two related scatter matrices: the between class and within class scatter matrices. Understanding in geometric terms what is happening when one solves for the linear discriminants becomes more difficult. It is well known that the problem of finding the linear discriminants is equivalent to solving a generalized Eigenproblem. However, quoting [Strang], "Geometrically, this has a meaning which we do not understand very well." In this statement, Strang is addressing all generalized Eigenproblems. Fortunately, both PCA and LDA induce positive definite and symmetric scatter matrices. These restrictions in turn lend themselves to a much clearer geometric interpretation.

3 Gaussian Random Variables and Principal Components

3.1 Axis Aligned Gaussians

Consider the equation that defines the probability density function for a Gaussian random variable in 3D.

$$p(x) = \frac{1}{(2\pi)^{(3/2)} \Omega^{(1/2)}} \exp\left(-\frac{1}{2}\right) (x - \mu)^t \Omega^{(-1)} (x - \mu)$$

where

$$x = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \mu = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix}, \Omega = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}$$

In the special case that the cross terms in the covariance matrix are zero, then the pdf reduces to:

$$p(x) = \frac{1}{4} \frac{\sqrt{2} e^{(-1/2 \frac{x^2}{\sigma_{xx}} - 1/2 \frac{y^2}{\sigma_{yy}} - 1/2 \frac{z^2}{\sigma_{zz}})}}{\pi^{(3/2)} \sqrt{\sigma_{xx} \sigma_{yy} \sigma_{zz}}}$$

Collecting terms:

$$p(x) = \frac{1}{(2\pi)^{(1/2)} \sqrt{\sigma_{xx}}} e^{(-\frac{1x^2}{2\sigma_{xx}})} \frac{1}{(2\pi)^{(1/2)} \sqrt{\sigma_{yy}}} e^{(-\frac{1y^2}{2\sigma_{yy}})} \frac{1}{(2\pi)^{(1/2)} \sqrt{\sigma_{zz}}} e^{(-\frac{1z^2}{2\sigma_{zz}})}$$

When the cross terms in the covariance matrix are zero, then the pdf is simply the product of the independent probabilities along each of the three axes. This also means, from a practical standpoint, that samples from an axis-aligned 3D Gaussian random variable may be generated by independently calling a 1D Gaussian random number generator three times, once for each of the three independent components.

3.2 Example of an Axis Aligned Gaussian

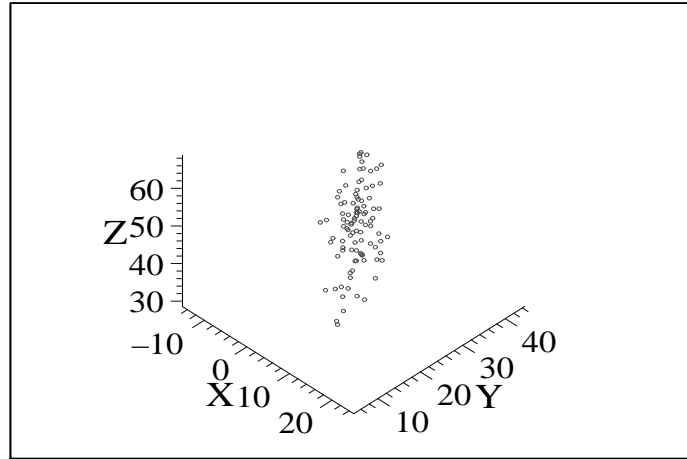
Consider the Gaussian random variable defined by the following mean and standard deviation:

$$\mu = \begin{bmatrix} 5 \\ 25 \\ 50 \end{bmatrix}, \sigma = \begin{bmatrix} 1 \\ 3 \\ 10 \end{bmatrix}$$

Note the sigmas along the diagonal of the covariance matrix have been placed in a column vector.

Here is a plot of 100 points sampled from this distribution:

3.2.1 3D Plot of Points



One sees from these plots that the distribution shown has minimal variation along the x dimension, modest variation along y, and the greatest variation along z.

3.3 The Scatter Matrix, Sample Mean and Sample Covariance

To begin to draw the connection between our sampling from a Gaussian random variable and classification using PCA and LDA, let us begin by defining a data matrix A as the collection of points.

$$A = \begin{bmatrix} p_1 & p_2 & \dots & p_n \end{bmatrix}$$

$$A = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \\ z_1 & z_2 & \dots & z_n \end{bmatrix}$$

The sample mean and sample covariance for this set of points are:

$$\mu_s = \frac{1}{n} \sum_{i=1}^n p_i \quad \Sigma_s = \frac{1}{n-1} S \quad S = \sum_{i=1}^n (p_i - \mu_s)(p_i - \mu_s)^t$$

The matrix S is commonly called the scatter matrix or moment matrix for the set of points. For the specific points in the example above, the sample mean, scatter matrix and covariance matrix are:

$$\mu_s = \begin{bmatrix} 5.10 \\ 25.10 \\ 49.40 \end{bmatrix}, S = \begin{bmatrix} 97.50 & -13.50 & -96.30 \\ -13.50 & 973.50 & 350.80 \\ -96.30 & 350.80 & 9344.70 \end{bmatrix}, \Omega_s = \begin{bmatrix} 1.0 & -.10 & -1.0 \\ -.10 & 9.80 & 3.50 \\ -1.0 & 3.50 & 94.40 \end{bmatrix}$$

The standard deviation along the x, y and z axis are the square roots of the diagonal elements of the sample covariance matrix:

$$\sigma_{xx} = 1.0, \sigma_{yy} = 3.10, \sigma_{zz} = 9.70$$

These values are close to the original. How close depends in part on how many points are sampled.

3.4 Example of a Gaussian Rotated with respect to Principal Axes

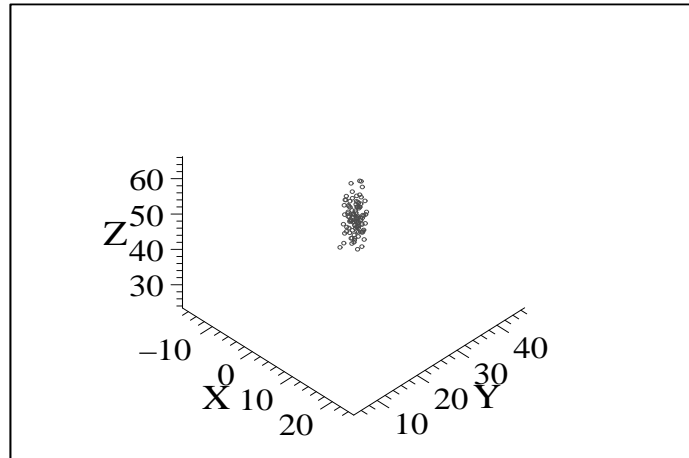
Now let us consider rotating the distribution about the x axis by 45 degrees and about the z axis by 45 degrees.

$$\mu = \begin{bmatrix} 5 \\ 25 \\ 50 \end{bmatrix}, \sigma = \begin{bmatrix} 1 \\ 3 \\ 10 \end{bmatrix}, \text{angles} = \begin{bmatrix} .7854 \\ 0 \\ .7854 \end{bmatrix}$$

Again, the sigmas along the diagonal of the covariance matrix have been placed in a column vector.

Here is a plot of 100 points sampled from this distribution:

3.4.1 3D plot of points.



Now we see the point cloud is moving up in a diagonal direction relative to the x, y and z coordinates. As we will see in the next section, we can analyze the data matrix of this new set of points to recover the principal axes of this distribution.

3.5 Scatter and Covariance Matrices for Rotated Distribution

In the same fashion as above, we can compute the sample mean vector, the scatter matrix, and the associated sample covariance matrix.

$$\mu_s = \begin{bmatrix} 5.00 \\ 24.90 \\ 50.00 \end{bmatrix}, S = \begin{bmatrix} 2868.80 & -2904.90 & 3326.10 \\ -2904.90 & 3154.10 & -3472.90 \\ 3326.10 & -3472.90 & 5502.40 \end{bmatrix}, \Omega_s = \begin{bmatrix} 29.00 & -29.30 & 33.60 \\ -29.30 & 31.90 & -35.10 \\ 33.60 & -35.10 & 55.60 \end{bmatrix}$$

While the sample mean has not been changed, the scatter matrix is very different; the off diagonal elements now have significant value indicating the variance along one axis is correlated with that along another.

3.6 Recovering the Rotation and Original Principal Axes

By construction, the covariance and scatter matrices are symmetric positive definite. Thus, they may be diagonalized in the following manner:

$$\Omega_s = R\Delta R^t$$

Here, R is an orthogonal matrix consisting of unit length row (and column) vectors. Thus, it is essentially a rotation matrix. The only reason we need qualify our statement is that it may include reflection about axes as well as rotation. The matrix Δ is a diagonal matrix:

$$\Delta = \begin{bmatrix} \sigma_{xx}^2 & 0 & 0 \\ 0 & \sigma_{yy}^2 & 0 \\ 0 & 0 & \sigma_{zz}^2 \end{bmatrix}$$

For reasons that will become apparent shortly, let us write Δ as

$$\Delta = SS$$

So now our diagonalization may be written as

$$\Omega_s = RSSR^t$$

The actual R and S matrices for our sample covariance matrix are:

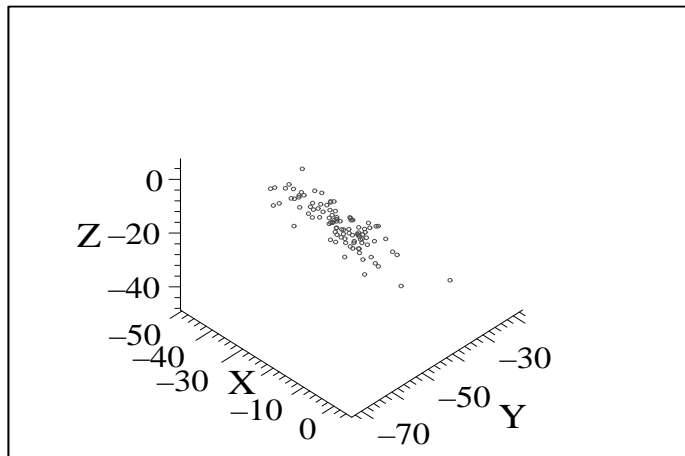
$$R = \begin{bmatrix} -.500 & .470 & -.730 \\ .520 & -.510 & -.680 \\ -.690 & -.720 & 0. \end{bmatrix}, S = \begin{bmatrix} 10.310 & 0. & 0. \\ 0. & 3.010 & 0. \\ 0. & 0. & 1.020 \end{bmatrix}$$

The diagonal terms of the S matrix are in fact the sample standard deviations for the set of points after they have been rotated by R.

$$\sigma_{xx} = 10.30, \sigma_{yy} = 3.00, \sigma_{zz} = 1.00$$

We can see this visually if we actually rotate and plot the points in the new coordinate system

3.6.1 3D Plot of Points after Rotation to Principal Components



As further confirmation, here is the sample mean, the scatter matrix, and the sample covariance matrix for the new set of points Q .

$$\mu_s = \begin{bmatrix} -27.30 \\ -44.30 \\ -20.90 \end{bmatrix}, S = \begin{bmatrix} 10480.30 & -518.70 & -406.70 \\ -518.70 & 924.90 & 52.00 \\ -406.70 & 52.00 & 120.10 \end{bmatrix}, \Omega_s = \begin{bmatrix} 105.90 & -5.20 & -4.10 \\ -5.20 & 9.30 & .50 \\ -4.10 & .50 & 1.20 \end{bmatrix}$$

$$\sigma_{xx} = 10.30, \sigma_{yy} = 3.10, \sigma_{zz} = 1.10$$

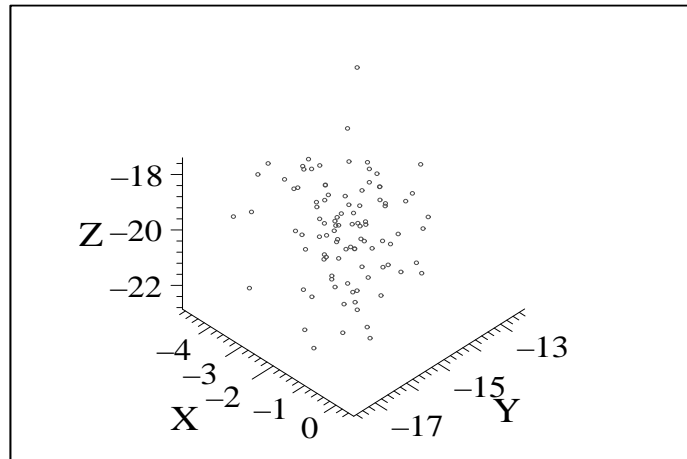
The matrix R defines the principal components of the point cloud generated by our rotated Gaussian distribution. The row vectors in R are also the Eigenvectors of the covariance or scatter matrix. The covariance and scatter matrices may be used interchangeably to find R since these matrices only differ by a constant scale factor.

3.7 Interpreting Variance as Scale

The matrix S may be described as a diagonal matrix containing the sample estimates for the standard deviation of the points along each of the principal components. Another way to view this is to observe that scaling the points by the inverse of this matrix will generate a new point set with unit variance along each of the principals axes. To illustrate, let us apply this transformation to the data matrix P containing our sample points. The result will be a new set of points Q .

$$Q = S^{(-1)}RP$$

3.7.1 3D Plot of Points after Rotation and Scaling



Looking at the sample mean, scatter matrix and sample covariance for the resulting set of points Q we will see the sample standard deviations are now all close to one.

$$\mu_s = \begin{bmatrix} -2.70 \\ -14.70 \\ -20.50 \end{bmatrix}, S = \begin{bmatrix} 98.60 & -16.70 & -38.70 \\ -16.70 & 102.00 & 16.90 \\ -38.70 & 16.90 & 115.50 \end{bmatrix}, \Omega_s = \begin{bmatrix} 1.0 & -.20 & -.40 \\ -.20 & 1.00 & .20 \\ -.40 & .20 & 1.20 \end{bmatrix}$$

$$\sigma_{xx} = 1.0, \sigma_{yy} = 1.00, \sigma_{zz} = 1.10$$

In essence, what we have done with the rotation and scale derived from the original covariance matrix is create a transformation to a new space in which the points have unit variance in all directions.

3.8 Principal Components Subspace Maximizes Variance

Often we do not bother to make explicit the criteria that the principal components optimize. However, because it will play a role later in how the Fisher Linear Discriminants are defined, it is worthwhile to review this basic material. Commonly it is stated that principal components represent axes of maximal variance. To put this more precisely, if one sought a single dimension over which the variance of the data in a data matrix was maximized, it would be the axis that maximized the following function V .

$$V(W) = W\Omega W^t$$

For a data matrix of points in 3D space, this product becomes.

$$V(W) = [w_x, w_y, w_z] \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \\ w_z \end{bmatrix}$$

Looking at the general case where Ω is not diagonal, it is not obvious what unit length basis vector to choose for W . However, for the case where Ω is diagonal, the choice is obvious given.

$$V(W) = [w_x \quad w_y \quad w_z] \cdot \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{bmatrix} \cdot \begin{bmatrix} w_x \\ w_y \\ w_z \end{bmatrix}$$

Assume that $\sigma_{xx} > \sigma_{yy} > \sigma_{zz}$, then the W that maximizes $V(W)$ is $[1, 0, 0]$. Now recall our earlier diagonalization of the general case.

$$\Omega = RSSR^t$$

We observed that R is a rotation matrix that shifts us from the original coordinates system, where Ω is not diagonal, to the principal components space, where the covariance matrix is diagonal. So, the axis in our original space that maximizes variance is the backward mapping of $[1, 0, 0]$ into the original space. Putting this more simply,

$$W = R^t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

and even more simply, we see that the first column of R is the unit length basis vector that defines the axis of maximum variance. This axis is what we typically call the first principal component of the data matrix.

The extension to k axes that collectively maximize variance generalizes the criterion function to

$$V(W) = |W\Omega W^t|$$

Now we are maximizing the determinant of the $k \times k$ matrix formed from product of the $k \times n$ matrix W and the covariance matrix Ω . By a generalization of the argument made above, the basis vectors that maximize this determinant are the first k columns of R .

4 Fisher Discriminants

Fisher's Linear Discriminants are defined as the k basis vectors in a $k \times n$ matrix W that maximizes the following function

$$J(W) = \frac{|W M_B W^t|}{|W M_W W^t|}$$

Where M_B is the between class scatter matrix and M_W is the within class scatter matrix [Duda]. More particularly, the within class scatter matrix is

$$M_W = \sum_{i=1}^c M_i \quad \text{where} \quad M_i = \sum_{j=1}^{n_i} (p_j - (\mu_i) \cdot ((p_j - \mu_i)^t))$$

and n_i is the number of elements in class i , p_j is the j th point in class i , and μ_i is the mean vector for class i .

The between class scatter matrix is

$$M_B = \sum_{i=1}^c ((n_i) (\mu_i - \mu) ((\mu_i - \mu)^t))$$

Now, quoting from [Duda], "The problem of finding a rectangular matrix W that minimizes $J(\cdot)$ is tricky." They, and other common references, state without elaboration that the optimal W may be found by solving the generalized eigenvector problem

$$M_B w_i = \lambda_i M_W w_i$$

Specifically, the optimal W consists of the eigenvectors associated with the k largest eigenvalues.

The problem with stopping at this point is two fold. First, many of us have no geometric intuition for what it means to solve this general eigenvector problem. Second, while general, this approach may not always be the most efficient/robust means of finding the Fisher Discriminants. This latter observation has been made by [Zhao] and others. They provide a general means of transforming a generalized eigenvector problem to a symmetric eigenvector problem, and this transformation is the basis for developing a geometric intuition for what is actually taking place when computing the Fisher Discriminants.

Here is the transformation with many of the intermediate steps omitted from [Zhao] made explicit.

4.1 Transformation to a Standard Eigenvector Problem

Begin with the generalized eigen vector problem:

$$M_b W = M_w W \Lambda$$

Now use singular value decomposition to express M_w in diagonal form, and indeed, go one step further and explicitly identify the square root of the diagonal as a scale matrix. So

$$M_w = R_w S_w S_w R_w^T$$

The original problem may now be written as

$$M_b W = R_w S_w S_w R_w^T W \Lambda$$

Take the inverse scale and rotation and right multiply both sides

$$S_w^i R_w^T M_b W = S_w R_w^T W \Lambda$$

Define a new matrix that will become the eigenvectors of a symmetric eigenvector problem

$$V = S_w R_w^T W$$

and substitute V on the right hand side

$$S_w^i R_w^T M_b W = V \Lambda$$

To simplify the left hand side, introduce a single matrix G that combines the scale and rotation derived from the within class scatter matrix.

$$G = S_w^i R_w^T \quad \text{and} \quad G^T = R_w S_w^i$$

Expand the left hand side by left multiplying by a sequence of scales and rotations that together are the identity transformation.

$$S_w^i R_w^T M_b R_w S_w^i S_w R_w^T W = V \Lambda$$

Now substitute G and V in where appropriate

$$G M_b G^T V = V \Lambda$$

The final result is a symmetric eigenvector problem in N_b where $N_b = G M_b G^T$:

$$N_b V = V \Lambda$$

Once we solve for V , W is found directly by the equality

$$W = R_w S_w^i V$$

We will see below that G is a transformation that takes us into a space where the fisher discriminants are related directly to the eigenvectors of the scatter matrix associated with the transformed points.

4.2 How this Transformation Acts Upon Fisher's Criterion

Now we can draw the connection between the general algebraic manipulation above and the Fisher Criterion we wish to optimize.

Recall Fisher's criterion

$$J(W) = \frac{|W M_B W^t|}{|W M_W W^t|}$$

Consider how much easier this problem would be to solve if the denominator were a constant with respect to our choice of W . We can map to a new space where this is true by exploiting the rotation and scale transformations obtained from diagonalizing M_W . Thus, reiterating the diagonalization used in the previous section.

$$M_w = R_w S_w S_w R_w^T$$

Applying the transformation G to our points before creating the within and between class scatter matrices leads to a new problem where only the numerator varies as a function of W . Showing the algebra of this transformation,

$$J(V) = \frac{\left| V R^T S_w^{(-1)} M_B R S_w^{(-1)} V^T \right|}{\left| V R^T S_w^{(-1)} R_w S_w S_w R_w^T R S_w^{(-1)} V^T \right|}$$

The denominator of this new criterion function collapses to $V V^T$, and this further collapses and the entire denominator becomes 1.

In geometric terms, we have transformed our problem to a space where the covariance of the within class scatter is one in all principal directions. This is exactly the same as we illustrated above when we showed that the R and S derived from the diagonalization of a covariance matrix could be used to remap a data matrix into a space where the points formed a compact ball with variance 1 in all directions.

One loose end is why maximizing J in the transformed space is equivalent to maximizing it in the original space. The answer lies in observing that the rotations leave the determinant unchanged, and the two scale matrices alter both the numerator and denominator by the same constant factor. Thus, maximizing one ratio is equivalent to maximizing the other. We are now ready to illustrate this method of finding the Fisher Discriminants on a specific 3D example.

5 Example of a Three Class Problem

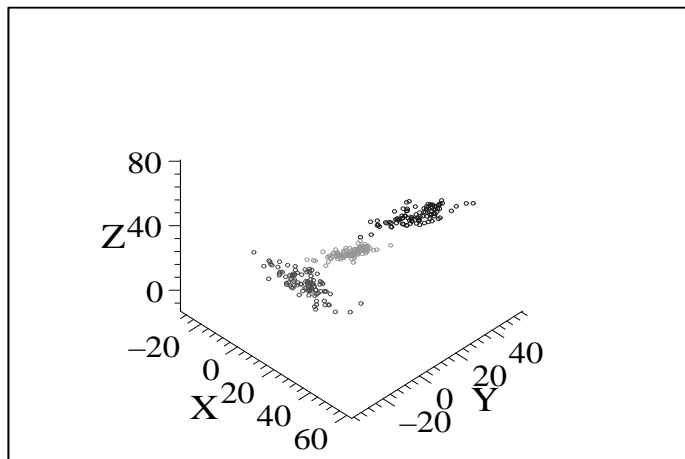
Consider a three class recognition problem where samples from each class are drawn from distinct distributions.

5.1 Different Covariance Structures

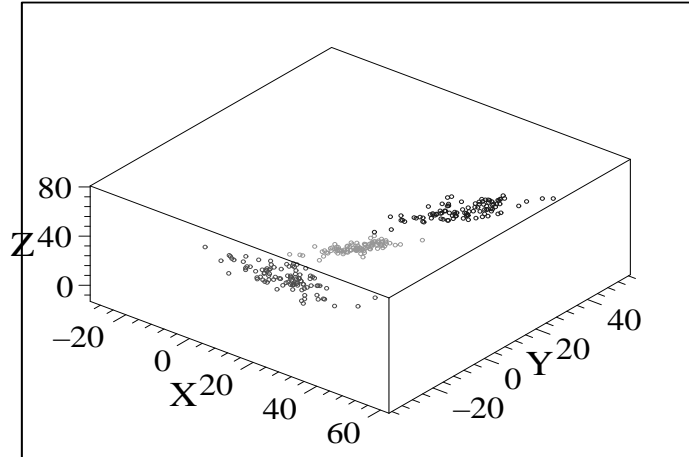
Here are three distinct tri-variate Gaussian random variables and sample points from each. These three classes give rise to two Fisher Linear Discriminants in 3D.

$$\left[\begin{array}{c} \mu = \begin{bmatrix} 0 & 20 & 50 \\ 0 & 5 & 10 \\ 0 & 30 & 70 \end{bmatrix}, \sigma = \begin{bmatrix} 10 & 10 & 10 \\ 3 & 3 & 3 \\ 1 & 1 & 1 \end{bmatrix}, \text{angles} = \begin{bmatrix} .4 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & 0 & 1.2 \end{bmatrix} \right]$$

5.1.1 3D Plot of Points from Three Classes



5.1.2 3D Animated Plot of Points from Three Classes



The sample means and scatter matrices for the three classes are:

$$\mu_1 = \begin{bmatrix} -.60 \\ -.20 \\ 0. \end{bmatrix}, \mu_2 = \begin{bmatrix} 20.30 \\ 4.20 \\ 30.30 \end{bmatrix}, \mu_3 = \begin{bmatrix} 50.40 \\ 11.00 \\ 70.00 \end{bmatrix}$$

$$\Omega_1^2 = \begin{bmatrix} 98.30 & -1.40 & .50 \\ -1.40 & 7.50 & 2.90 \\ .50 & 2.90 & 2.10 \end{bmatrix}, \Omega_2^2 = \begin{bmatrix} 43.00 & 2.20 & 35.70 \\ 2.20 & 11.20 & 1.30 \\ 35.70 & 1.30 & 31.20 \end{bmatrix}, \Omega_3^2 = \begin{bmatrix} 17.90 & 24.50 & -.20 \\ 24.50 & 72.70 & 0. \\ -.20 & 0. & 1.20 \end{bmatrix}$$

6 Illustrating the Fisher Linear Discriminants

The within and between class scatter matrices are shown below. Next, the rotation and scale associated with the within and between class scatter matrices are shown. The rotation and scale for the within class scatter matrix is used to compute the transformation G described above that will cause the variance within classes to be unity in all directions. The Fisher Basis Vectors (Discriminants) are then computed in this space by solving a standard eigenvector problem, and subsequently transformed into the original data space. The resulting discriminants computed in this manner are compared against discriminants computed using the more commonly prescribed manner of solving a generalized eigenvector problem: as expected, they are the same. Finally, the 3D points are shown projected into the 2D space defined by the 2 Fisher discriminants, thus illustrating how these discriminants concentrate samples within classes while at the same time separating the classes.

6.1 The Within and Between Class Scatter Matrices

The within class scatter matrix is the sum of the scatter matrices for the three classes. The between class scatter matrix is formed from three points that are the centroids (means) of the three classes.

$$S_W = \begin{bmatrix} 15919.60 & 2524.10 & 3601.00 \\ 2524.10 & 9142.30 & 413.30 \\ 3601.00 & 413.30 & 3453.40 \end{bmatrix}, S_B = \begin{bmatrix} 131618.70 & 29098.50 & 180227.20 \\ 29098.50 & 6435.20 & 39826.20 \\ 180227.20 & 39826.20 & 246952.40 \end{bmatrix}$$

6.1.1 Computer R and S for the within class and between class matrices

Just for interest, compute the diagonalization of the within class and between class scatter matrices

$$R_{S_W} = \begin{bmatrix} -.930 & .270 & -.270 \\ -.290 & -.960 & .040 \\ -.240 & .110 & .960 \end{bmatrix}, S_{S_W} = \begin{bmatrix} 132.840 & 0. & 0. \\ 0. & 91.600 & 0. \\ 0. & 0. & 49.770 \end{bmatrix}$$

$$R_{S_B} = \begin{bmatrix} -.590 & -.730 & -.350 \\ -.130 & -.340 & .930 \\ -.800 & .590 & .110 \end{bmatrix}, S_{S_B} = \begin{bmatrix} 620.440 & 0. & 0. \\ 0. & 8.070 & 0. \\ 0. & 0. & 0. \end{bmatrix}$$

6.2 Change Coordinates by the G Transformation

As derived above, G is the composition of inverse rotation followed by the inverse scale derived from the within class scatter matrix.

$$G = \begin{bmatrix} -.00700 & -.00220 & -.00180 \\ .00290 & -.01050 & .00130 \\ -.00530 & .00080 & .01940 \end{bmatrix}$$

6.2.1 Look at within and between scatter matrices after transformation

Here we confirm with an example that indeed the within class scatter matrix after this change is now the identity matrix. Thus, searching for the basis vectors that maximize Fisher's criterion in this new space amounts to maximizing the numerator for the new between class scatter matrix.

$$S_W = \begin{bmatrix} 1.000000 & 0. & 0. \\ 0. & 1.000000 & 0. \\ 0. & 0. & 1.000000 \end{bmatrix}, S_B = \begin{bmatrix} 13.060530 & -3.019950 & -27.984910 \\ -3.019950 & .699000 & 6.476270 \\ -27.984910 & 6.476270 & 60.005060 \end{bmatrix}$$

The next step is to diagonalize the between class matrix to obtain the principal components of the between class scatter matrix. The first two of these are the Fisher Linear Discriminants expressed in this transformed space.

$$V = \begin{bmatrix} -.420 & -.910 & .050 \\ .10 & -.10 & -.990 \\ .900 & -.410 & .130 \end{bmatrix}, \Lambda = \begin{bmatrix} 8.590 & 0. & 0. \\ 0. & .090 & 0. \\ 0. & 0. & 0. \end{bmatrix}, W = \begin{bmatrix} 0. & 0. & 0. \\ 0. & 0. & .010 \\ .020 & 0. & 0. \end{bmatrix}$$

6.3 The Fisher Basis Vectors in Transformed and Original Space

The Fisher Basis Vectors in the transformed space are FB_v and in the original space they are FB_w .

$$FB_v = \begin{bmatrix} -.4210 & -.9060 \\ .0970 & -.0960 \\ .9020 & -.4120 \end{bmatrix}, FB_w = \begin{bmatrix} -.0870 & .7640 \\ .0340 & .2430 \\ .9960 & -.5970 \end{bmatrix}$$

The angle between the vectors in the two spaces are

$$\theta_v = 90.000$$

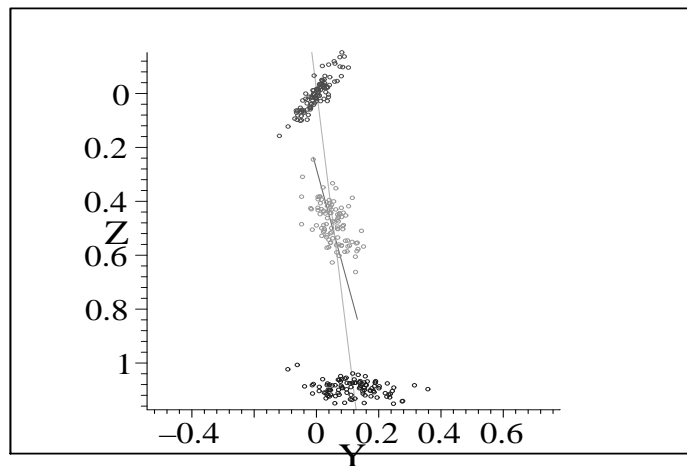
$$\theta_w = 49.280$$

As expected, these vectors are orthogonal in the intermediate space, but not in the final space.

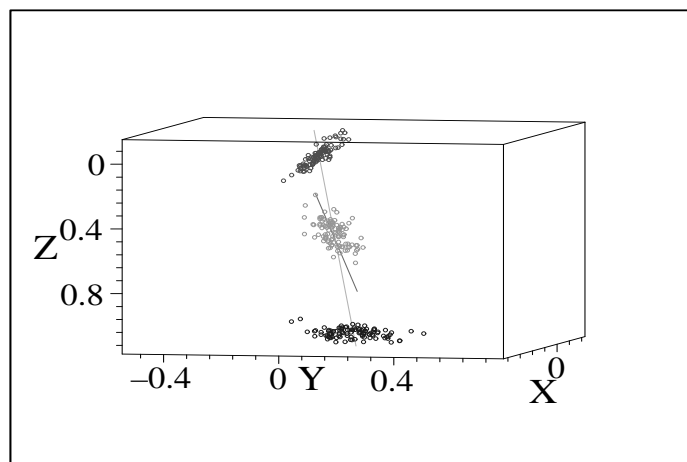
6.4 Plot Data Matrix with Fisher Basis Vectors

Generate 3D scatter plots with the 2 Fisher Basis Vectors drawn in as line segments.

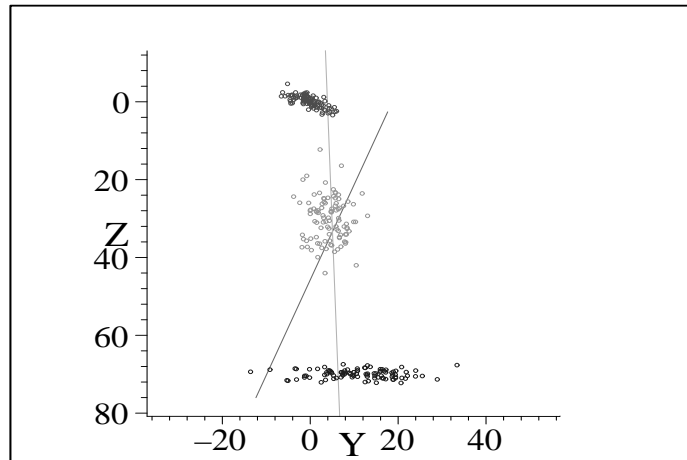
6.4.1 Fisher Basis Vectors in Transformed Space



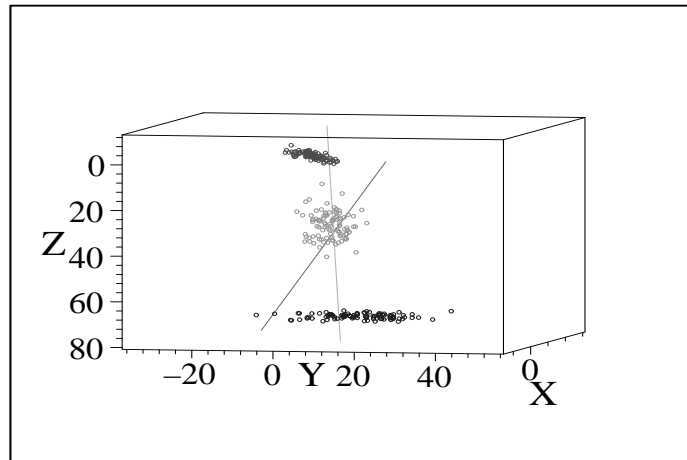
6.4.2 Animation of Fisher Basis Vectors in Transformed Space



6.4.3 Fisher Basis Vectors in Original Space



6.4.4 Animation of Fisher Basis Vectors in Original Space



6.4.5 What is the fisherCriterion for the resulting W

$$J(W) = .55250$$

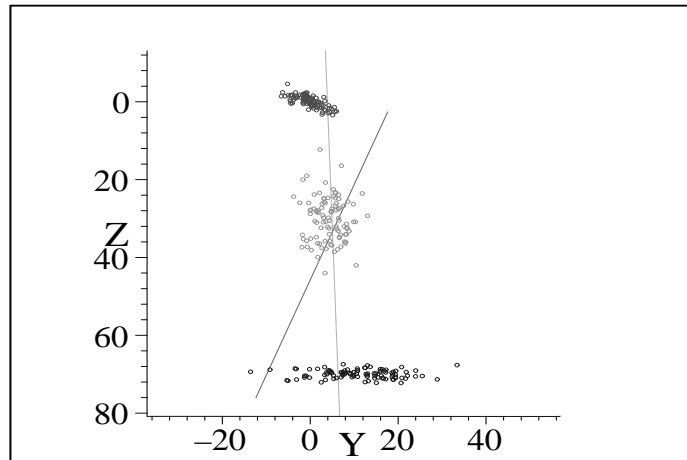
6.5 Compare to Generalized Eigenvector Method

As discussed above, the standard way recommended in [Duda] for finding the Fisher Basis Vectors is to solve a generalized Eigenvector problem. Here we check to see if the results we obtain through this method are comparable.

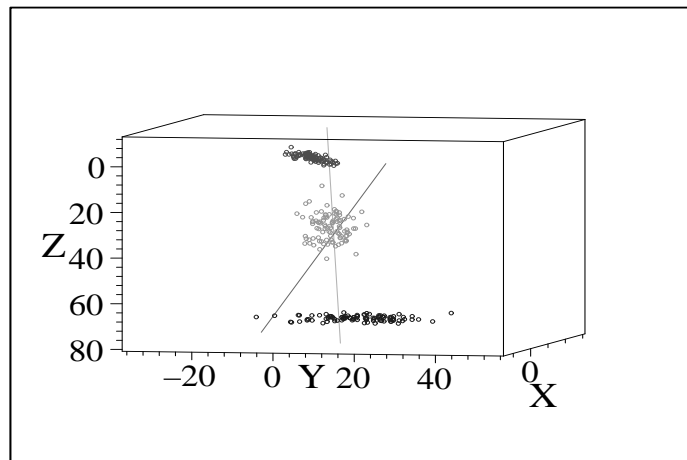
$$S_W = \begin{bmatrix} 15919.60 & 2524.10 & 3601.00 \\ 2524.10 & 9142.30 & 413.30 \\ 3601.00 & 413.30 & 3453.40 \end{bmatrix}, S_B = \begin{bmatrix} 131618.70 & 29098.50 & 180227.20 \\ 29098.50 & 6435.20 & 39826.20 \\ 180227.20 & 39826.20 & 246952.40 \end{bmatrix}$$

$$G_R = \begin{bmatrix} .0870 & .7640 & -.3500 \\ -.0340 & .2430 & .9310 \\ -.9960 & -.5970 & .1050 \end{bmatrix}, G_S = \begin{bmatrix} 8.5880 & 0. & 0. \\ 0. & .0870 & 0. \\ 0. & 0. & 0. \end{bmatrix}, W = \begin{bmatrix} -.0020 & .0080 & -.0040 \\ 0. & .0030 & .0100 \\ .0180 & -.0060 & .0010 \end{bmatrix}$$

6.5.1 3D Plot of Fisher Basis Vectors Found using Generalized Eigenvector Method



6.5.2 3D Animation of Fisher Basis Vectors Found using Generalized Eigenvector Method



Of interest is the angle between the Fisher Basis Vectors computed using the generalized eigenvector method θ_w . This should be the same as the angle between the vectors as found using the geometric transformation method. Also of interest is the value of the Fisher Criterion. These should be the same for the generalized eigenvector and geometric transformation methods, denoted as $J(W_b)$ and $J(W_a)$ respectively.

$$\theta_w = 49.280$$

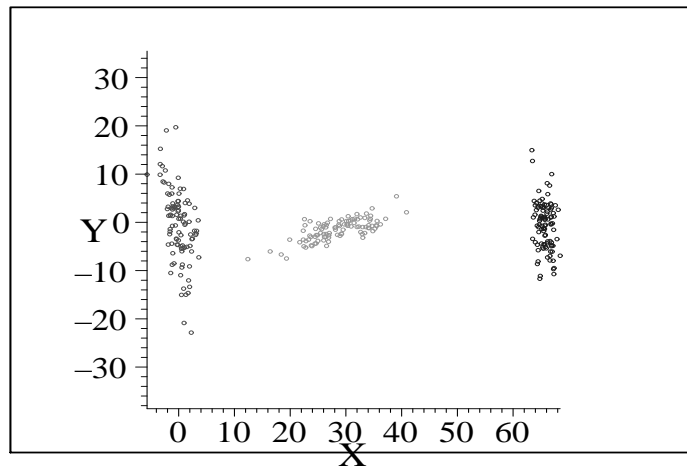
$$J(W_b) = .55252863079274794867566029410447$$

$$J(W_a) = .55252863079289922603574657131343$$

6.6 View 2D projection using Alternative Bases

Another way to look at the result of finding Fisher Discriminants is to look at the projected points. As expected, there is good separation between the resulting 2D points.

6.6.1 2D Plot of classes projected onto Fisher Discriminants



7 References

Duda "Pattern Classification", Richard O. Duda, Peter E. Hart and David G. Stork, Wiley-Interscience, 2001.

Strang "Introduction to Linear Algebra", Gilbert Strang, Wellesley-Cambridge Press, 1998.

Zhao "Subspace Linear Discriminant Analysis for Face Recognition", Zhao, Chellapa and Phillips, Center for Automation Research, University of Maryland, College Park, Technical Report CAR-TR-914, 1999.