The Mathematical Process of Classification

Brandon M. Turner

Missouri State University

The paper will focus on one of the most computationally intensive types of problems in industrial-organizational psychology: classification of individuals among groups. The paper will look at the bivariate case as well as the *p*-variate case. The paper also covers the probability distribution function for particular scores and ranges of scores. The last section will explore the probabilities behind classification as well as the probabilities of misclassification.

The paper will focus on one of the most computationally intensive types of problems in industrial-organizational psychology: classification of individuals among groups. We will look at the *p*-variate case only after I have convinced you of the effectiveness of the bivariate case. The paper will follow the same basic process that any psychologist might use to increase the success of hiring individuals for any particular type of work. I will start with the predicting equations and then merge into the probability distributions. Once a particular distribution equation is effectively constructed, we can then use this information to establish confidence regions. I will then progress into decision procedures for assigning individuals into groups. This section will explore the probabilities behind classification as well as the probabilities of misclassification.

Often, industrial-organizational (I/O) psychologists are hired to increase the overall efficacy of the hiring procedures for certain corporations. To accomplish this task, the psychologist would first be interested in assigning certain positions a combination of attributes that would hopefully optimize this position's productivity or efficiency. The psychologist will need to consult with the subject-matter experts (SME) to determine which attributes are most important to the company for all of the positions in question. The psychologist needs to prepare an assessment with questions that specifically target those attributes in question. From this, the psychologist can use the scores on the assessment along with corresponding information about that particular person's overall effectiveness or yield. The yield can be hypothetical, but the yield is usually a concrete number like an employee's score on annual evaluations. This number may get more complicated. For instance, if the employer wanted to hire people that would produce a high yield but was willing to sacrifice a high yield for a longer career, then the psychologist would put a 'weight' in the equation so that the yield was higher for individuals that stayed with the company longer (assuming that everything else is equal). I have found that the concepts are better understood with computational examples.

Suppose the psychologist has records from a previous similar study in which the data is distributed below. I will call the average of each employee's evaluation the yield or *y*-component. The *x*-component will be the scores that the employee achieved on the developed assessment. For graphical purposes, the example will consist of only two attributes or scales. Let the first attribute be verbal ability and the second attribute be mathematical ability.

$$\mathbf{X} = \begin{bmatrix} 1 & 100 & 60 \\ 1 & 91 & 69 \\ 1 & 32 & 83 \\ 1 & 50 & 86 \\ 1 & 62 & 58 \\ 1 & 26 & 44 \\ 1 & 93 & 89 \\ 1 & 86 & 82 \\ 1 & 15 & 20 \\ 1 & 92 & 99 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 67 \\ 73 \\ 70 \\ 74 \\ 59 \\ 32 \\ 100 \\ 97 \\ 12 \\ 100 \end{bmatrix}$$

As mentioned, the psychologist will use the data to predict how well an individual will perform (yield) given his or her score on the developed assessment. To do this, I will fit the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

by using the equation for estimation

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

The $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ matrix is

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \left(\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 100 & 91 & \ldots & 92 \\ 60 & 69 & \ldots & 99 \end{bmatrix}\begin{bmatrix} 1 & 100 & 60 \\ 1 & 91 & 69 \\ \ldots & \ldots & \ldots \\ 1 & 92 & 99 \end{bmatrix}\right)^{-1} = \begin{bmatrix} 10 & 647 & 690 \\ 647 & 51059 & 48712 \\ 690 & 48712 & 52792 \end{bmatrix}^{-1} = \begin{bmatrix} 1.0372 & -.0018 & -.0119 \\ -.0018 & .0002 & -.0001 \\ -.0119 & -.0001 & .0003 \end{bmatrix}$$

and the $\mathbf{X}^T\mathbf{y}$ matrix is

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 100 & 91 & \ldots & 92 \\ 60 & 69 & \ldots & 99 \end{bmatrix}\begin{bmatrix} 67 \\ 73 \\ \ldots \\ 100 \end{bmatrix} = \begin{bmatrix} 684 \\ 50795 \\ 53055 \end{bmatrix}.$$

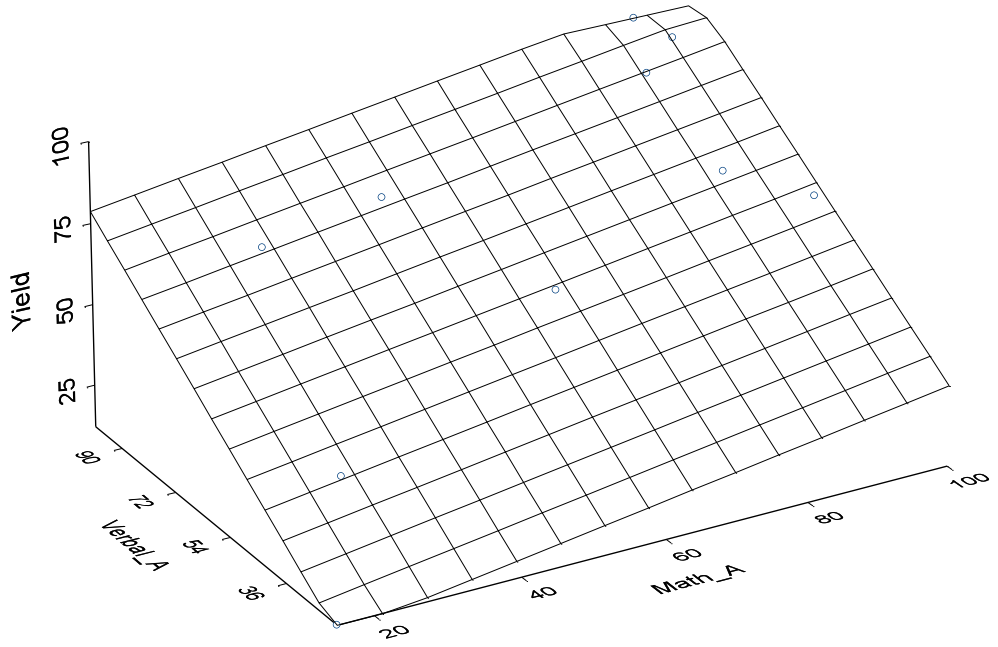
Since the linear regression model will never be perfect, we must use the least squares estimate of the vector β with the above equation for estimation:

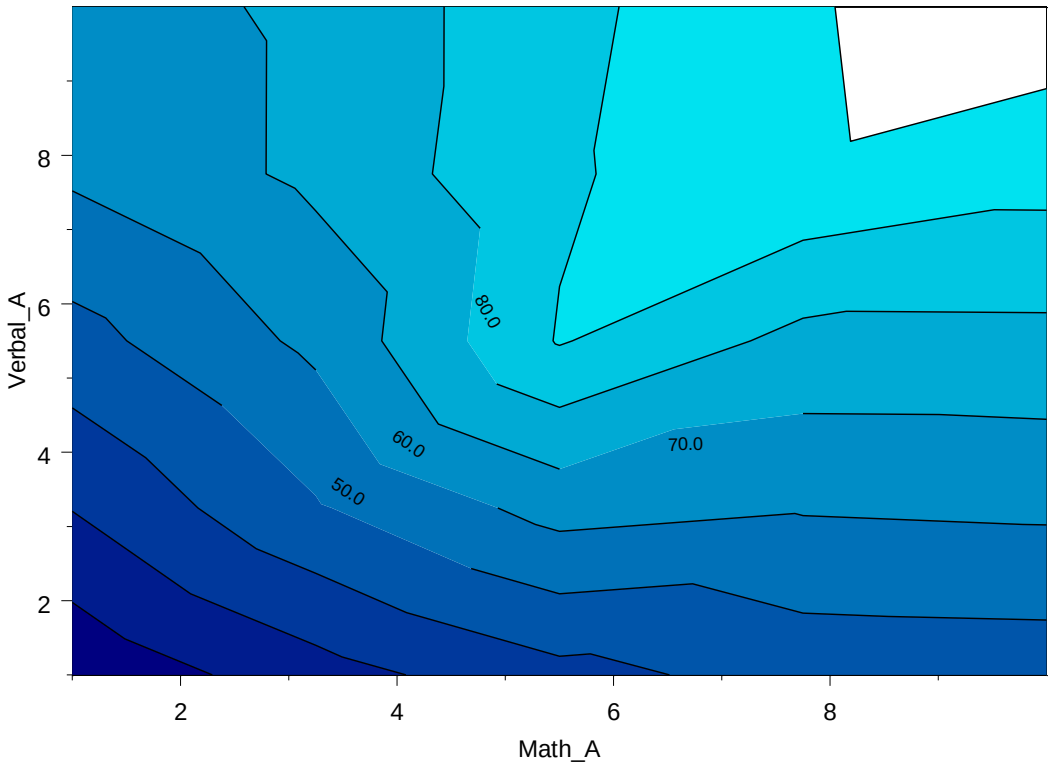
$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1.0372 & -.0018 & -.0119 \\ -.0018 & -.0018 & -.0001 \\ -.0119 & -.0001 & .0003 \end{bmatrix}\begin{bmatrix} 684 \\ 50795 \\ 53055 \end{bmatrix} = \begin{bmatrix} -13.0135 \\ .3231 \\ .8769 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Thus, the least squares estimate for this example is

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = -13.0135 + .3231 x_1 + .8769 x_2.$$

The plot of the regression model (also called the surface response model) is shown below with the corresponding contour plot.


Yield
100
75
50
25
90
72
54
36
Verbal_A
20
40
60
80
100
Math_A


Verbal_A
8
6
4
2
80.0
70.0
60.0
50.0
2
4
6
8
Math_A

We will take a quick look at the second-order regression models only for the illustration of finding the maximum value and the corresponding score vector, **x**. In many optimization problems, the regression model needs to take into account the correlation between variables. These types of models incorporate curvature and generally fit the regression models better than the first-order models. The equation is given by

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^{\,2} + \sum\sum_{i<j} \beta_{ij} x_i x_j + \varepsilon .$$

We might also be interested in the point of maximum yield. If we first write the above equation into matrix notation, we will have

$$y = \beta_0 + \mathbf{x}^T\mathbf{b} + \mathbf{x}^T\mathbf{B}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}; \qquad \mathbf{b} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}; \qquad \mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12}/2 & \cdots & \beta_{1k}/2 \\ \beta_{21}/2 & \beta_{22} & \cdots & \beta_{2k}/2 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1}/2 & \beta_{k2}/2 & \cdots & \beta_{kk} \end{bmatrix}.$$

The derivative of *y* with respect to the elements of the vector **x** when set to **0** is

$$\frac{\partial y}{\partial \mathbf{x}} = \mathbf{b} + 2\mathbf{B}\mathbf{x} = 0 \Rightarrow \mathbf{x} = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} .$$

Finally, we can find the predicted response or maximum value (yield) at the stationary point by the equation

$$y = \beta_0 + \frac{1}{2}\mathbf{x}^T\mathbf{b} .$$

We can use the contour plot to discover the region in which the *p*-percentile of the scores is. Having the regression estimate makes the job very nice because after the scores are reported, we can immediately approximate the overall yield that applicant might have. For example, suppose an applicant scores a 76 on verbal ability and an 81 on mathematical reasoning. Given this, we can estimate that individual's potential productivity in the company as being -13.0135 + .3231(76) + .8769(81) = 82.571. Because the yield ranges from 100 to 0, we can interpret the applicant's score or predicted yield as being 82.571. Reiterating, if hired, this applicant is predicted to have a yield equal to 82.571 units of yield. Depending of course on the company's hiring status, they might want to consider scheduling an interview for this applicant.

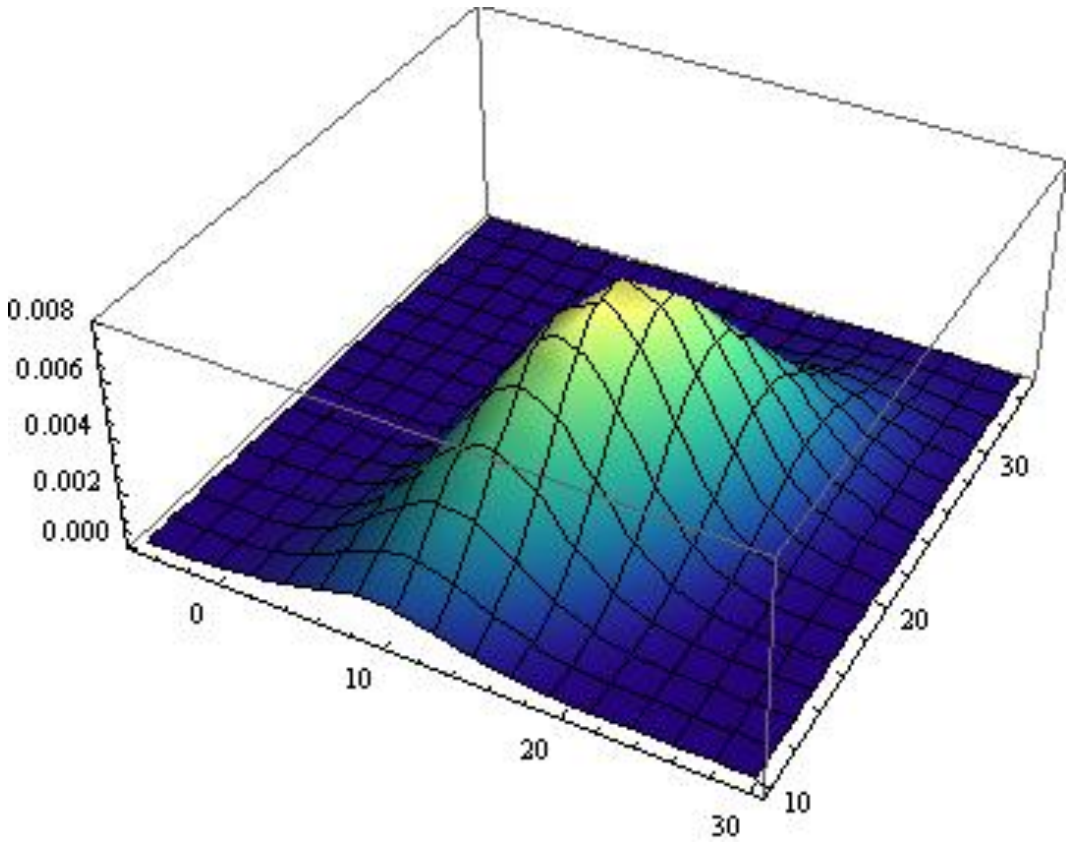
Once we have enough data that we feel sufficiently samples the population, we can determine the probability distribution of the assessment. For both complexity and notation purposes, I will start with the bivariate normal density function given by

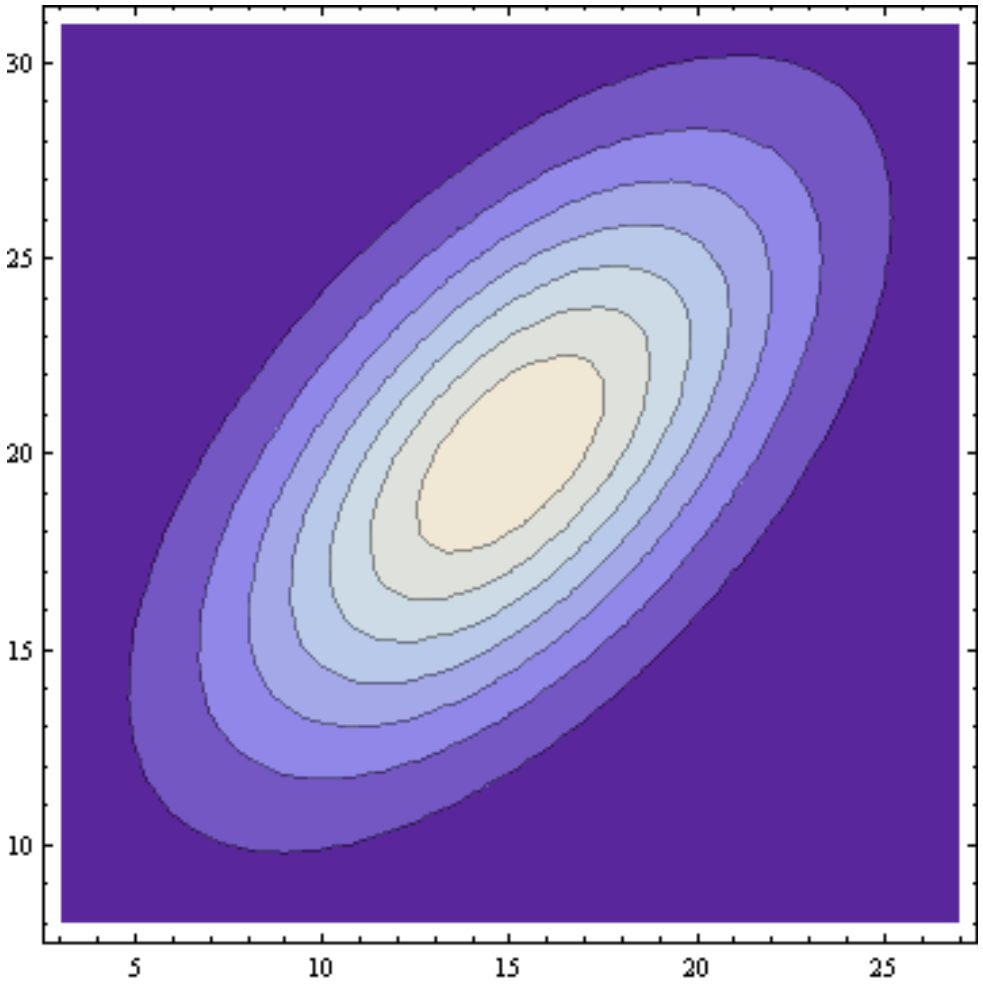
$$\phi(\mathbf{X}_1,\mathbf{X}_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[\frac{-1}{2(1-\rho^2)}\left\{\frac{(\mathbf{X}_1-\mu_1)^2}{\sigma_1^{\,2}}+\frac{(\mathbf{X}_2-\mu_2)^2}{\sigma_2^{\,2}}-2\rho\frac{(\mathbf{X}_1-\mu_1)(\mathbf{X}_2-\mu_2)}{\sigma_1\sigma_2}\right\}\right]$$

where $\sigma_i^{\,2}$ is the variance, μ_i is the mean, and $\rho=\dfrac{\mathrm{cov}(\mathbf{X}_1,\mathbf{X}_2)}{\sigma_1\sigma_2}=\dfrac{E(\mathbf{X}_1-\mu_1)E(\mathbf{X}_2-\mu_2)}{\sigma_1\sigma_2}$ is the correlation coefficient. Also note that the equation

$$\frac{(\mathbf{X}_1-\mu_1)^2}{\sigma_1^{\,2}}+\frac{(\mathbf{X}_2-\mu_2)^2}{\sigma_2^{\,2}}-2\rho\frac{(\mathbf{X}_1-\mu_1)(\mathbf{X}_2-\mu_2)}{\sigma_1\sigma_2}=C$$

represents an ellipse with center at point $(\mu_1,\mu_2)=(15,20)$; this point is known as the centroid of the bivariate population. Later, we will make use of the constant *C* in finding certain percentiles of distribution given the means for the bivariate population. Below is a graph of a bivariate normal distribution followed by its contour plot.





To continue to the *p*-variate case, it is convenient to rewrite the quantities in the equation for the bivariate case in matrix notation. The variance-covariance matrix or the dispersion matrix for the bivariate population can be written as

$$\Sigma_2 = \begin{bmatrix} \sigma_1^{\,2} & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^{\,2} \end{bmatrix}.$$

This matrix essentially tells us how the each attribute is related to each other. The dispersion matrix has similar mathematical results as a correlation matrix (Bernstein, 1988). The determinant of matrix Σ_2 is

$$|\Sigma_2| = \sigma_1^{\,2}\sigma_2^{\,2}\left(1-\rho^2\right)$$

and the inverse of matrix Σ_2 is given by

$$\Sigma_2^{-1} = \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_2\sigma_1 & \sigma_1^2 \end{bmatrix} = \frac{1}{\left(1-\rho^2\right)}\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_2\sigma_1} & \frac{1}{\sigma_2^2} \end{bmatrix}.$$

Now we can see that the expression in the exponent of the bivariate normal density function is equivalent to

$$\left(-\frac{1}{2}\right)\left[\frac{1}{\left(1-\rho^2\right)}\left\{\frac{\left(\mathbf{X}_1-\mu_1\right)^2}{\sigma_1^2}+\frac{\left(\mathbf{X}_2-\mu_2\right)^2}{\sigma_2^2}-2\rho\frac{\left(\mathbf{X}_1-\mu_1\right)\left(\mathbf{X}_2-\mu_2\right)}{\sigma_1\sigma_2}\right\}=\begin{bmatrix}\mathbf{X}_1-\mu_1 & \mathbf{X}_2-\mu_2\end{bmatrix}\Sigma_2^{-1}\begin{bmatrix}\mathbf{X}_1-\mu_1 \\ \mathbf{X}_2-\mu_2\end{bmatrix}\right]$$

Disregarding the -1/2 for now (exp(-1/2) is an irrelevant constant), let χ^2 represent this equation

$$\chi^2 = \mathbf{x}^T\Sigma_2^{-1}\mathbf{x} \quad \text{where} \quad \mathbf{x}^T = \begin{bmatrix}\mathbf{X}_1-\mu_1 & \mathbf{X}_2-\mu_2\end{bmatrix}.$$

We can also write the constant term of $\phi\left(\mathbf{X}_1,\mathbf{X}_2\right)$ as

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = \left(2\pi\right)^{-1}\left|\Sigma_2\right|^{-1/2}.$$

Finally, we can rewrite $\phi\left(\mathbf{X}_1,\mathbf{X}_2\right)$ more compactly as

$$\phi\left(\mathbf{X}_1,\mathbf{X}_2\right) = \left(2\pi\right)^{-1}\left|\Sigma_2\right|^{-1/2}\exp\left(\frac{-\chi^2}{2}\right).$$

We can now progress into the *p*-variate case with the new terms following

$$\Sigma_p = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \ldots & \rho_{2p}\sigma_2\sigma_p \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_2 & \ldots & \sigma_p^2 \end{bmatrix},$$

$$\chi^2 = \mathbf{x}^T\Sigma_p^{-1}\mathbf{x},$$

$$\mathbf{x}^T = \begin{bmatrix}\mathbf{X}_1-\mu_1 & \mathbf{X}_2-\mu_2 & \ldots & \mathbf{X}_p-\mu_p\end{bmatrix},$$

$$\phi(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p) = (2\pi)^{\frac{-p}{2}} |\Sigma_p|^{-\frac{1}{2}} \exp\left(\frac{-\chi^2}{2}\right) \sim N(\mu, \Sigma)$$

With the hyperellipsoid of the distribution being given by

$$\chi^2 = \mathbf{x}^T \Sigma_p{}^{-1} \mathbf{x} = C .$$

Note that the constant term is taken from comparing the bivariate case to the univariate case. The constant in the univariate normal density function is

$$(2\pi)^{-\frac{1}{2}} \left(\sigma^2\right)^{-\frac{1}{2}} .$$

From this we can show that the power of 2π seems to be (-1/2) times the number of variables, whereas the power of $|\Sigma_1| = \sigma^2$ is (-1/2) regardless of the number of variables.

The reason for using χ^2 to denote its expression is because when the $\mathbf{X}_i$ are statistically independent, χ^2 is a chi-square variate with *p* degrees of freedom. This is always true because even when the $\mathbf{X}_i$ are not independently distributed, a new set of variables $\mathbf{Y}_i$ can be produced that are independently distributed (Tatsuoka, 1971).

Once we have enough data to properly establish a probability distribution, we can then supply confidence regions to aid in the classification of applicants. For example, if we wanted our applicants to meet 90% of the population of scores for a particular trait, we could look up the χ^2 value that corresponded to .90. In this case, $P(\chi^2 \leq 4.605) = .90$. From this we can denote $R_2(4.605)$ as the region bounded by and including the ellipse

$$\frac{1}{1-p^2}\left[\frac{(\mathbf{X}_1 - \mu_1)^2}{\sigma_1{}^2} + \frac{(\mathbf{X}_2 - \mu_2)^2}{\sigma_2{}^2} - 2\rho\frac{(\mathbf{X}_1 - \mu_1)(\mathbf{X}_2 - \mu_2)}{\sigma_1\sigma_2}\right] = 4.605$$

where the subscript of *R* indicates the bivariate case. The geometrical interpretation of this is if we were to take many observations and assign them values $(\mathbf{X}_1, \mathbf{X}_2)$, then 90% of the scores would lie on or in the region $R_2(4.605)$.

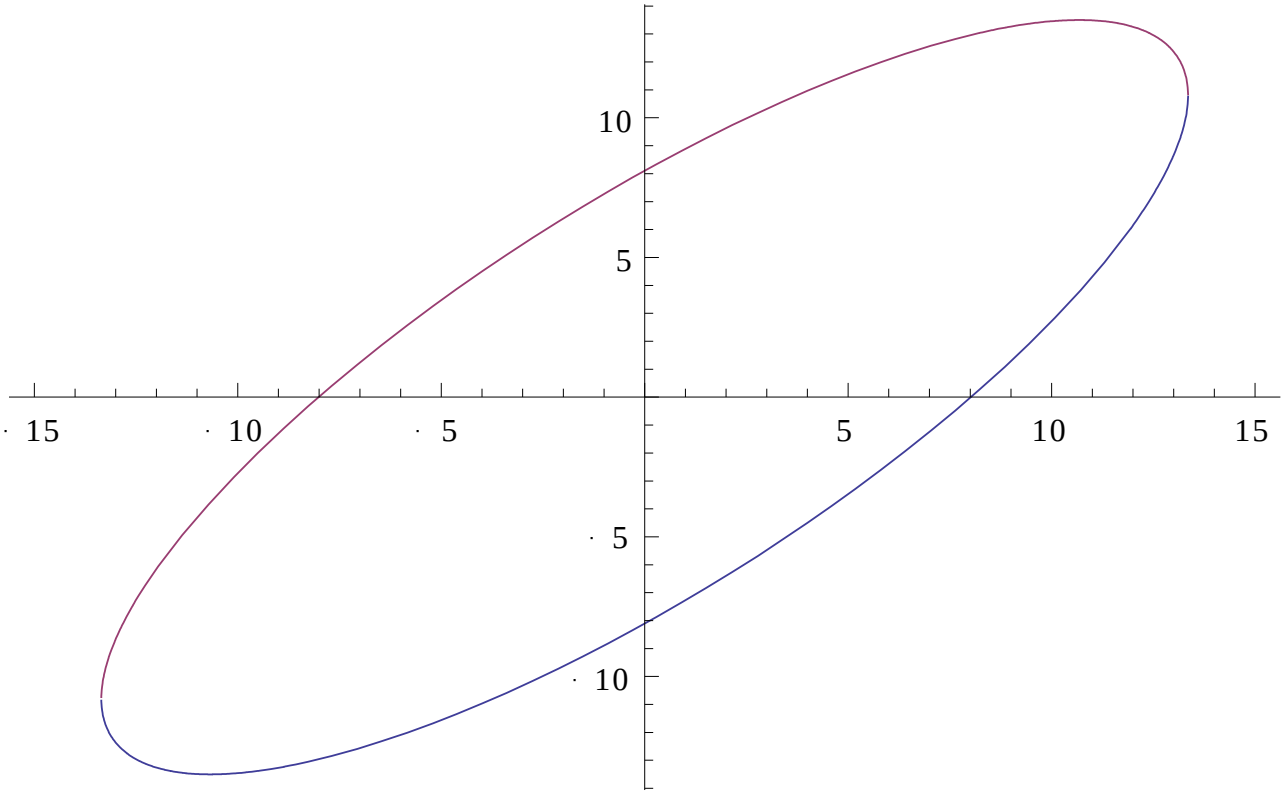
We can use this technique to determine other useful information such as the likelihood that any particular score has in the population. The following data is arbitrary. Consider a population where

$$\mu = \begin{bmatrix} 70 \\ 65 \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} 225 & 182 \\ 182 & 230 \end{bmatrix}.$$

Suppose we are interested in the 20%, that is, the equation of the ellipse inside or on which 20% of the population exists. $P(\chi^2 \le .4463) = .20$ from the equation we know that the population has the distribution

$$\mathbf{x}^T \Sigma^{-1} \mathbf{x} = \frac{1}{1-.8^2}\left[\frac{(\mathbf{X}_1 - 70)^2}{225} + \frac{(\mathbf{X}_2 - 65)^2}{230} - 1.6\frac{(\mathbf{X}_1 - 70)(\mathbf{X}_2 - 65)}{15 \cdot 15.17}\right] = .4463$$

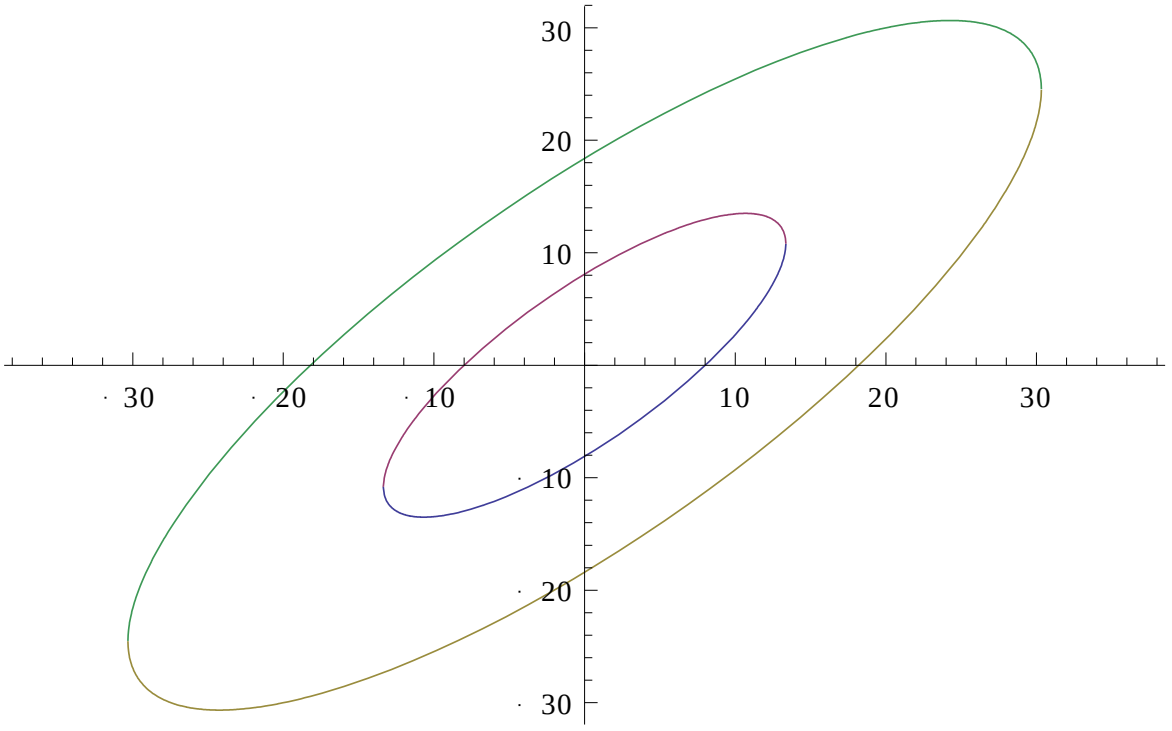
We can also use this to show the *p* percentile of scores. For example, the ellipse enclosing the 20% is shown below.



This procedure can be reversed to find the likelihood of a particular score. Suppose an applicant scores a (90, 95) on an assessment of ability with the population mean and variance-covariance matrix from above. We will determine the probability of this score.

$$\mathbf{x}^T \Sigma^{-1} \mathbf{x} = \frac{1}{1-.8^2}\left[\frac{(90 - 70)^2}{225} + \frac{(95 - 65)^2}{230} - 1.6\frac{(90 - 70)(95 - 65)}{15 \cdot 15.17}\right] = 4.089$$

Using the χ^2 distribution table with two degrees of freedom, we find that this point lies on the 87.05% range. This means that 87.05% of all applicants lie within – geometrically closer to the centroid or exactly the same distance – the region generated by the ellipse passing through the point (90, 95) and only 12.95% of all applicants fall on the outside of this ellipse. Below is the graph of this ellipse plotted with the 20% from above (interior ellipse).



Note: The ellipses in the above graphs are not centered at their original centroids. I have made the following substitutions for both ellipses for graphical and computational purposes: $\mathbf{X}_i - \mu_i = x_i$.

Because of the nature of the bivariate distribution, we can't be sure yet precisely what this means. We need to know in which region the score lies and how it relates to both the assessment and the centroid. According to Tatsuoka (1971), the angle of the axis of the ellipse is given by

$$\theta = \begin{cases} \frac{1}{2}\tan^{-1}\frac{2\rho\sigma_1\sigma_2}{{\sigma_1}^2 - {\sigma_2}^2} & \text{when} \quad \sigma_1 \neq \sigma_2 \\ 45^{\circ} & \text{when} \quad \sigma_1 = \sigma_2 \end{cases}$$

A line corresponding to this angle contains the major axis if $\rho > 0$, and the minor axis if $\rho < 0$.

It is important to note that the percentile of the applicant does not give us an indication of his or her abilities. We must make a few more computations to be able to actually use this statistic. Consider Applicant A, who scores a low score on both attributes. For arguments sake, Applicant A's score is in the 10% (again, that is only 10% of the population score the same or geometrically closer to the population mean). Furthermore, Applicant B, who scores very well on both attributes also scores in the 10%. Here it is clear that the percentile simply shows the likelihood of the particular score, not the aptitude. To find this, we must consider the assessment. If a higher score on an attribute can be interpreted as better, say mathematical abilities, then we have only to find the difference between the centroid and the score on the corresponding attribute. One could create another set of data in which the new score is the difference between the centroid and the score on the corresponding attribute multiplied by the percentile:

$$P\left(\mathbf{X} \le \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) \cdot \left(x_1 - \mu_1\right) = S_1 .$$

Note: this model does not consider the x_2 attribute in the computation of S_1. If both attributes increase the overall score continuously, then a possible model might be

$$P\left(\mathbf{X} \le \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) \cdot \left(\sqrt{\left(x_1 - \mu_1\right)^2 + \left(x_2 - \mu_2\right)^2} \right) \cdot \cos\left(\theta - \tan^{-1}\left(\frac{x_2 - \mu_2}{x_1 - \mu_1} \right) \right) = S_2$$

where θ is given previously. The model we decide to use is not as important as the concept that I am trying to exemplify: the model needs to behave in such a way that as the distance from the centroid increases, so does the score and as the angle the score vector minus the mean vector makes with θ increases, the score decreases. We want the score to decrease as it deviates from the major axis because the highest score will be the one that is farthest from the centroid and closest to the major axis. Note that the distance as described means the distance in coordination with an angle between 0 and 90 degrees; otherwise, we could have a long distance in the 'low score direction' and still have a good S_2. The problem of nonlinear scales now comes into play. Consider a scale that measures outgoingness. Image an applicant who is very talkative – to a fault. We might develop an assessment that screens out particular applicants on both ends of the score spectrum (i.e. a score of 8 has a higher yield than 10 or 5). This does not present a problem, because we need only to redefine the regression model. The contour plots can again be used to determine the score a person obtains and the least squares estimate can be used to predict the yield an applicant could potentially have.

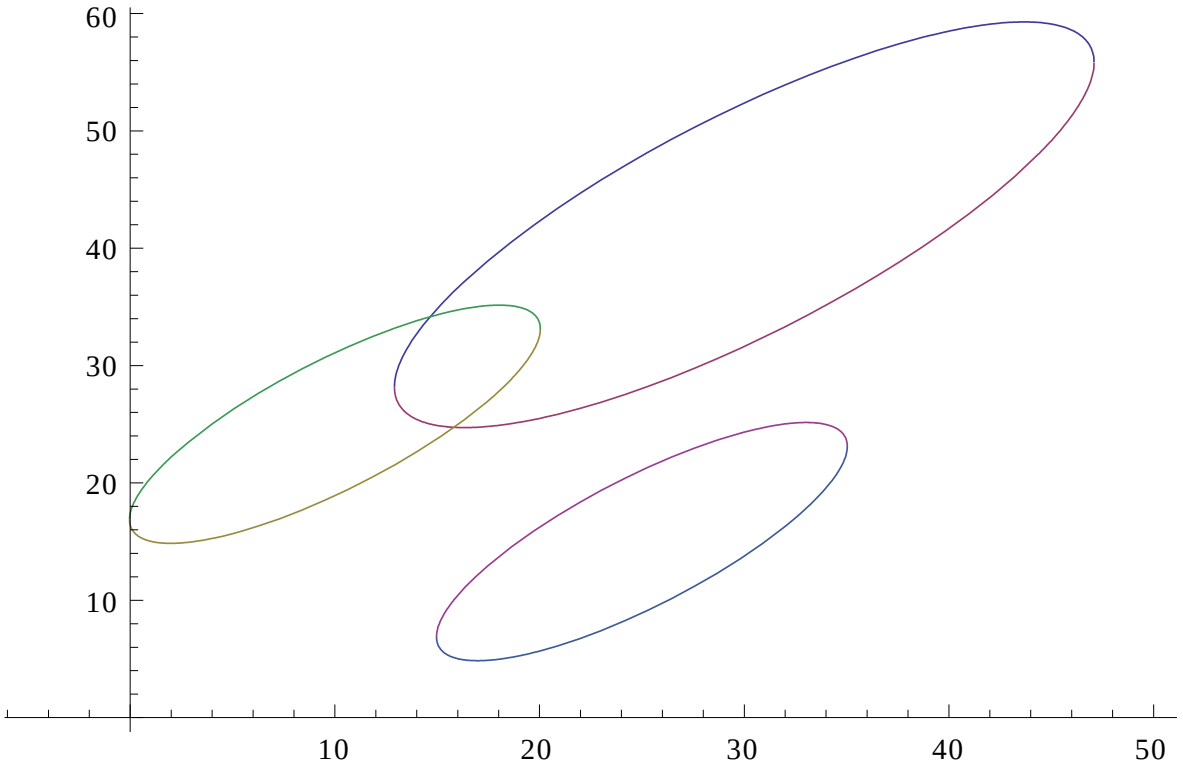
Perhaps a more useful application is to determine what percent of the population has a yield that falls within a range of scores. This is a useful technique because we can set realistic expectations for the people taking our assessment. For example, if we wanted to know how unlikely it would be for a person to score an 80% or above on our assessment, we simply find the volume of the multivariate normal distribution with the region on the x_1-x_2 plane being the cross-section of the regression model that corresponds to yield of 80. We use the formula

$$\iint_A \phi(\mathbf{X}_1, \mathbf{X}_2)dx_1dx_2 \quad \text{where} \quad A: \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 80$$

For the *p*-variate case,

$$\int \cdots \int_A \phi(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p)dx_1 \cdots dx_p \quad \text{where} \quad A: \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = 80.$$

Now I will focus more on the idea of classifying applicants based on his or her score on the assessment. In classification, we are interested in finding the minimum distance from a given score to a group's region, π_i. This makes intuitive sense in that a given score vector is closest to a particular mean. The image below is a good two dimensional geometric representation of how the mathematics works. Restating, we will be considering any particular score vector, **x**, and determining which group the applicant should be assigned to based on the distance the score is from the centroid of the group distributions. Below is a graph of three ellipses that might represent three sample distributions of a non specified likelihood.



This region can be altered by a priori probabilities, and by the cost of misclassification. A priori probabilities are probabilities that are logically determined from existing information. We would likely use priori probabilities to account for our need for particular positions. For example, the accounting department might only staff 8 people, but the sales department might staff 30

people. In this simple example, the priori probability of the accounting department is 8/38 and the priori probability of the sales department is 30/38. This type of alteration can be used based on the number of people we are hiring for each position. However, it is not necessary to know relative staffing information in every instance.

The cost of misclassification is simply the cost of classifying an individual incorrectly. Consider two positions, an accounting position and a cashiering position. The cost of misclassifying a potential accountant as a cashier is not as high as the cost of classifying a cashier as an accountant. The cashier might be unskilled in financial mathematics and would thus cost the company more than an individual who is not meeting his or her fullest potential. Such costs are usually determined by company guidelines and then used computationally by the psychologist.

To find each classification region, we will first need to compare the ratio of the densities of the two multivariate normal populations:

$$\frac{\phi(\mathbf{X}_1, \mathbf{X}_2)}{\phi(\mathbf{Y}_1, \mathbf{Y}_2)} \to \frac{\phi_1(x)}{\phi_2(x)} = \frac{\exp\left(\frac{-\chi_1^{\,2}}{2}\right)}{\exp\left(\frac{-\chi_2^{\,2}}{2}\right)} = \exp\left[-\frac{1}{2}\left(\chi_1^{\,2} - \chi_2^{\,2}\right)\right].$$

Please note the change in notation. For this model, we must assume that the two populations have equal variance-covariance matrices, $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. The region of classification is the set of *x*'s that make the above equation greater than or equal to *k* (Anderson, 1984). We choose the logarithmic function because it is monotonically increasing; thus, the resulting inequality becomes

$$-\frac{1}{2}\left(\chi_1^{\,2} - \chi_2^{\,2}\right) \geq \log k\,.$$

As a consequence, the best regions of classification are given by

$$R_1 : -\frac{1}{2}\left(\chi_1^{\,2} - \chi_2^{\,2}\right) \geq \log k$$
$$R_2 : -\frac{1}{2}\left(\chi_1^{\,2} - \chi_2^{\,2}\right) < \log k$$

If priori probabilities q_1 and q_2 are known, then *k* is given by

$$k = \frac{q_2 C(1 \mid 2)}{q_1 C(2 \mid 1)}.$$

Where $C(i \mid j)$ is the cost of misclassification into group *i* given that it was taken from or belongs to *j*. Otherwise, ln *k* = *c,* for some *c* that is suitably chosen.

To determine the probability of misclassification with two groups, we must first introduce the Mahalanobis squared distance between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$:

$$\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2).$$

From this, we find that the distribution of the ratio of the densities is $N\left(\frac{1}{2}\Delta^2, \Delta^2\right)$, and thus the probabilities of misclassification for R_1 and R_2 respectively are

$$R_1 : P(2 \mid 1) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}\Delta} \exp \frac{-\frac{1}{2}\left(z - \frac{1}{2}\Delta^2\right)^2}{\Delta^2} dz = \int_{-\infty}^{l} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy; l = \frac{\left(c - \frac{1}{2}\Delta^2\right)}{\Delta}$$

$$R_2 : P(1 \mid 2) = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}\Delta} \exp \frac{-\frac{1}{2}\left(z + \frac{1}{2}\Delta^2\right)^2}{\Delta^2} dz = \int_{m}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy; m = \frac{\left(c + \frac{1}{2}\Delta^2\right)}{\Delta}.$$

For the minimax solution, we find *c* such that

$$C(1 \mid 2) \cdot \int_{m}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = C(2 \mid 1) \cdot \int_{-\infty}^{l} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

where *m* and *l* are defined.

Note that *l* and *m* are found from the transformation that we are familiar with: If $\mathbf{X} \sim N(\mu, \sigma^2)$, then

$$P(a < \mathbf{X} < b) = P\left(\frac{a - \mu}{\sigma} < \mathbf{X} < \frac{b - \mu}{\sigma}\right).$$

If the costs of misclassification are equal, then the probability of misclassification becomes

$$\int_{\Delta/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

Now, we will progress into the case of classification into several multivariate normal populations. For now, we will assume that the means and the variance-covariance matrices are different but the cost of misclassification for each group is equal. First, we examine the Mahalanobis squared distance for the general case:

$$\Delta_{ji}^{\ 2} = (\mu_j - \mu_i)^T \Sigma^{-1} (\mu_j - \mu_i)$$

and define the Mahalanobis squared distance function as

$$\Delta_i^{\ 2}(x) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i).$$

Now to classify simply means to find the minimum distance between a raw score and a mean score. We obtain this by comparing the distances $\Delta_i^{\ 2}(x)$ for each π_i. We will use the discriminant function to classify **x** to π_i when the linear discriminant score is largest:

$$d_i(x) = (\mu_i)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_i)^T \Sigma^{-1} (\mu_i) + \ln(p_i)$$

We can use the above equation to compare two scores at the same time. The term $\ln\left(\frac{p_2}{p_1}\right)$ is meant to place the plane of separation closer to μ_1 than μ_2 if p_2 is greater than p_1. We assign **x** to π_k if

$$(\mu_k - \mu_i)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_k - \mu_i)^T \Sigma^{-1} (\mu_k + \mu_i) \geq \ln\left(\frac{p_i}{p_k}\right) \forall i = 1,2,\ldots,n.$$

For example, if we want to partition a space into three regions (n = 3), we have only to define the regions so that R_1 consists of all **x** satisfying

$$R_1 : (\mu_k - \mu_i)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_k - \mu_i)^T \Sigma^{-1} (\mu_k + \mu_i) \geq \ln\left(\frac{p_i}{p_k}\right); i = 2,3.$$

Restating, R_1 consists of all **x** satisfying both of the following equations simultaneously (in which I will introduce a new notation, $d_{ij}(x)$):

$$d_{12}(x) = (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) \geq \ln\left(\frac{p_2}{p_1}\right)$$

$$d_{13}(x) = (\mu_1 - \mu_3)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_1 - \mu_3)^T \Sigma^{-1} (\mu_1 + \mu_3) \geq \ln\left(\frac{p_3}{p_1}\right)$$

Anderson (1984) uses similar calculations, which seem to take into accout a priori probabilities better that does Johnson and Wichern (1988) and similarly Tatsuoka (1971). The calculations to compute the statistics for Anderson are more difficult and some are beyond the scope of this paper. However, I will show the formulas used in Anderson's book for comparison reasons.

First, he makes use of the function

$$u_{jk}(x) = \log \frac{\phi_j(x)}{\phi_k(x)} = \left[x - \frac{1}{2}(\mu_j - \mu_k) \right]^T \Sigma^{-1}(\mu_j - \mu_k).$$

If priori probabilities (given by q_i) are known, then the region R_j is defined by the vector **x** satisfying

$$R_j : u_{jk}(x) > \log \frac{q_k}{q_j}, \qquad k = 1,...,m; k \neq j .$$

If no priori probabilities are known, the then region is defined by

$$R_j : u_{jk}(x) \geq c_j - c_k, \qquad k = 1,...,m; k \neq j .$$

If priori probabilities are known, and they are equal, then the region is defined by

$$R_j : u_{jk}(x) \geq 0, \qquad k = 1,...,m; k \neq j$$

In this case, we will need to find the probabilities of correct classification. The ultimate goal or is to find the constants so that $P(i \mid i, R_i)$ for each region are equal. If **X** is a random observation then let

$$U_{ji} = \left[\mathbf{X} - \frac{1}{2}(\mu_i + \mu_j) \right]^T \Sigma^{-1}(\mu_j - \mu_i).$$

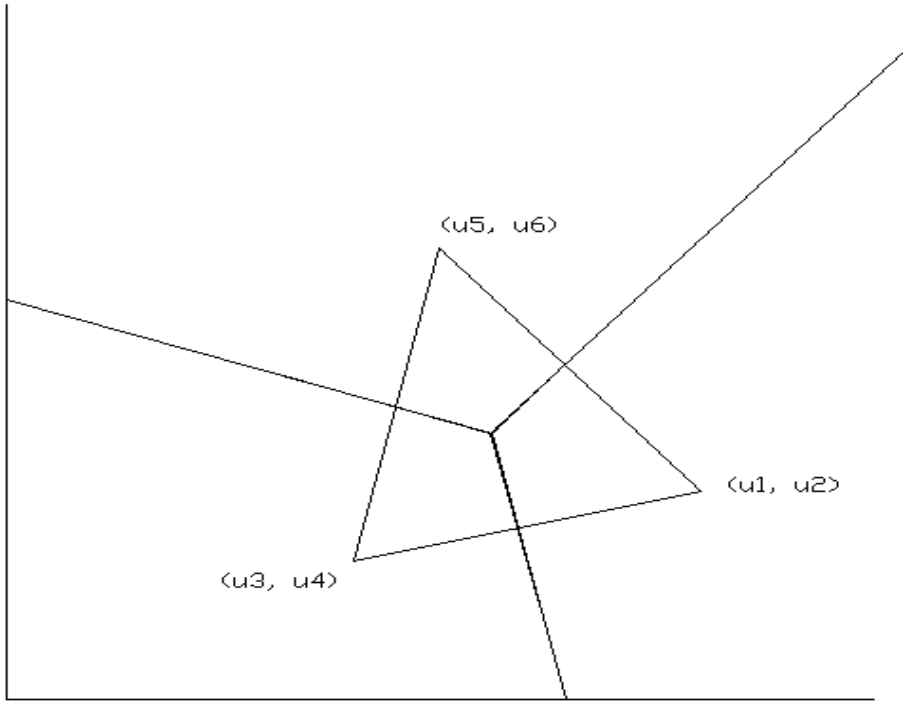
The distances will be the same for U_{ji} and U_{ij} but because it is a vector, $U_{ji} = -U_{ij}$. Again, the Mahalanobis squared distance becomes

$$\Delta_{ji}^{\ 2} = (\mu_j - \mu_i)^T \Sigma^{-1}(\mu_j - \mu_i).$$

At this point, we have only to determine the constants that make the integrals equal:

$$\int_{c_m - c_j}^{\infty} \cdots \int_{c_j - c_1}^{\infty} f_j \, du_{j1} \cdots d_{j,j-1} \cdots du_{jm}$$

where f_j is the density of U_{ji}, $i = 1,2,...,m, i \neq j$. The application of integration can be interpreted geometrically as the equal division of regions in space. The illustration below should help illustrate this point. This is the bivariate case with three groups.



The verticies of the triangle are the means of the three groups.

Now I will create an example of classifying an individual into one of three groups. Let the groups consist of accountants (π_1), cashiers (π_2), and managers (π_3). Let an assessment be developed so that it measures verbal ability (x_1), mathematical ability (x_2), and integrity (x_3). Suppose the following table has modeled a company for many years, and acquired a large value of *N*. The mean scores are shown below.

	Accountants (π_1)	Cashiers (π_2)	Managers (π_3)
Verbal Ability (x_1)	63	85	81
Math Ability (x_2)	96	61	70
Integrity (x_3)	85	43	87

$$\Sigma = \begin{bmatrix} 25 & 18 & 19 \\ 16 & 27 & 17 \\ 13 & 15 & 23 \end{bmatrix}$$

Also suppose priori probabilities are known, and they are shown below.

$$p_1 = .25$$
$$p_2 = .65$$
$$p_3 = .1$$

For actual results, we need a raw score. Suppose an applicant's score vector is shown below.

$$\begin{bmatrix} 65 & 91 & 79 \end{bmatrix}^T$$

With this, we need only to make the calculations to find the minimum distance:

$$\text{Recall: } d_{ij}(x) = \left(\mu_i - \mu_j\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i + \mu_j\right) \geq \ln\left(\frac{p_j}{p_i}\right)$$

$$d_{13}(x) = \left(\mu_1 - \mu_3\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_1 - \mu_3\right)^T \Sigma^{-1}\left(\mu_1 + \mu_3\right) \geq \ln\left(\frac{p_3}{p_1}\right) \Rightarrow 36.8 \geq -.9163$$

$$d_{12}(x) = \left(\mu_1 - \mu_2\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_1 - \mu_2\right)^T \Sigma^{-1}\left(\mu_1 + \mu_2\right) \geq \ln\left(\frac{p_2}{p_1}\right) \Rightarrow 117 \geq .9555$$

$$d_{23}(x) = \left(\mu_2 - \mu_3\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_2 - \mu_3\right)^T \Sigma^{-1}\left(\mu_2 + \mu_3\right) \geq \ln\left(\frac{p_3}{p_2}\right) \Rightarrow -68.95 \leq -1.872$$

$$d_{21}(x) = \left(\mu_2 - \mu_1\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_2 - \mu_1\right)^T \Sigma^{-1}\left(\mu_2 + \mu_1\right) \geq \ln\left(\frac{p_1}{p_2}\right) \Rightarrow -117 \leq -.9555$$

$$d_{31}(x) = \left(\mu_3 - \mu_1\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_3 - \mu_1\right)^T \Sigma^{-1}\left(\mu_3 + \mu_1\right) \geq \ln\left(\frac{p_1}{p_3}\right) \Rightarrow -36.8 \leq .9163$$

$$d_{32}(x) = \left(\mu_3 - \mu_2\right)^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_3 - \mu_2\right)^T \Sigma^{-1}\left(\mu_3 + \mu_2\right) \geq \ln\left(\frac{p_2}{p_3}\right) \Rightarrow 68.95 \geq 1.872$$

From the calculations, you should see a few things: $d_{ij} = -d_{ji}$, and $\mathbf{x}$ only satisfied one pair of equations, namely d_{13} and d_{12}. Therefore, we make the classification that this applicant would be best suited as an accountant. We can then take this information and calculate what percentage

of the population scores closer to the mean for accountants and make an inference as to whether or not we should consider this individual for employment.

I will continue with the same example using Anderson's technique to show his method for defining the regions until the probability density function needs to be integrated. Using our formulas from above, we get

$$u_{12}(x)=\left[x-\frac{1}{2}(\mu_1+\mu_2)\right]^T\Sigma^{-1}(\mu_1-\mu_2)=-4.8375x_1+2.1916x_2+3.1311x_3+14.456=-u_{21}(x)$$
$$u_{13}(x)=\left[x-\frac{1}{2}(\mu_1+\mu_3)\right]^T\Sigma^{-1}(\mu_1-\mu_3)=-2.1679x_1+2.5975x_2-.5556x_3-11.7221=-u_{31}(x)$$
$$u_{23}(x)=\left[x-\frac{1}{2}(\mu_2+\mu_3)\right]^T\Sigma^{-1}(\mu_2-\mu_3)=2.6696x_1+.4059x_2-3.6867x_3-8.5278=-u_{32}(x)$$

For the first part, if we assume that priori probabilities are known, and they are equal, then the best set of regions of classification are

$$\begin{aligned}R_1&=u_{12}(x)\geq 0,u_{13}(x)\geq 0\\R_2&=u_{21}(x)\geq 0,u_{23}(x)\geq 0\\R_3&=u_{31}(x)\geq 0,u_{32}(x)\geq 0\end{aligned}.$$

If any priori probabilities are known, but not equal, then the resulting equations become:

$$R_1:u_{12}(x)>\ln\frac{p_2}{p_1};u_{13}(x)>\ln\frac{p_3}{p_1}$$
$$R_2:u_{23}(x)>\ln\frac{p_3}{p_2};u_{21}(x)>\ln\frac{p_1}{p_2}$$
$$R_3:u_{31}(x)>\ln\frac{p_1}{p_3};u_{32}(x)>\ln\frac{p_2}{p_3}$$

If priori probabilities exist but are unknown, then the computations become more difficult. This will be completed in another work. However, the density functions are the combined distribution functions whose corresponding random functions are shown below:

$$\begin{aligned}f_1&:U_{12},U_{13}\\f_2&:U_{23},U_{21}\\f_3&:U_{31},U_{32}\end{aligned}$$

At this point, we have only to integrate each density function and solve the system of integrals and solve for the constants c_k so that the constants set each region to be the same in size (the integrals must be equivalent for this to happen).

$$R_j : u_{jk}(x) \geq c_j - c_k, \qquad k = 1,...,m; k \neq j$$

In this paper, you have seen the mathematics involved in the delicate procedure of classification. It is important to note that there have been many different procedures designed to further analyze the ideas discussed here. To construct the paper, I have examined some of these procedures but decided to only use those selected. Additional works will analyze the effectiveness of the remaining procedures as well as show the calculations from Anderson in their full extent.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). New York: John Wiley and Sons.

Bernstein, I. H. (1988). *Applied Multivariate Analysis*. New York: Springer-Verlag New York, Inc.

Johnson, R. A., & Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis* (2nd ed.). Prentice – Hall, Inc.

Montgomery, D. C. (2005). *Design and Analysis of Experiments* (6th ed.). John Wiley and Sons, Inc.

Overall, J. E., & Klett, C. J. (1972). *Applied Multivariate Analysis.* New York: McGraw-Hill Book Company.

Rosner, B. (2005). *Fundamentals of Biostatistics* (6th ed.). Thomson Brooks/Cole.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics* (4th ed.). Boston: Allyn and Bacon.

Tatsuoka, M. M. (1971). *Techniques for Educational and Psychological Research.* New York: John Wiley & Sons, Inc.