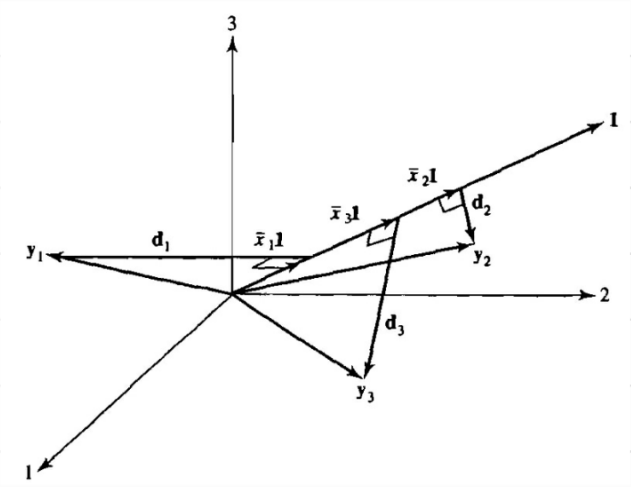
3.2 The Geometry of the Sample

The scatter plot of n points in p-dimensional space provides information on the locations and variability of the points.

- coefficient of determination (R^2) is the determinant of variance-covariance matrix.
- mean vector $\bar{x}$ is the center of the n points in p-dimension
- although graphical representation is limited to 3-dimensions, geometrical relationships and the associated statistical concepts remain valid for extended dimensions.



$\bar{x}_i$'s and corresponding d_i's are orthogonal.

Decomposition of y_i vectors

$$d_i = y_i - \bar{x}_i 1$$

$$y_i = d_i + \bar{x}_i 1$$

Let $y = [x_{1i}, x_{2i}, \dots x_{ni}]$. The projection of y_i on the unit vector $\frac{1}{\sqrt{n}}1$ is

$$y_i'\left(\frac{1}{\sqrt{n}}1\right)\frac{1}{\sqrt{n}}1 = \frac{x_{1i} + x_{2i} + \cdots x_{ni}}{n} = \bar{x}_i 1$$

The squared length of the deviation vectors equals the sum of squared deviations

- the length of deviation vector is proportional to the standard deviation.

Squared Lengths of the deviation vectors

$$L_{d_i}^2 = d_i' d_i = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2$$

$L_{d_i}^2$: (Length)^2 — $\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2$: sum of squared deviations

- note that the squared length L^2 is proportional to the variance of the measurements on the i^{th} variable. Equivalently, the length is proportional to the standard deviation. Longer vectors represent more variability than shorter vectors.

For any two deviation vectors d_i and d_k,

$d_i' d_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$, and let θ_{ik} be the angle between d_i and d_k,

$d_i' d_k = L_{d_i} L_{d_k} \cos(\theta_{ik}) = \sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2} \cdot \cos(\theta_{ik})$, then we notice that

$$\frac{d_i' d_k}{L_{d_i} L_{d_k}} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 \cdot \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}} = \cos(\theta_{ik}) = r_{ik} = Cor(i, k)$$

$\cos(0°) = 1$, $\cos(90°) = 0$, ... 등등 Cosine의 각도를 생각해보면 당연한 말...! 썸그해...!

3.3 Random Samples and the Expected Values of the Sample Mean and Covariance Matrix

- Euclidean distance appears appropriate when samples are independent and have the same variances.
- If not, the influence of individual measurements on location is asymmetric so the use of statistical distances or quadratic forms are appropriate.

* Refer to the proof at pg 142-143

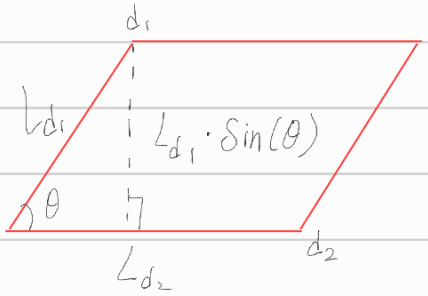
↓

Unbiased Sample Variance - Covariance Matrix

$$S = \frac{n}{n-1} S_n = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})' \qquad (E(S_n))$$

3.4 Generalized Variance

- $|S|$, the determinant of a variance - covariance matrix, is a single numerical value for the variation expressed (Generalized Variance)
- Consider the area of a trapezoid,



then, $Area = |L_{d_1} \sin(\theta)| L_{d_2}$

$= L_{d_1} L_{d_2} \sqrt{1 - \cos^2\theta}$, and

we know $L_{d_1} = \sqrt{\sum_{j=1}^{n} (x_{j1} - \bar{x})^2} = \sqrt{(n-1) S_{11}}$,

$L_{d_2} = \sqrt{\sum_{j=1}^{n} (x_{j2} - \bar{x})^2} = \sqrt{(n-1) S_{22}}$,

$\cos(\theta) = r_{12}$, so

$Area = (n-1)\sqrt{S_{11}}\sqrt{S_{22}}\sqrt{1 - r_{12}^2} = (n-1)\sqrt{S_{11} S_{22}(1 - r_{12}^2)}$, and note that

$S_{12} = S_{21} = \sqrt{S_{11}}\sqrt{S_{22}}\, r_{12}$, so b/c $|S| = S_{11} S_{22} - S_{11} S_{22} r_{12}^2 = S_{11} S_{22}(1 - r_{12}^2)$,

$$|S| = \frac{(Area)^2}{(n-1)^2} = (n-1)^{-p} (volume)^2$$

determinant

Generalized sample variance

The generalized variance with 0 is indicative of degeneracy.

- when at least one deviation vector lies in the (hyper) plane formed by linear combinations of the others $\Leftrightarrow$ linearly dependent (Column Degeneracy)

* this means that one of the deviation vectors lies in the plane generated by the other residual vectors of the others. (the volume or area is 0)

Conditions for generalized variance = 0 (either all true or all false)

1. $S\alpha = 0$, α is a scaled eigenvector of S with eigenvalue 0.
2. $\alpha'(X_j - \bar{X}) = 0$, linear combination of the mean corrected data is 0.
3. $\alpha' X_j = C$, $C = \alpha'\bar{X}$, linear combinations of the original data is a constant

- In any statistical analysis, $|S| = 0$ means that the measurements on some variables should be removed from the study as far as the mathematical computations are concerned.

Some Pre-alerts of $|S| = 0$ cases

1. $n \leq p$ ← ? covariance 성분 고려X

2. If the linear combination $\alpha' X_j$ has positive variance for each constant vector $\alpha \neq 0$, then, provided $p < n$, S has full rank with probability 1 and $|S| > 0$.
3. If, with probability 1, $\alpha' X_j$ is a constant, then $|S| = 0$.

Generalized Variance Determined by $|R|$ and its Geometrical Interpretation

- The generalized sample variance is unduly affected by the variability of measurements on a single data

Generalized sample variance of the standardized variables = $|R| = (n-1)^{-p}(\text{volume})^2$

- the generalized sample variance of the standardized variables will be large when these vectors are nearly perpendicular and will be small when two or more of these vectors are in almost the same direction

$|S| = (S_{11} S_{22} S_{33} \cdots S_{pp})|R|$

$(n-1)^p |S| = (n-1)^p (S_{11} S_{22} S_{33} \cdots S_{pp})|R|$

Another Generalization of variance

Total Sample Variance = $\sum_{i=1}^{n} S_{ii} = tr(S)$

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations

$$\bar{x} = \frac{1}{n} X' \mathbf{1}$$

$$\mathbf{1}\bar{x}' = \frac{1}{n}\mathbf{1}\mathbf{1}X = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & & & \\ \vdots & & & \\ \bar{x}_1 & \cdots & & \bar{x}_p \end{bmatrix}$$

$$X - \frac{1}{n}\mathbf{1}\mathbf{1}'X = \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & & \\ \vdots & & & \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}$$, from we can derive

$$(n-1)S = \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & & \\ \vdots & & & \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & & \\ \vdots & & & \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}$$

$$= \left(X - \frac{1}{n}\mathbf{1}\mathbf{1}'X\right)'\left(X - \frac{1}{n}\mathbf{1}\mathbf{1}'X\right) = X'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)X$$

$$= \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)$$

$$\therefore \bar{x} = \frac{1}{n}X\mathbf{1}$$

$$S = \frac{1}{n-1}X'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)X$$

Let $D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & & \\ \vdots & & & \\ 0 & \cdots & & \sqrt{s_{pp}} \end{bmatrix}$, $D^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} & & \\ \vdots & & & \\ 0 & \cdots & & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$, and

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & & & \\ s_{1p} & \cdots & & s_{pp} \end{bmatrix} = D^{\frac{1}{2}} R D^{\frac{1}{2}}$$

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{p1} \\ r_{21} & 1 & & \\ \vdots & & & \\ r_{p1} & \cdots & & \end{bmatrix} = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

3.6 Sample Values of Linear Combinations of Variables

If $C'X = C_1X_1 + C_2X_2 + \cdots + C_pX_p$,

sample mean $= C'\bar{X}$, and sample variance of $C'X = C'SC$

If $b'X = b_1X_1 + b_2X_2 + \cdots + b_pX_p$,

covariance of $C'X$ = covariance $b'X = b'SC$