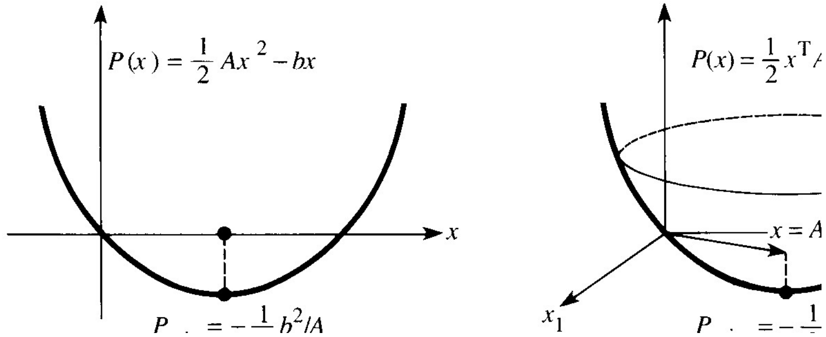
Minimum and Solving $Ax=b$



- **If A is symmetric positive definite, then** $P(x) = \frac{1}{2}x^{\mathrm{T}}Ax - x^{\mathrm{T}}b$ **reaches its minimum at the point where $Ax=b$. At that point** $P_{\min} = -\frac{1}{2}b^{\mathrm{T}}A^{-1}b$

***Proof*: Let x be the solution of $Ax=b$. For any vector y:**

$$\begin{aligned} P(y) - P(x) &= \tfrac{1}{2}y^{\mathrm{T}}Ay - y^{\mathrm{T}}b - \tfrac{1}{2}x^{\mathrm{T}}Ax + x^{\mathrm{T}}b \\ &= \tfrac{1}{2}y^{\mathrm{T}}Ay - y^{\mathrm{T}}Ax + \tfrac{1}{2}x^{\mathrm{T}}Ax \\ &= \tfrac{1}{2}(y-x)^{\mathrm{T}}A(y-x) > 0 \end{aligned}$$

since A is positive definite

Example: Minimize $P(x) = x_1^2 - x_1x_2 + x_2^2 - b_1x_1 - b_2x_2$

Calculus: $\begin{aligned} \partial P/\partial x_1 &= 2x_1 - x_2 - b_1 = 0 \\ \partial P/\partial x_2 &= -x_1 + 2x_2 - b_2 = 0 \end{aligned}$

Linear algebra: solve $Ax=b$ where $P(x)$=$\frac{1}{2}x^{\mathrm{T}}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}x - x^{\mathrm{T}}b$

Minimum/Maximum and Solving $Ax=\lambda x$

- ***Rayleigh's quotient*:**

$$R(x) = \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x}$$

- ***Rayleigh's Principle:***

The quotient $R(x)$ is maximized by the first eigenvector $x=x_1$ of A corresponding to the largest eigenvalue λ_1 and its maximum value is λ_1: $R(x_1) = \dfrac{x_1^{\mathrm{T}}Ax}{x_1^{\mathrm{T}}x_1} = \dfrac{x_1^{\mathrm{T}}\lambda_1x_1}{x_1^{\mathrm{T}}x_1} = \lambda_1$

***Geometrically*:**

Fix numerator at 1: $x^{\mathrm{T}}Ax=1$ ellipsoid

$\Rightarrow$ denominator $x^{\mathrm{T}}x=\|x\|^2$ as small as possible $\Rightarrow$ shortest axis $\Rightarrow$

smallest $1/\sqrt{\lambda_i}$ **$\Rightarrow$ largest eigenvalue λ_1**

***Algebraically*: Diagonalize $A \Rightarrow A=Q\Lambda Q^{\mathrm{T}}$ ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$)**

$$R(x) = \frac{\left(Q^{\mathrm{T}}x\right)^{\mathrm{T}}\Lambda\left(Q^{\mathrm{T}}x\right)}{\left(Q^{\mathrm{T}}x\right)^{\mathrm{T}}\left(Q^{\mathrm{T}}x\right)} = \frac{y^{\mathrm{T}}\Lambda y}{y^{\mathrm{T}}y} = \frac{\lambda_1y_1^2 + \cdots + \lambda_ny_n^2}{y_1^2 + \cdots + y_n^2} \leq \lambda_1 \; (\geq \lambda_{\mathrm{n}})$$

since $\lambda_1\left(y_1^2 + y_2^2 + \cdots + y_n^2\right) \geq \left(\lambda_1y_1^2 + \lambda_2y_2^2 + \cdots + \lambda_ny_n^2\right)$

The maximum R must be at $y_1=1$ and $y_2=y_3=\ldots=y_n=0$

- **Rayleigh quotient is never above λ_1 and never below λ_n**

Multivariate Analysis

- **In multivariate analysis, we try to find some pattern, or some natural structure, by observing data of more than one variable.**
- **An example of typical unsupervised learning problem: we observe <u>53 blood and urine measurements without knowing who are alcoholic</u>. With data of so many measurements, is there any pattern we can find to distinguish alcoholics from non-alcoholics?**
- **Let us take blood and urine samples from 65 persons (33 alcoholics and 32 non-alcoholics) and obtain 53 measurements from each person.**

	Measurement 1	Measurement 2	…	Measurement 53
Person 1	x_{11}	x_{12}	…	$x_{1,53}$
⋮	⋮	⋮	⋮	⋮
Person 65	$x_{65,1}$	$x_{65,2}$	…	$x_{65,53}$

- **Each measurement (say, *i*th measurement) varies over different persons, (e.g. x_{1i}, x_{2i}, …, $x_{(65)i}$ are different).**
- **The sample mean (average) of each measurement:** $\bar{x}_i = \dfrac{\sum_{k=1}^{65} x_{ki}}{65}$

Sample Variances and Sample Co-variances

- **The variation of the *i*th measurement over different persons is measured by sample variance: (the average squared distance from the sample mean)**

$$\textbf{Sample Variance} = SVar(x_i) = \sigma_i^2 = \frac{\sum_{k=1}^{65}(x_{ki} - \bar{x}_i)^2}{65-1};$$

$$\textbf{Sample Standard Deviation} = \sqrt{SVar(x_i)} = \sqrt{\frac{\sum_{k=1}^{65}(x_{ki} - \bar{x}_i)^2}{65-1}}\ \sqrt{\sigma_i^2} = \sigma_i$$

- **There must be correlations among measurements. For example, the higher blood pressure level often comes with the higher cholesterol level. A larger body weight often comes with a greater height.**
- **The co-variation of two measurements, say the *i*th and the *j*th, is measure by sample covariance: (the trend that one is larger then the other is larger or one is smaller then the other is smaller)**

$$SCov(x_i, x_j) = \sigma_{ij} = \frac{\sum_{k=1}^{65}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{65-1}$$

- ***i*th ↑, *j*th↑ ⇒ SCov>0: positively correlated**
- ***i*th ↑, *j*th↓ ⇒ SCov<0: negatively correlated**

Sample Covariance Matrix

- The *symmetric* sample covariance matrix is formed by variances and covariances The diagonal elements are variances and the off-diagonal elements are covariances:

$$\Sigma = \begin{bmatrix} \sigma_1^2 = \dfrac{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)^2}{65-1} & \sigma_{12} = \dfrac{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)(x_{k2}-\bar{x}_2)}{65-1} & \cdots & \sigma_{1,53} \\ \sigma_{21} & \sigma_2^2 & & Cov(2,53) \\ \vdots & & \ddots & \vdots \\ \sigma_{53,1} & \sigma_{53,2} & \cdots & \sigma_{53}^2 \end{bmatrix}$$

- Let all measurements x_{kj} are moved to be around zero (centering: $x_{ki}-\bar{x}_i$) and let

$$A = \begin{bmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1,53}-\bar{x}_{53} \\ \vdots & \vdots & \vdots \\ x_{65,1}-\bar{x}_1 & \cdots & x_{65,53}-\bar{x}_{53} \end{bmatrix}$$

Then,

$$\mathbf{\Sigma} = \frac{1}{65-1}A^T A$$

Normalized x_{ki} and Correlation Coefficient

- Let all measurements x_{ki} be *centered* to be around zero (centering: $x_{ki}-\bar{x}_i$) and be *scaled* to have equal variance (=1):

$$y_{ki} = \frac{x_{ki}-\bar{x}_i}{\sigma_i} = \frac{x_{ki}-\bar{x}_i}{\sqrt{\sum_{k=1}^{65}(x_{ki}-\bar{x}_i)^2/(65-1)}}$$

- y_{ki} is called *normalized* measurements.

- $$\overline{y_i} = \frac{\sum_{k=1}^{65} y_{ki}}{65} = \frac{\sum_{k=1}^{65}\frac{x_{ki}-\bar{x}_i}{\sigma_i}}{65} = \frac{\sum_{k=1}^{65}x_{ki}-65\bar{x}_i}{65\sigma_i} = \sigma_i\left(\frac{\sum_{k=1}^{65}x_{ki}}{65}-\bar{x}_i\right) = 0$$

- $SVar(y_i) = 1$

$$SCov(y_i, y_j) = \frac{\sum_{k=1}^{65}(y_{ki}-\bar{y}_i)(y_{kj}-\bar{y}_j)}{65-1} = \frac{\sum_{k=1}^{65}(y_{ki}-0)(y_{kj}-0)}{65-1}$$

$$= \frac{1}{65-1}\sum_{k=1}^{65}\frac{x_{ki}-\bar{x}_i}{\sigma_i}\frac{x_{kj}-\bar{x}_j}{\sigma_j} = \frac{\sum_{k=1}^{65}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)/(65-1)}{\sigma_i\sigma_j} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}$$

- Correlation coefficient is defined as ($-1 \le \rho_{ij} \le 1$):

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\sigma_j} = \frac{\sum_{k=1}^{65}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{65}(x_{ki}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{65}(x_{kj}-\bar{x}_j)^2}}$$

- $-1 \le \rho_{ij} < 0$: negatively correlated; $0 < \rho_{ij} \le 1$: positively correlated
- $\rho_{ij}=0$: no correlation; $\rho_{ij}=\pm 1$: perfect correlation (all $\rho_{ii}=1$)

Correlation Matrix

- **The *symmetric* correlation matrix ρ is formed by all correlation coefficients (ρ_{ij}). All diagonal elements are 1 (ρ_{ii}=1) and all the off-diagonal elements are between −1 and 1 ($-1\le\rho_{ij}\le 1$ where $i\neq j$)**

- **Let** $\mathbf{B}=\begin{bmatrix} y_{11} & \cdots & y_{1,53} \\ \vdots & \vdots & \vdots \\ y_{65,1} & \cdots & y_{65,53} \end{bmatrix}$

Then,

$$\rho=\begin{bmatrix} \rho_{11}=\dfrac{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)(x_{k1}-\bar{x}_1)}{\sqrt{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)^2}\sqrt{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)^2}}=1 & \rho_{12} & \cdots & \rho_{1,53} \\ \rho_{21}=\dfrac{\sum_{k=1}^{65}(x_{k2}-\bar{x}_2)(x_{k1}-\bar{x}_1)}{\sqrt{\sum_{k=1}^{65}(x_{k2}-\bar{x}_2)^2}\sqrt{\sum_{k=1}^{65}(x_{k1}-\bar{x}_1)^2}} & \rho_{22} & & \rho_{2,53} \\ \vdots & & \ddots & \vdots \\ \rho_{53,1} & \rho_{53,2} & \cdots & \rho_{53,53} \end{bmatrix}$$

$$=\frac{1}{65-1}\mathbf{B}^T\mathbf{B}$$

Weighted Index for Analysis

- **Our purpose now is to find a weighted index (linear combination of measurements), $z_k=a_1y_{k1}+a_2y_{k2}+a_3y_{k3}+\ldots+a_{53}y_{k,53}=a^{\mathrm{T}}y_k$, to maximize the differences among people's z_k ($=a^{\mathrm{T}}y_k$), i.e., Max $SVar(z)$. $z_k(=a^{\mathrm{T}}y_k)$ can then be a possible measurement to distinguish the alcoholics from non-alcoholics;**
- **Average of z_k is zero:** $\bar{z}=a_1\bar{y}_1+a_2\bar{y}_2+\cdots+a_{53}\bar{y}_{53}=0$ ($\because\ \bar{y}_i=0$)
- $Var(z)=\dfrac{\sum_{k=1}^{65}(z_k-\bar{z})^2}{65-1}=\dfrac{\sum_{k=1}^{65}(a^Ty_k-0)^2}{65-1}=\dfrac{a^T\mathbf{B}^T\mathbf{B}a}{65-1}=\boldsymbol{a^T\rho a}$

where $\mathbf{B}a=\begin{bmatrix} y_{11} & \cdots & y_{1,53} \\ \vdots & \vdots & \vdots \\ y_{65,1} & \cdots & y_{65,53} \end{bmatrix}\begin{bmatrix} a_1 \\ \vdots \\ a_{53} \end{bmatrix}=\begin{bmatrix} a^Ty_1 \\ \vdots \\ a^Ty_{65} \end{bmatrix}$

- **The weights a_i are relative weights. There could be infinite possible choices for the same relative weights. Example: 1:2:1≡2:4:2≡**$1/\sqrt{6}:2/\sqrt{6}:1/\sqrt{6}\equiv\cdots$

  ⇒ Choose a unit length a : $a_1^2+a_2^2+\cdots+a_{53}^2=1$
- **Thus, our problem become:**

  Max: $a^T\mathbf{B}^{\mathrm{T}}\mathbf{B}a$ Subject to: $a^{\mathrm{T}}a$=1

Solving Weights *a*

Max: $a^T\mathbf{B}^T\mathbf{B}a$

Subject to: $a^Ta=1$

- **First method: Max** $a^T\mathbf{B}^T\mathbf{B}a \equiv \frac{a^T\mathbf{B}^T\mathbf{B}a}{a^Ta}$ $(\because a^Ta=1)$

  Max $\frac{a^T\mathbf{B}^T\mathbf{B}a}{a^Ta}$ **(Rayleigh Quotient)**

  $\Rightarrow \frac{a^T\mathbf{B}^T\mathbf{B}a}{a^Ta}$ **is maximized by the first eigenvector of $\mathbf{B}^T\mathbf{B}$**

- **Since $\mathbf{B}^T\mathbf{B}=\rho$ is real symmetric and positive semidefinite, it can be diagonalized with orthonormal eigenvectors. Let eigenvalue-eigenvector pairs (λ_1, e_1), (λ_2, e_2),…., (λ_{53}, e_{53}) where $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_{53} \geqq 0$ (positive semidefinite). That is,**

$$Q^T\mathbf{B}^T\mathbf{B}Q = \Lambda = \begin{bmatrix} - & e_1 & - \\ & \vdots & \\ - & e_{53} & - \end{bmatrix} \mathbf{B}^T\mathbf{B} \begin{bmatrix} | & & | \\ e_1 & \cdots & e_{53} \\ | & & | \end{bmatrix} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{53} \end{bmatrix}$$

- **Rayleigh's principal: $a=e_1$ maximizes** $\frac{a^T\mathbf{B}^T\mathbf{B}a}{a^Ta} = a^T\mathbf{B}^T\mathbf{B}a=\lambda_1$

- **Second method: Max $a^T\mathbf{B}^T\mathbf{B}a - \lambda(a^Ta-1)$ (by Lagrange multiplier)**

  $\Rightarrow$ **Max $a^T(\mathbf{B}^T\mathbf{B} - \lambda\mathbf{I})a-\lambda$ (Recall: Max** $P(x)=\frac{1}{2}x^TAx-x^Tb$**)**

  $\Rightarrow$ **$(\mathbf{B}^T\mathbf{B} - \lambda\mathbf{I})a=0$**

  $\Rightarrow$ **$\mathbf{B}^T\mathbf{B}a=\lambda a \Rightarrow a$: eigenvector λ: corresponding eigenvalue.**

Principal Component Analysis (PCA)

- **Our problem: Max $a^T\mathbf{B}^T\mathbf{B}a$ subject to $a^Ta=1$**

- **Let $a=e_1$ and $\lambda=\lambda_1$:**

$$z_k = \mathbf{e}_1^T\mathbf{y} = e_{11}y_{k1} + e_{12}y_{k2} + \ldots + e_{153}y_{k,53}$$

  where $\mathbf{e}_1^T = [e_{11}, e_{12}, \ldots e_{1,53}]$

- ***Var* (z)=$e_1^T\mathbf{B}^T\mathbf{B}e_1 = \mathbf{e}_1^T\lambda_1\mathbf{e}_1 = \lambda_1\mathbf{e}_1^T\mathbf{e}_1=\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_{53} \geqq 0$**

  $\Rightarrow$ **$a=e_1$ and $\lambda=\lambda_1$ is the solution for our max variation problem**

- $\mathbf{e}_1^T\mathbf{y} = e_{11}y_{k1} + e_{12}y_{k2} + \ldots + e_{153}y_{k,53}$ **is called the *First Principal Component***

- **The Second Principal Component:**

  $\mathbf{e}_2^T\mathbf{y} = e_{21}y_{k1} + e_{22}y_{k2} + \ldots + e_{2,53}y_{k,53}$

- **Properties of the 2nd principal component**

  1. $\lambda_1 \geqq$ *Var* (z_k)=$\mathbf{e}_2^T\rho\mathbf{e}_2 = \mathbf{e}_2^T\lambda_2\mathbf{e}_2 = \lambda_1\mathbf{e}_2^T\mathbf{e}_2=\lambda_2 \geqq \lambda_3 \ldots \geqq \lambda_{53} \geqq 0 \Rightarrow$ has the second largest variance

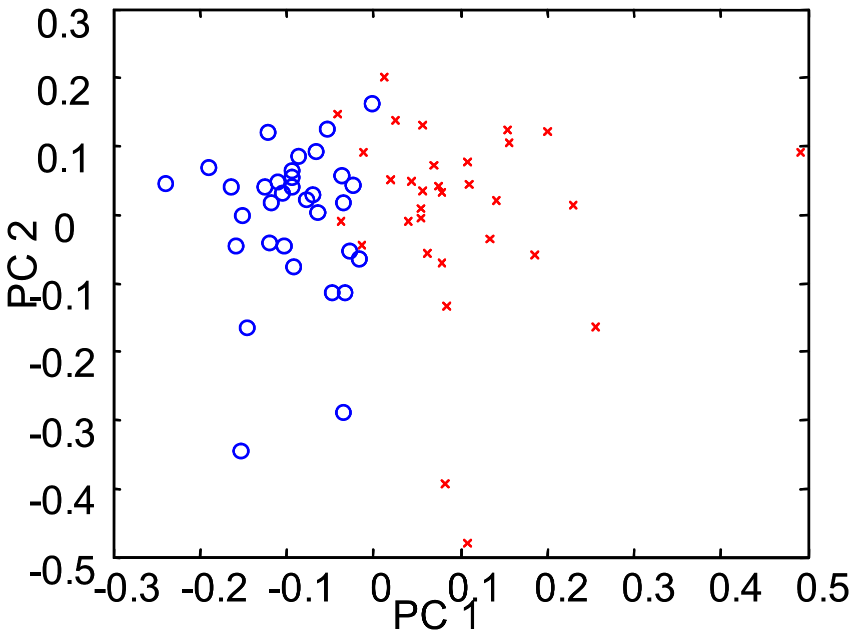
  2. Cov($\mathbf{e}_2^T\mathbf{y}, \mathbf{e}_1^T\mathbf{y}$)=$\mathbf{e}_2^T\mathbf{B}^T\mathbf{B}\mathbf{e}_1=0 \Rightarrow$ The 2nd principal component is not correlated to the 1st principal component

- **The *i*th principal component=$\mathbf{e}_i^T\mathbf{y} = e_{i1}y_{k1} + e_{i2}y_{k2} + \ldots + e_{i,53}y_{k,53}$**

- **Sum of normalized measurement variances=Trace(ρ)=Sum of λ_i=Sum of all principal components variances =53**

Back to Alcoholic Distinguishing Problem

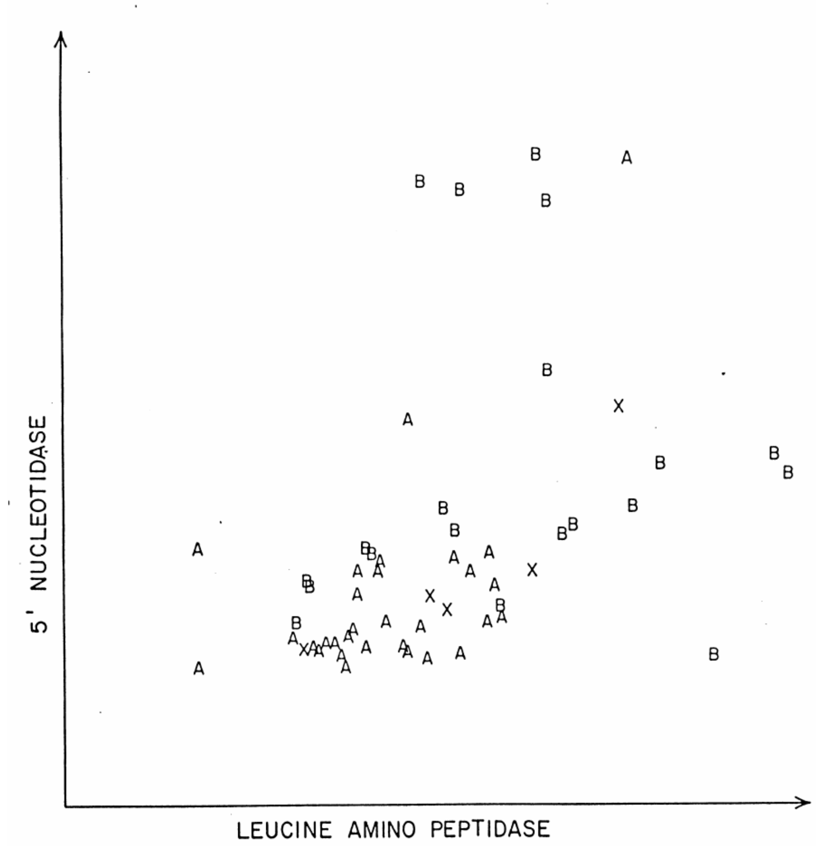
- Use the First and Second Principal Components to distinguish the alcoholics from the non-alcoholics:



Blue circles are non-alcoholics and red crosses are alcoholics

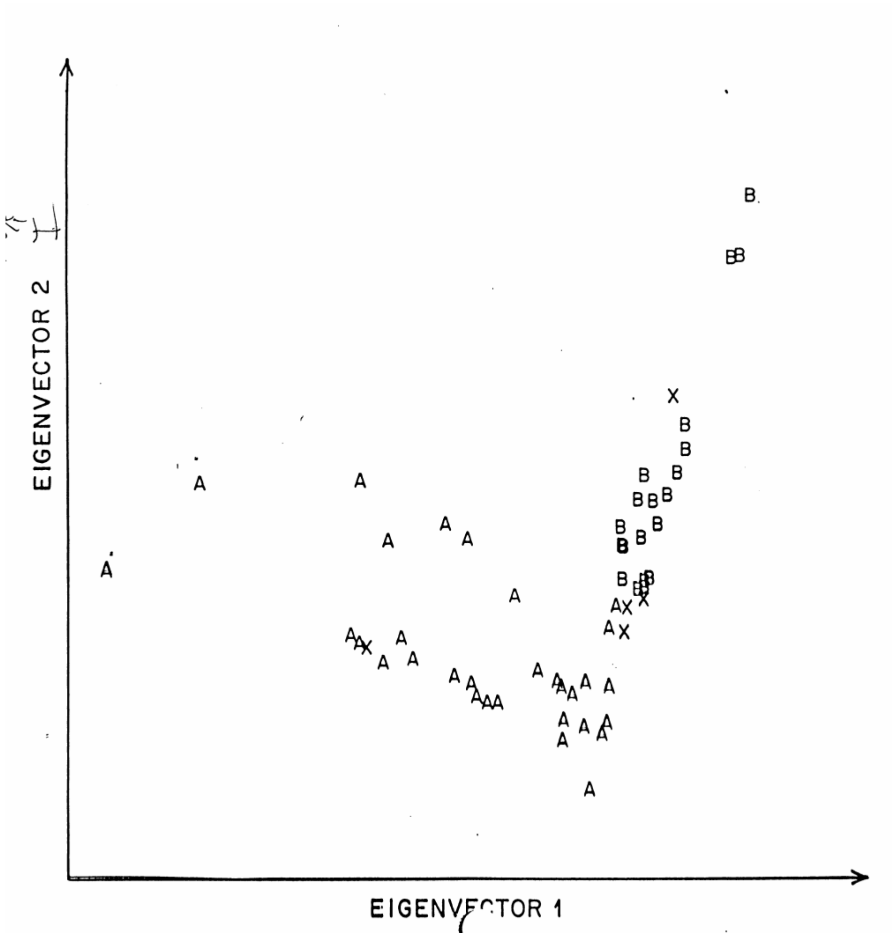
Clinical Study Example: Liver Disorders

- Two types of liver disorders: "A" and "B" ("X" normal)
- Eight enzyme concentrations in blood of patients are observed
- Conventional diagnosis: use only two enzyme concentrations to diagnose whether a patient has "A" or "B" liver disorder:



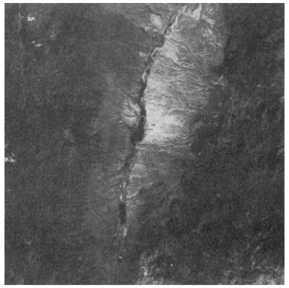
PCA of Liver Disorders

- **Diagnosis using first two principal components:**



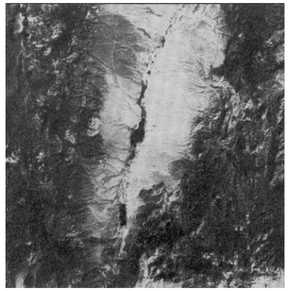
Example: Computer Vision and Remote Sensing

- **Satellite visions from three remote sensing spectrums: blue, near infrared, mid infrared rays.**



200

200 **blue**



near infrared

mid infrared

- **Vision intensity data:**

$$A = \begin{bmatrix} b_1 & i_1 & m_1 \\ b_2 & i_2 & m_2 \\ \vdots & \vdots & \vdots \\ b_{40000} & i_{40000} & m_{40000} \end{bmatrix}$$

PCA of Remote Sensing Data

- **Covariance matrix:**

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}$$

- **Eigenvalues:** $\lambda_1 = 7614.23 \geq \lambda_2 = 427.63 \geq \lambda_3 = 98.10$

- **Eigenvectors:**

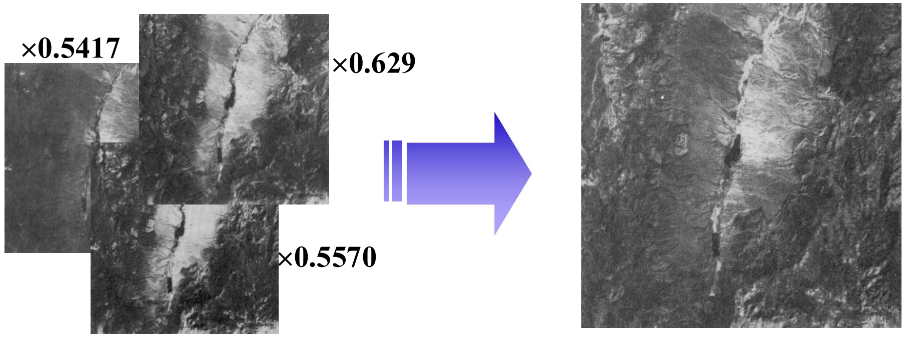
$$u_1 = \begin{bmatrix} .5417 \\ .6295 \\ .5570 \end{bmatrix}, u_2 = \begin{bmatrix} -.4894 \\ -.3026 \\ .8179 \end{bmatrix}, u_3 = \begin{bmatrix} .6834 \\ -.7157 \\ .1441 \end{bmatrix}$$

- **The first principal component:**

$$z_k = \mathbf{e}_1 \mathbf{y} = .5417 y_1 + .6295 y_2 + .5570 y_3$$

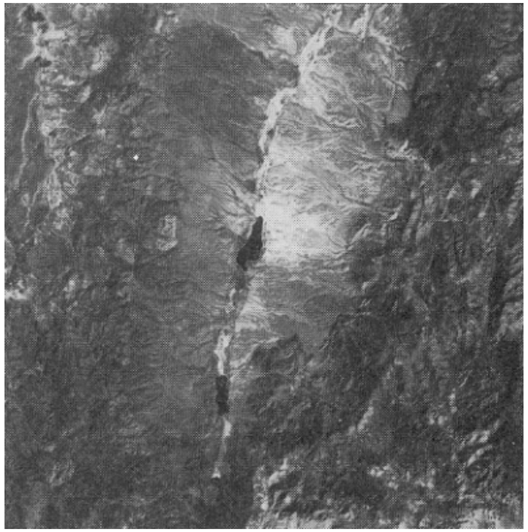
- **Every pixel is calculated by the linear combination of pixels from three spectrums**

- ***This 1st component maximizes the intensity differences among the pixels and gives us the best overall clarity in the vision.***

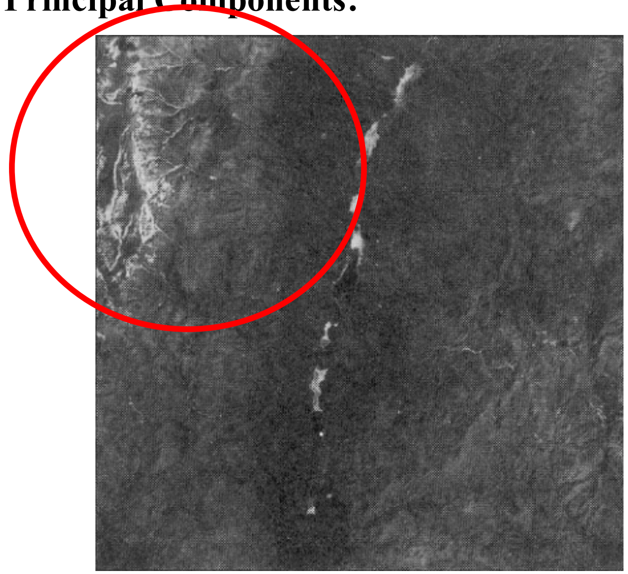


1st and 2nd PC of Remote Sensing Data

- **1st Principal Component:**



- **2nd Principal Components:**



Typical Supervised Learning Problems

	Group 1	Group 2	…	Group g
Sample	$x_{11}, x_{21}, \ldots, x_{n_1 1}$	$x_{12}, x_{22}, \ldots, x_{n_2 2}$	…	$x_{1g}, x_{2g}, \ldots, x_{n_g g}$
Mean Vector	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_g$
Overall Mean	$\bar{x}$			

where $x_{ik} = \begin{bmatrix} x_{1ik} \\ x_{2ik} \\ \vdots \\ x_{pik} \end{bmatrix}$ $\quad$ i : sample number $1,\ldots,n_k$; k : group number $1,\ldots,g$; characteristics : $1,\ldots,p$

$$\bar{x}_k = \begin{bmatrix} \frac{\sum_{i=1}^{n_k} x_{1ik}}{n_k} \\ \vdots \\ \frac{\sum_{i=1}^{n_k} x_{pik}}{n_k} \end{bmatrix} \qquad \bar{x} = \begin{bmatrix} \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k} x_{1ik}}{g n_k} \\ \vdots \\ \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k} x_{pik}}{g n_k} \end{bmatrix}$$

- **Examples: Credit card users are classified into three types: impulsive, mild and conservative spender (g=1, 2 and 3); Liver disorders have three types: Disorder A, Disorder B and Normal.**
- **We collected $n_1 = n_2 = n_3 = 100$ for each type of spenders**
- **For each spender, we observe the card user's 5 characteristics: age? income? house value? average neighborhood house value? total family income? (*p*=5)**
- **For example, for the 20th user (*i*=20) in the impulsive group (*k*=1), we have observed his characteristics:** $\mathbf{x}_{20,1} = \begin{bmatrix} x_{1,20,1} \\ \vdots \\ x_{5,20,1} \end{bmatrix}$

Discriminating Different Types of Spenders

- **Problem: find a linear combination of spending characteristics**

$$y_{ik} = b^T x_{ik} = b_1 x_{1ik} + b_2 x_{2ik} + \ldots + b_5 x_{5ik}$$

that can best discriminate three types of spenders

- **Ideas:**

  1. **$y=b^T x$ should make the difference within the same group as small as possible**
  2. **$y=b^T x$ should make the difference between different groups as large as possible**

- **Measuring the difference within the same group: *Within Group Sum of Squares Matrix***

$$W = \sum_{k=1}^{3}\sum_{i=1}^{100}(x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k)^T$$

- **Measuring the difference among different groups: *Among Group Sum of Squares Matrix***

$$G = \sum_{k=1}^{3}\sum_{i=1}^{100}(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T = \sum_{k=1}^{g} 100(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$$

Discriminant Analysis

- **The linear combination of spending characteristics $y = b^T x$:**

	Group 1	Group 2	Group 3
Sample	$b^T x_{11}, \ldots, b^T x_{100,1}$	$b^T x_{12}, \ldots, b^T x_{120,2}$	$b^T x_{13}, \ldots, b^T x_{80,3}$
Sample Mean	$b^T \bar{x}_1$	$b^T \bar{x}_2$	$b^T \bar{x}_3$
Overall Mean	$b^T \bar{x}$		

- **Sum of squares of y within groups = $b^T W b$**
- **Sum of squares of y among groups = $b^T G b$**
- **We want to minimize $b^T W b$ while maximize $b^T G b$**

  $\Rightarrow$ **Max $b^T G b \,/\, b^T W b \Rightarrow$ Max $b^T W^{-1} G b \,/\, b^T W^{-1} W b = b^T W^{-1} G b \,/\, b^T b$**

- **That is, we like to find a b such that $b^T W^{-1} G b \,/\, b^T b$ is the largest**
- **This is equivalent to a generalized eigenvalue problem:**

$$W^{-1} G b = \lambda b \Rightarrow G b = \lambda W b \quad (\textit{Rayleigh's Principle})$$

- **The largest eigenvalue will be the maximum value and the corresponding eigenvector is the best discriminant b^***
- **We can predict the spender type of a new card user with observed 5 characteristics by $b^{*T} x$**