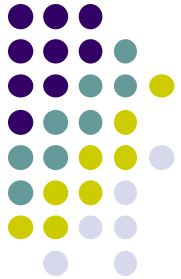
Ch 3. Sample Geometry and Random Sampling



- Assume that data are a sample of size n from a p-variate population.

  - If n observations on p different variables have been obtained, the data set can be expressed by an $n \times p$ array (matrix):

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

- Data can be plotted in two different ways:

  (1) For the p-dimensional scatterplot, the rows of X represent n points in p-dimensional space.

  (2) For the n-dimensional scatterplot, the columns of X represent p vectors in n-dimensional space.

3.2 The Geometry of the Sample

- In the ***p*-dimensional scatterplot**,

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \begin{matrix} \longleftarrow \\ \\ \\ \longleftarrow \end{matrix}$$

1^{st} (multivariate) observation

n^{th} (multivariate) observation

  - The scatterplot of n points in p-dimensional space provides information on the locations and variability of the points.
  - The sample mean vector $\bar{x}$ is the center of balance.
  - Variability is quantified by the sample variance-covariance matrix S_n.

- Example 3.1.

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

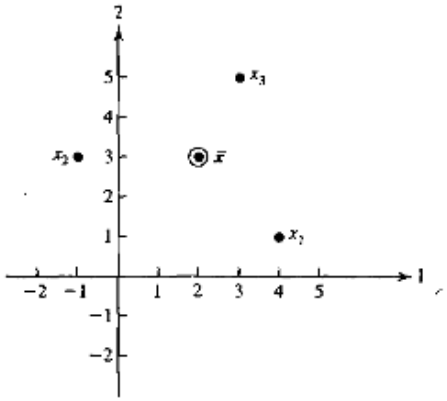


Figure 3.1 A plot of the data matrix **X** as $n = 3$ points in $p = 2$ space.

3.2 The Geometry of the Sample

- In the ***n*-dimensional scatterplot**, data can be considered as *p* vectors in *n* dimensional space.

  - Take the elements of the *columns* of the data matrix to be the coordinates of the vectors:

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_p \end{bmatrix}$$

  - In the geometrical representation, we depict $y_1,\ldots, y_p$ as vectors rather than points, as in the *p*-dimensional scatter plot.

- Example 3.2. Data as *p* vectors in *n* din

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

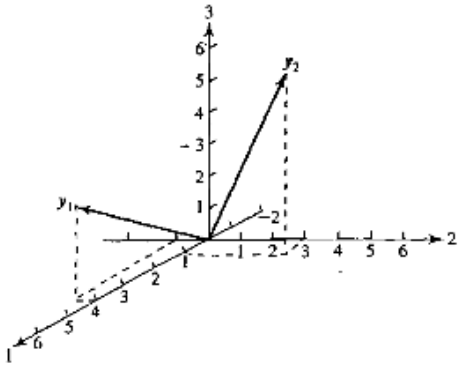
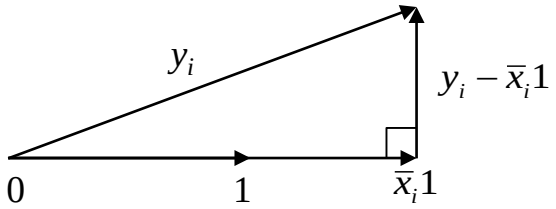


Figure 3.2 A plot of the data matrix $\mathbf{X}$ as $p = 2$ vectors in $n = 3$ space.

3.2 The Geometry of the Sample

- Define $n \times 1$ vector $1_n' = 1' = [1, 1, ..., 1]$.
  Consider the vector $y_i' = [x_{1i}, x_{2i}, ..., x_{ni}]$.
  The projection of y_i on the unit vector $\left(1/\sqrt{n}\right)1$ is

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



  - The sample mean $\bar{x}_i = \left(x_{1i} + x_{2i} + \cdots + x_{ni}\right)/n = y_i' 1/n$ corresponds to the multiple of **1** required to give the projection of y_i onto the line determined by **1**.

- The deviation, or mean corrected, vector is

$$d_i = y_i - \bar{x}_i 1 = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}.$$

  - The elements of d_i are the deviations of the measurements on the ith variable from their sample mean.

3.2 The Geometry of the Sample

- Decompose y_i into $\bar{x}_i 1$ and $d_i = y_i - \bar{x}_i 1$.

  - See Figure 3.3 (p. 115).

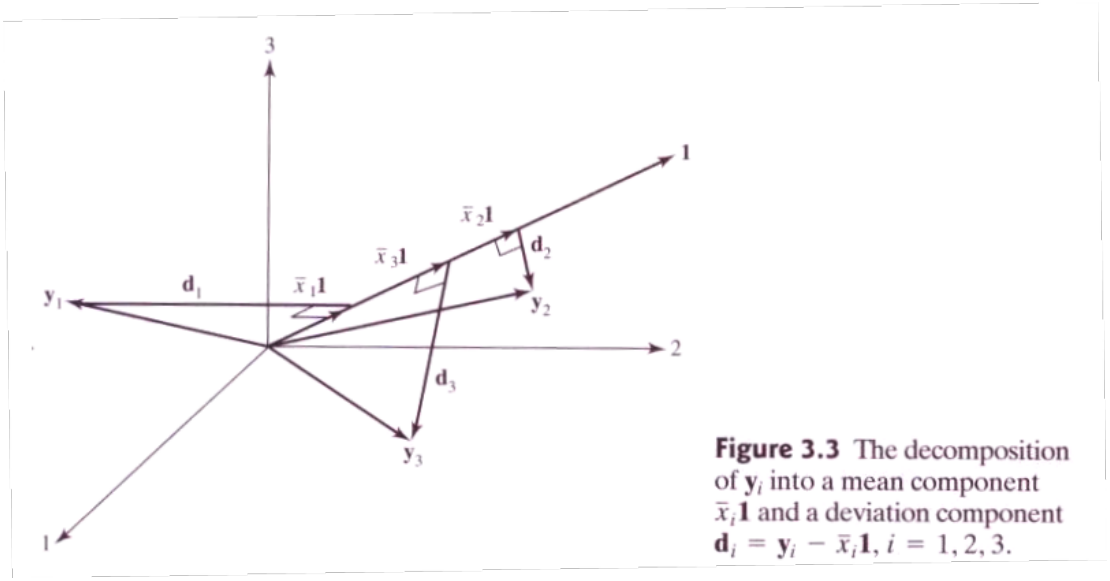
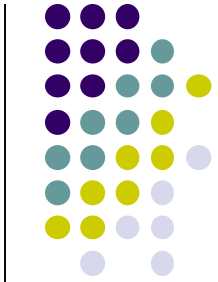


Figure 3.3 The decomposition of $\mathbf{y}_i$ into a mean component $\bar{x}_i\mathbf{1}$ and a deviation component $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$, $i = 1, 2, 3$.

3.2 The Geometry of the Sample



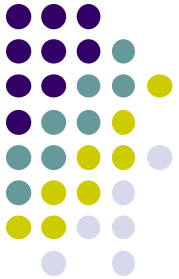
- Consider the squared lengths of the deviation vectors

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

(Length of deviation vector)^2 = sum of squared deviations

- The squared length is proportional to the variance of the measurements on the *i*th variable.
- The *length* is proportional to the standard deviation.
- Longer vectors represent more variability than shorter vectors.

3.2 The Geometry of the Sample



- For two deviation vectors d_i and d_k,

$$d_i' d_k = \sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right).$$

- $d_i' d_k = L_{d_i} L_{d_k} \cos(\theta_{ik})$,

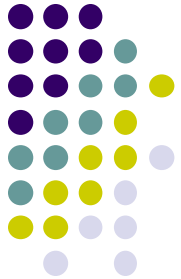
  where θ_{ik} is the angle formed by the vectors d_i and d_k.

- $$\sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) = \sqrt{\sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i\right)^2} \sqrt{\sum_{j=1}^{n} \left(x_{jk} - \bar{x}_k\right)^2} \cos(\theta_{ik})$$

$$\therefore r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik})$$

- The ***cosine*** of the angle is the sample **correlation coefficient**.
- If the two deviation vectors have nearly the same orientation, the sample correlation will be close to 1.
- If the two vectors are nearly perpendicular, the sample correlation will be approximately 0.
- If the two vectors are oriented in nearly opposite directions, the sample correlation will be close to -1.

3.2 The Geometry of the Sample



- The concept of length, angle, and projection provide geographic interpretation of the sample.

- Geometric Interpretation of the Sample

1. The projection of a column y_i of the data matrix X onto the equal angular vector 1 is the vector $\bar{x}_i 1$. The vector $\bar{x}_i 1$ has length $\sqrt{n}|\bar{x}_i|$. Therefore, the ith sample mean, $\bar{x}_i$, is related to the length of the projection of y_i on 1.
2. The information comprising S_n is obtained from the deviation vectors

$$d_i = y_i - \bar{x}_i 1 = \left[x_{1i} - \bar{x}_i, x_{2i} - \bar{x}_i, ..., x_{ni} - \bar{x}_i\right]'.$$

   The square of the length of d_i is ns_{ii}, and the (inner) product between d_i and d_k is ns_{ik}.
3. The sample correlation r_{ik} is the cosine of the angle between d_i and d_k.

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

- Consider the data

$$X_{n\times p}=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}=\begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix}.$$

If the vectors $X_1', X_2', ..., X_n'$ represent *independent* observations from a *common* joint distribution with density function $f(x) = f(x_1, x_2, \ldots, x_p)$, then $X_1, X_2, \ldots, X_n$ are said to form a *random sample* from $f(x)$.

- Mathematically, $X_1, X_2, \ldots, X_n$ form a random sample if their joint density function is given by the product $f(x_1)\, f(x_2)\cdots f(x_n)$, where $f(x_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$ is the density function for the jth row vector.

- The measurements of the p variables in a *single* trial, such as $X_j' = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, will usually be correlated. However, the measurements from *different* trials must be independent.

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

- The notion of statistical independence has important implications for measuring distance.

  - Euclidean distance is appropriate if the components of a vector are independent and have the same variances.

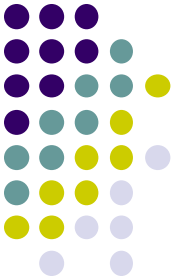
- Consider the location of the kth column as $Y_k' = [X_{1k}, X_{2k}, ..., X_{nk}]$ of X, regarded as a point in n dimensions.

  - The location of this point is determined by the joint probability distribution $f(y_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk})$.

  - When the measurements $X_{1k}, X_{2k}, \ldots, X_{nk}$ are a random sample, $f(y_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk}) = f_k(x_{1k})\, f_k(x_{2k}) \cdots f_k(x_{nk})$. Each coordinate x_{jk} contributes equally to the location through the identical marginal distributions $f_k(x_{jk})$.

  - If the n components are not independent or the marginal distributions are not identical, the influence of individual measurements (coordinates) on location is asymmetrical. Then, need to consider a distance function in which the coordinates are weighted unequally.

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



- Result 3.1.

  Let $X_1, X_2, \ldots, X_n$ be a random sample from a joint distribution that has mean vector μ and covariance matrix Σ. Then $\bar{X}$ is an unbiased estimator of μ, and its covariance matrix is $\frac{1}{n}\Sigma$.

  That is, $E(\bar{X}) = \mu$,

  $$Cov(\bar{X}) = \frac{1}{n}\Sigma.$$

  For the covariance matrix S_n, $E(S_n) = \frac{n-1}{n}\Sigma = \Sigma - \frac{1}{n}\Sigma.$

  Thus, $E\left(\frac{n}{n-1}S_n\right) = \Sigma$

  so $[n/(n\text{-}1)]\, S_n$ is an unbiased estimator of Σ, while S_n is a biased estimator with (bias) = $E(S_n) - \Sigma = -(1/n)\,\Sigma$.

- (**Unbiased**) Sample variance-covariance matrix is

  $$S = \left(\frac{n}{n-1}\right)S_n = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})'.$$

3.4 Generalized Variance

- When p variables are observed on each unit, the variation is described by the sample variance-covariance matrix

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i \right) \left(x_{jk} - \bar{x}_k \right) \right\}.$$

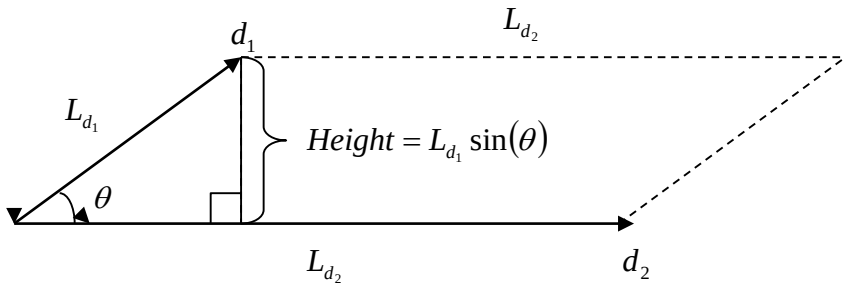
- Sometimes it is desirable to assign a single numerical value for the variation expressed by S.

  - One choice for a value is the determinant of S, which is also called the generalized sample variance:

Generalized sample variance = $|S|$.

  - When $p > 1$, some information about the sample is lost in the process.

3.4 Generalized Variance



$$\text{Area} = L_{d_1} L_{d_2} \sqrt{1-\cos^2(\theta)}.$$

Since $L_{d_1} = \sqrt{\sum_{j=1}^{n}\left(x_{j1}-\bar{x}_1\right)^2} = \sqrt{(n-1)s_{11}}$, $L_{d_2} = \sqrt{\sum_{j=1}^{n}\left(x_{j2}-\bar{x}_2\right)^2} = \sqrt{(n-1)s_{22}}$, and $\cos(\theta) = r_{12}$,

$$\text{Area} = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1-r_{12}^2} = (n-1)\sqrt{s_{11}s_{22}\left(1-r_{12}^2\right)}.$$

On the other hand,

$$|S| = \left|\begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix}\right| = \left|\begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}\,r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}\,r_{12} & s_{22} \end{bmatrix}\right|$$

$$= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 = s_{11}s_{22}\left(1-r_{12}^2\right).$$

Therefore, $|S| = (area)^2/(n-1)^2$.

Generalized sample variance = $|S| = (n-1)^{-p}(\text{volume})^2$.

3.4 Generalized Variance

- The generalized sample variance, for a fixed set of data, is proportional to the square of the volume generated by the p deviation vectors $d_1 = y_1 - \bar{x}_1 1, d_2 = y_2 - \bar{x}_2 1, \ldots, d_p = y_p - \bar{x}_p 1$.

  - See Figure 3.6 (p. 125).

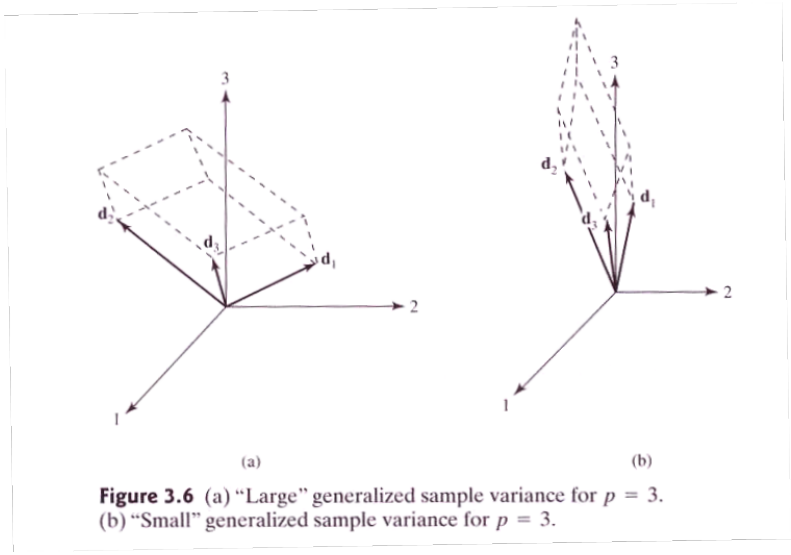


Figure 3.6 (a) "Large" generalized sample variance for $p = 3$. (b) "Small" generalized sample variance for $p = 3$.

- For a fixed sample size, the volume, or $|S|$, increases when the length of any $d_i = y_i - \bar{x}_i 1 \left(\text{or } \sqrt{s_{ii}}\right)$ increases.

  - The volume increases if the residual vectors of fixed length are moved until they are at right angles to one another, as in Figure 3.6(a).

  - The volume is small if just one of the s_{ii} is small or one of the deviation vectors lies nearly in the (hyper) plane formed by the others, or both, as in Figure 3.6(b).

3.4 Generalized Variance

- Generalized variance also has interpretation in the *p*-space scatterplot representation of the data.
- The coordinates $x' = \left[x_1, x_2, ..., x_p\right]$ of the points with a constant distance *c* from $\bar{x}$ satisfy

$$(x - \bar{x})' S^{-1} (x - \bar{x}) = c^2.$$

- It can be shown that the volume of this hyperellipsoid is related to |*S*|.
  - Volume of $\left\{x : (x - \bar{x})' S^{-1} (x - \bar{x}) \le c^2\right\} = k_p |S|^{1/2} c^p$

  (Volume of ellipsoid)^2 = (constant) (generalized sample variance).
  - A large volume corresponds to a large generalized variance.

- Example 3.8. Interpreting the generalized variance
  - All three data sets have the same $\bar{x}' = \begin{pmatrix} 2 & 1 \end{pmatrix}$ and different covariance matrix as follows:

$$S = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, r = 0.8; \quad S = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, r = 0; \quad S = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, r = -0.8.$$

  - See Figure 3.7 (p. 127).
  - Note that different correlation structures are not detected by |*S*| = 9 for all three.

3.4 Generalized Variance

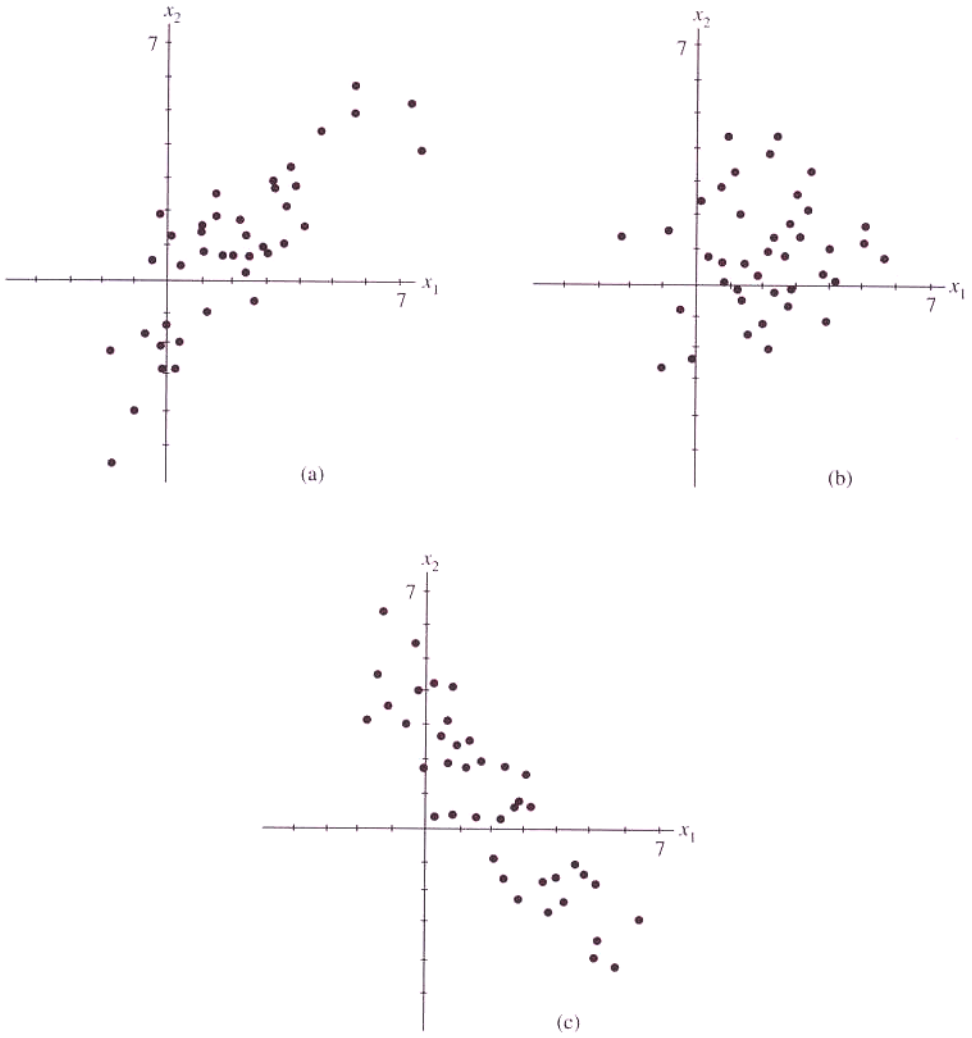
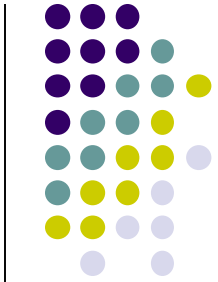


Figure 3.7 Scatter plots with three different orientations.

Situations in which the Generalized Sample Variance is Zero



- A generalized variance of zero is indicative of extreme degeneracy, in the sense that at least one column of the matrix of deviations

$$\begin{bmatrix} x_1' - \bar{x} \\ x_2' - \bar{x} \\ \vdots \\ x_n' - \bar{x}_i \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$= X_{n\times p} - 1_{n\times 1}\bar{x}'_{1\times p}$$

can be expressed as a linear combination of the other columns.

- This is a case where one of the deviation vectors, $d_i = [x_{1i} - \bar{x}_i, ..., x_{ni} - \bar{x}_i]'$, lies in the (hyper) plane generated by $d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_p$.

- $|S| = 0$ means that the measurements on some variables should be removed from the study, and the corresponding reduced data matrix will lead to a covariance matrix of full rank and a nonzero generalized variance.

Situations in which the Generalized Sample Variance is Zero

- If the columns of the mean corrected data matrix are linearly dependent,

$$\left(X - 1\bar{x}'\right)a = 0.$$

  - Whenever the columns of the mean corrected data matrix are linearly dependent,

$$(n-1)Sa = \left(X - 1\bar{x}'\right)'\left(X - 1\bar{x}'\right)a = \left(X - 1\bar{x}'\right)'0 = 0.$$

  - $Sa = \mathbf{0}$ establishes the linear dependency of the columns of S, and hence, $|S| = 0$.

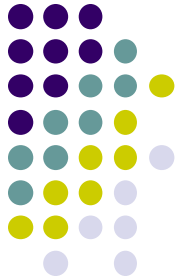
- When $Sa = \mathbf{0}$, $Sa = 0a$.

  That is, a is a scaled eigenvector of S associated with an eigenvalue of zero.

  - If we are unaware of extra variables that are linear combinations of the others, we can find them by calculating the eigenvectors of S and identifying the one associated with a zero eigenvalue.

- Result 3.3. If $n \leq p$, that is, (sample size) $\leq$ (number of variables), then $|S| = 0$ for all samples.

Generalized Variance Determined by |*R*|



- The generalized sample variance is unduly affected by the variability of measurements on a single variable.
  - If s_{ii} is large, $d_i = (y_i - \bar{x}_i 1)$ will be very long, affecting to the size of a volume.
  - It is useful to scale all the deviation vectors so that they have the same length.
- Scaling the residual vectors is equivalent to replacing each original observation x_{jk} by its standardized value $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$.
  - With the sample covariance matrix of the standardized variables, *R*,

    Generalized sample variance of the standardized variables = |*R*|

  - Since the resulting vectors

$$\left[(x_{1k} - \bar{x}_k)/\sqrt{s_{kk}}, (x_{2k} - \bar{x}_k)/\sqrt{s_{kk}}, \ldots, (x_{nk} - \bar{x}_k)/\sqrt{s_{kk}}\right] = (y_k - \bar{x}_k 1)'/\sqrt{s_{kk}}$$

    all have length $\sqrt{n-1}$, the generalized sample variance of the standardized variables is large when these vectors are nearly perpendicular and will be small when two or more of the vectors are in almost the same direction.
  - The cosine of the angle θ_{ik} between $(y_i - \bar{x}_i 1)'/\sqrt{s_{ii}}$ and $(y_k - \bar{x}_k 1)'/\sqrt{s_{kk}}$ is the sample correlation coefficient r_{ik}.
  - |*R*| is large when all the r_{ik} are nearly zero and it is small when one or more of the r_{ik} are nearly +1 or -1.

Generalized Variance Determined by $|R|$

- Let

$$\frac{(y_i - \bar{x}_i 1)}{\sqrt{s_{ii}}} = \begin{bmatrix} \frac{(x_{1i} - \bar{x}_i)}{\sqrt{s_{ii}}} \\ \frac{(x_{2i} - \bar{x}_i)}{\sqrt{s_{ii}}} \\ \vdots \\ \frac{(x_{ni} - \bar{x}_i)}{\sqrt{s_{ii}}} \end{bmatrix}, \quad i = 1,2,\ldots, p$$

be the deviation vectors of the standardized variables.

- The ith deviation vectors lie in the direction of d_i, but all have a squared length of n - 1.
- Generalized sample variance of the standardized variables

$$= |R| = (n - 1)^{-p}(\text{volume})^2.$$

- See Figure 3.10 (p.135) and compare it with Figure 3.6 (p.125).

Generalized Variance Determined by $|R|$

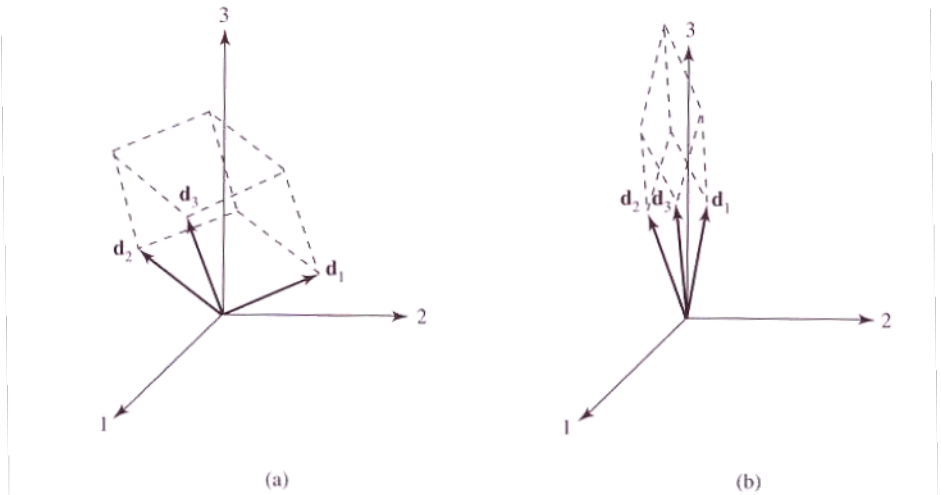


Figure 3.10 The volume generated by equal-length deviation vectors of the standardized variables.

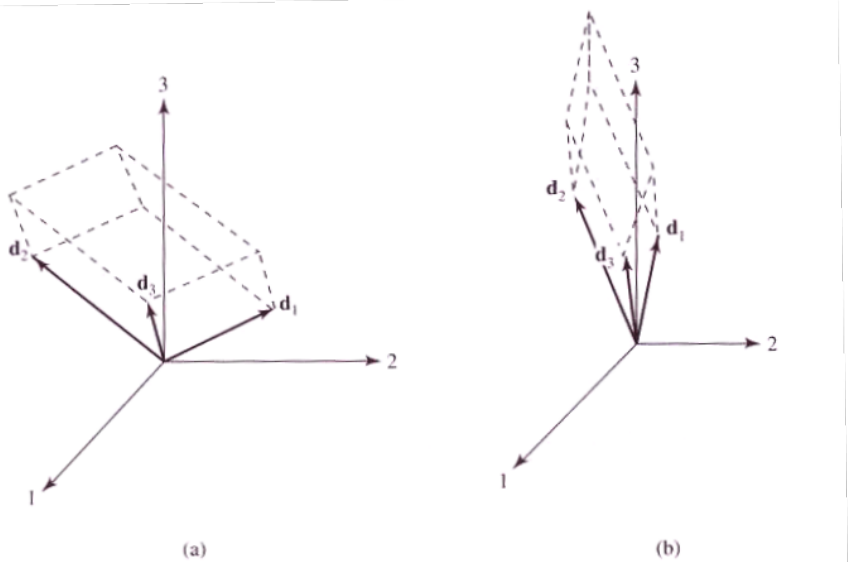


Figure 3.6 (a) "Large" generalized sample variance for $p = 3$. (b) "Small" generalized sample variance for $p = 3$.

Generalized Variance Determined by $|R|$

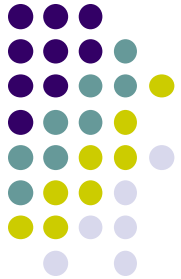
- $|S| = (s_{11}\, s_{22} \cdots s_{pp})\, |R|$

  Therefore, $(n - 1)^p\, |S| = (n - 1)^p\, (s_{11}\, s_{22} \cdots s_{pp})\, |R|$.

  - The squared volume $(n - 1)^p\, |S|$ is proportional to the squared volume $(n - 1)^p\, |R|$.
  - The constant of proportionality is the product of the variances.
  - Since $|R|$ is based on standardized measurements, it is unaffected by the change in scale. However, the relative value of $|S|$ will be changed whenever the multiplicative factor s_{11} changes.

- Example 3.11. Illustrating the relation between $|S|$ and $|R|$.

Another Generalization of Variance

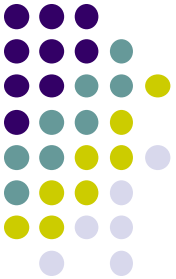


- **Total sample variance** = $s_{11} + s_{22} + \cdots + s_{pp} = tr(S)$

  - Geometrically, the total sample variance is the sum of the squared lengths of the p deviation vectors $d_1 = (y_1 - \bar{x}_1 1), \ldots, d_p = (y_p - \bar{x}_p 1)$ divided by n - 1.

  - The total sample variance criterion pays no attention to the orientation (correlation structure) of the residual vectors.

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations



- It is possible to link algebraically the calculation of $\bar{x}$ and S directly to X using matrix operations.

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \frac{y_1'1}{n} \\ \frac{y_2'1}{n} \\ \vdots \\ \frac{y_p'1}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n} X'1$$

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

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations

- Since $1\bar{x}' = \frac{1}{n}11'X = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}$,

  an $n \times p$ *matrix of deviations* (residuals) is

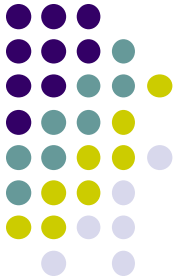
$$X - \frac{1}{n}11'X = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}.$$

- The sums of squares and cross product matrix, $(n-1)S$, is

$$(n-1)S = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}' \times \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\therefore (n-1)S = \left(X - \frac{1}{n}11'X\right)'\left(X - \frac{1}{n}11'X\right) = X'\left(I - \frac{1}{n}11'\right)X$$

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations



- Define the *p×p sample standard deviation matrix* $D^{1/2}$:

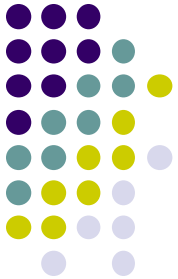
$$D^{1/2} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{s_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{s_{pp}} \end{bmatrix}$$

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

Since $$R = \begin{bmatrix} \frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}} & \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{s_{pp}}{\sqrt{s_{pp}}\sqrt{s_{pp}}} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix},$$

$R = D^{-1/2}SD^{-1/2}$ or $S = D^{1/2}RD^{1/2}$

3.6 Sample Values of Linear Combinations of Variables



- Result 3.5. The linear combinations

$$b'X = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$
$$c'X = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

have sample means, variances, and covariances that are related to $\bar{x}$ and S by

Sample mean of $b'X = b'\bar{x}$

Sample mean of $c'X = c'\bar{x}$

Sample variance of $b'X = b'Sb$

Sample variance of $c'X = c'Sc$

Sample covariance of $b'X$ and $c'X = b'Sc$.

- Consider the q linear combinations $a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$, $i = 1,2,\ldots, q$:

$$\begin{bmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + a_{q2}X_2 + \cdots + a_{qp}X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = AX$$

- Result 3.6. The q linear combinations AX have sample mean vector $A\bar{x}$ and sample covariance matrix ASA'.