Topic Three:
Sample Geometry and Random Sampling

Preview

- Motivation:
  - Geometrical representation makes the data easier to visualize and understand.
  - While there are many ways to obtain data, *random sampling* affords many advantages.
- Goals:
  - Demonstrate how multivariate data and operations on it can be viewed geometrically.
  - Define random sampling in the context of multivariate data and explore its implications.

Sample Geometry

A multivariate data set $\underset{(n\times p)}{\mathbf{X}}$ can be viewed in two ways:

- The rows $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ are a sample of size n from a p-variate population:
  - We depict the $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ as solid spheres in p dimensions (rather than vectors).
  - The sample mean vector $\bar{\mathbf{x}}$ is the center of balance.
  - The sample variance-covariance matrix quantifies the variability in all p dimensions.
- The columns $\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_p'$ are p n-variate vectors.
  - We depict the $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$ as vectors (rather than solid spheres).
  - Certain sample statistics and calculations can be handily represented using vector operations on the $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$.

Example: Consider

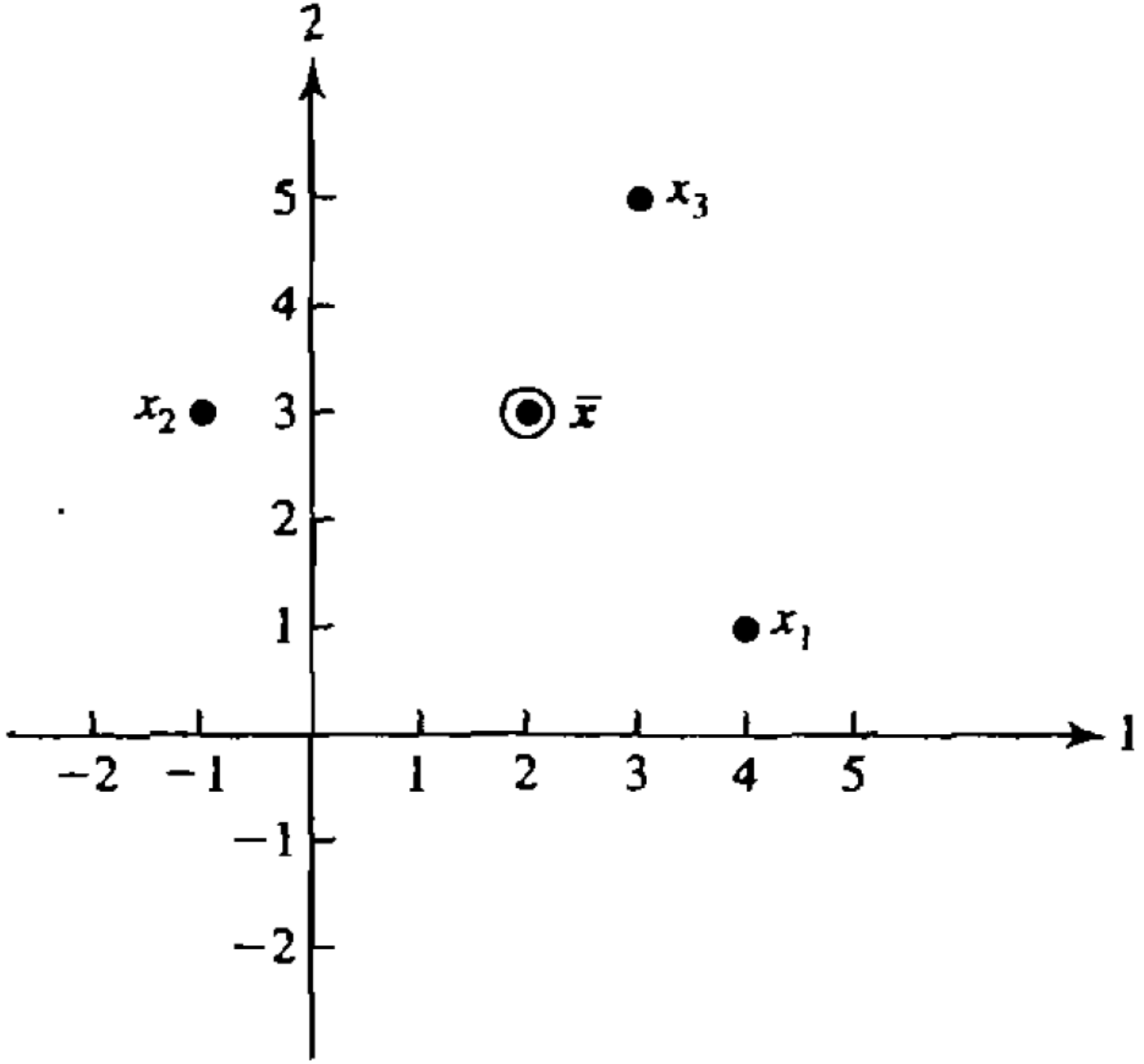
$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

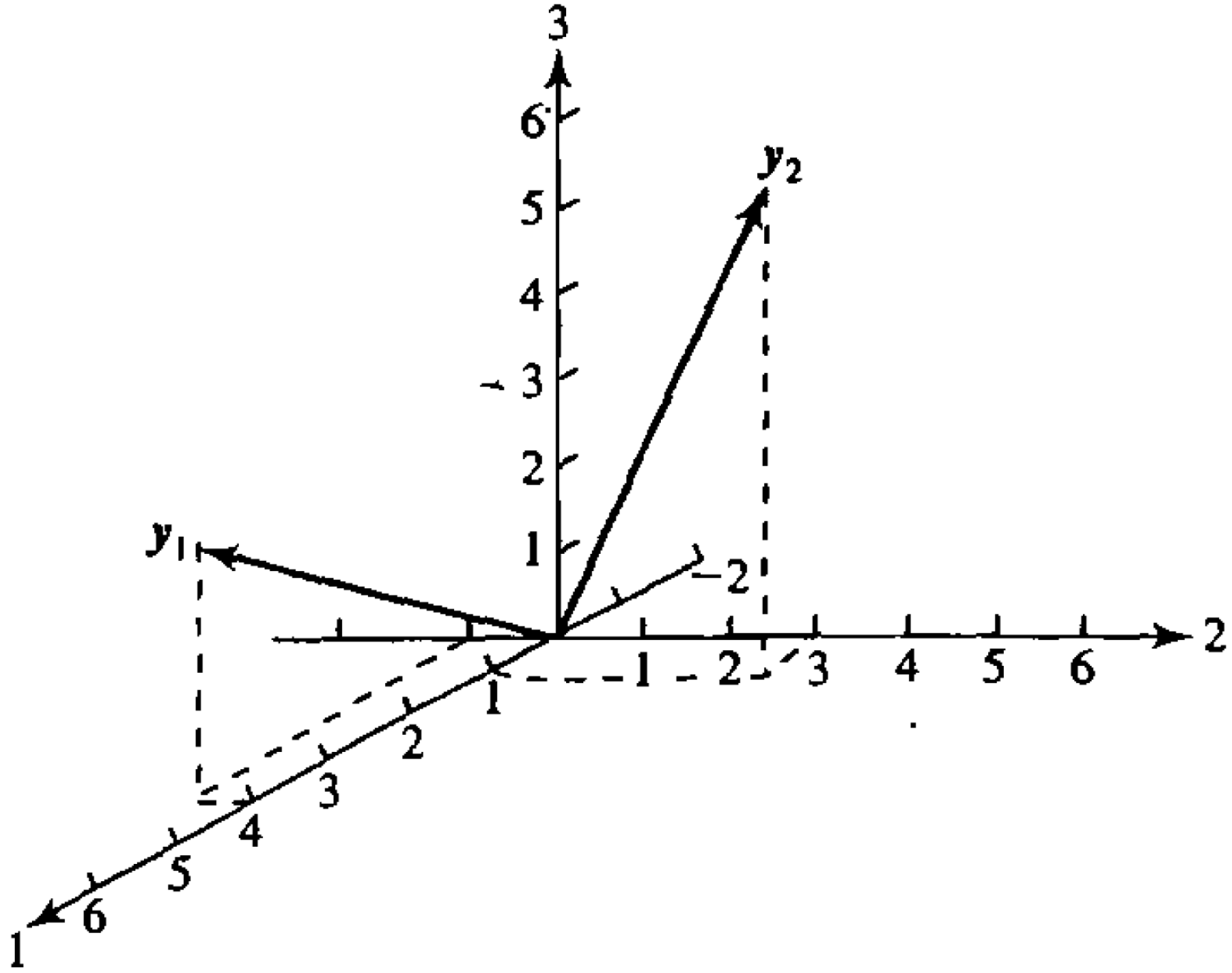
The row vectors are

$$\mathbf{x}_1 = \begin{bmatrix} 4 \\ 1 \end{bmatrix} \qquad \mathbf{x}_2 = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \qquad \mathbf{x}_3 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

Then $\bar{\mathbf{x}}' = [2, 3]$. And the column vectors are

$$\mathbf{y}_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} \qquad \mathbf{y}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$$

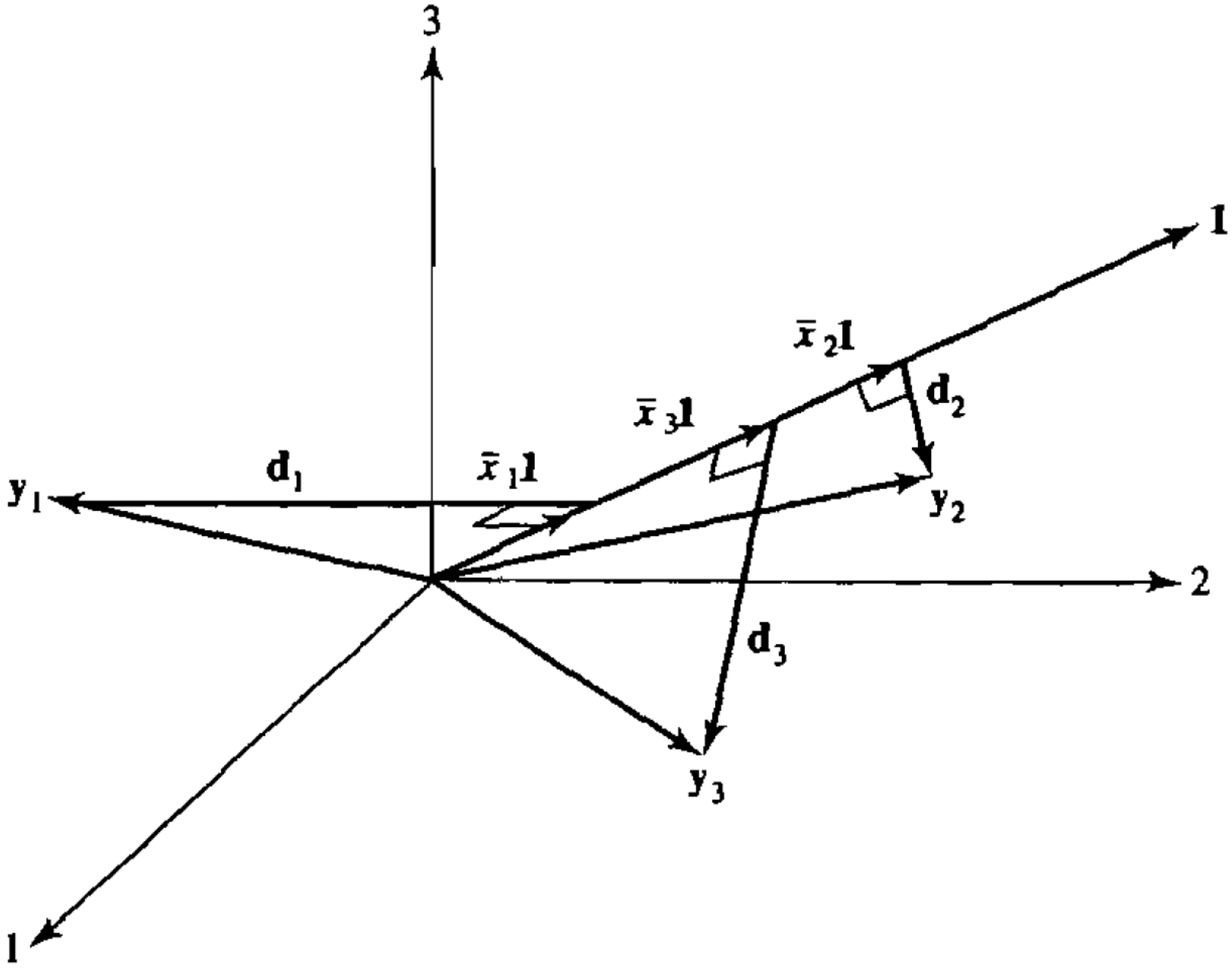

2
5
4
3
2
1
−1
−2
−2
−1
1
2
3
4
5
1
x_3
x_2
$\bar{x}$
x_1


3
6
5
4
3
2
1
y_2
y_1
2
1
2
3
4
5
6
1
2
3
4
5
6
1
2

The One Vector and Deviation Vectors: Let $\mathbf{1}_n' = [1, 1, \ldots, 1]$ (the n-vector of ones); when possible, we will just use $\mathbf{1}$ instead of $\mathbf{1}_n$. Note that $(1/n)\mathbf{1}'\mathbf{1} = 1$, so $(1/\sqrt{n})\mathbf{1}$ has unit length. Then, the projection of $\mathbf{y}_i$ on the unit vector $(1/\sqrt{n})\mathbf{1}$ is

$$\left(\frac{1}{n}\right)\mathbf{y}_i'\mathbf{1}\mathbf{1} = \frac{\mathbf{y}_i'\mathbf{1}}{\mathbf{1}'\mathbf{1}}\mathbf{1} = \bar{x}_i\mathbf{1}$$

This is the n-vector with every element equal to the sample mean of the ith variable. Thus, $\bar{x}_i$ is the multiple of $\mathbf{1}$ corresponding to the projection of $\mathbf{y}_i$ onto $\mathbf{1}$. Also, we can define the *deviation vector* $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$, so that $\mathbf{y}_i = \bar{x}_i\mathbf{1} + \mathbf{d}_i$ and $\bar{x}_i\mathbf{1}$ is perpendicular to $\mathbf{d}_i$.


3
1
$\bar{x}_2\mathbf{1}$
$\mathbf{d}_2$
$\bar{x}_3\mathbf{1}$
$\mathbf{d}_1$
$\bar{x}_1\mathbf{1}$
$\mathbf{y}_1$
$\mathbf{y}_2$
2
$\mathbf{d}_3$
$\mathbf{y}_3$
1

Example Revisited: Again let

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

with $\bar{\mathbf{x}}' = [2, 3]$. Then

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1} = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{d}_2 = \mathbf{y}_1 - \bar{x}_2\mathbf{1} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

We can then verify the decomposition

$$\mathbf{y}_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\mathbf{y}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

Deviation Vectors and Angles: Consider the squared lengths of the deviation vectors:

$$L^2_{\mathbf{d}_i} = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2$$

so that the square length is proportional to the sample variance of the ith variable, and the length is proportional to the standard deviation. Meanwhile, for any two deviation vectors $\mathbf{d}_i$ and $\mathbf{d}_k$,

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

so that the inner product of $\mathbf{d}_i$ and $\mathbf{d}_k$ is proportional to their covariance. Now let θ_{ik} be the angle between $\mathbf{d}_i$ and $\mathbf{d}_k$. We know that

$$\cos(\theta_{ik}) = \frac{\mathbf{d}_i'\mathbf{d}_k}{\sqrt{\mathbf{d}_i'\mathbf{d}_i}\sqrt{\mathbf{d}_k'\mathbf{d}_k}}$$

Since the proportionality constants cancel in the ratio, we have $\cos(\theta_{ik}) = r_{ik}$. The cosine of the angle equals the sample correlation.

Example Revisited: Recall that $\mathbf{d}_1' = [2, -3, 1]$ and $\mathbf{d}_2' = [-2, 0, 2]$. We now have

$$s_{11} = \left(\frac{1}{3}\right)\mathbf{d}_1'\mathbf{d}_1 = \left(\frac{1}{3}\right)\begin{bmatrix} 2 & -3 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = \frac{14}{3}$$

Similarly, $s_{22} = 8/3$ and $s_{12} = -2/3$, so that

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-\frac{2}{3}}{\sqrt{\frac{14}{3}}\sqrt{\frac{8}{3}}} = -0.189$$

This gives

$$\mathbf{S}_n = \begin{bmatrix} \frac{14}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{8}{3} \end{bmatrix} \qquad \text{and} \qquad \mathbf{R} = \begin{bmatrix} 1 & -0.189 \\ -0.189 & 1 \end{bmatrix}$$

Example Revisited: Recall that $\mathbf{d}_1' = [2, -3, 1]$ and $\mathbf{d}_2' = [-2, 0, 2]$. We now have

$$s_{11} = \left(\frac{1}{3}\right)\mathbf{d}_1'\mathbf{d}_1 = \left(\frac{1}{3}\right)\begin{bmatrix} 2 & -3 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = \frac{14}{3}$$

Similarly, $s_{22} = 8/3$ and $s_{12} = -2/3$, so that

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-\frac{2}{3}}{\sqrt{\frac{14}{3}}\sqrt{\frac{8}{3}}} = -0.189$$

This gives

$$\mathbf{S}_n = \begin{bmatrix} \frac{14}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{8}{3} \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} 1 & -0.189 \\ -0.189 & 1 \end{bmatrix}$$

Random Samples

Random Samples: A univariate collection of observations $x_1, x_2, \ldots, x_n$, drawn from a common distribution $f(x)$, is said to be a *random sample* if the observations are independent of one another; that is, if their joint density can be factored into the product $f(x_1)f(x_2)\cdots f(x_n)$. Similarly, for the matrix $\underset{(n\times p)}{\mathbf{X}}$ with row vectors $\mathbf{X}_j' = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, $j = 1, 2, \ldots, n$, the $\mathbf{X}_j$ form a random sample if they are independent observations from a common joint density; that is, if their joint density can be factored into the product $f(\mathbf{x}_1)f(\mathbf{x}_2)\cdots f(\mathbf{x}_n)$. Note that a random sample does not mean that the p variables for a single item are uncorrelated.

A couple of examples:

1. A random number generator is used to sample n devices from the latest batch produced by a manufacturing facility. For each device, four measures of quality were obtained. This is a random sample from the population of all devices in the latest batch.

2. For each of n children who participated in an experimental preschool program, measures of academic progress and social competence are obtained biannually. Because measurements on the same child at different times will be correlated, this is not a random sample.

Properties of $\bar{\mathbf{X}}$: Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We have

$$E(\bar{\mathbf{X}}) = \boldsymbol{\mu}$$

$$\text{Cov}(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$$

so the sample mean $\bar{\mathbf{X}}$ is an *unbiased* estimator of the population mean $\boldsymbol{\mu}$, and its variability is less than that of the population. Also,

$$E(\mathbf{S}_n) = \frac{n-1}{n}\boldsymbol{\Sigma}$$

so $\mathbf{S}_n$ is actually a *biased* estimator of the population covariance matrix $\boldsymbol{\Sigma}$. We can define $\mathbf{S} = ((n-1)/n)\mathbf{S}_n$ as an unbiased alternative. This is what is typically used in statistical analysis.

The proof that $E(\bar{\mathbf{X}}) = \boldsymbol{\mu}$ is simple. To prove that $\text{Cov}(\bar{\mathbf{X}}) = \boldsymbol{\Sigma}/n$, note that

$$\begin{aligned}
\text{Cov}(\bar{\mathbf{X}}) &= E(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})' \\
&= \frac{1}{n^2} E\left(\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})\right)\left(\sum_{l=1}^{n}(\mathbf{X}_l - \boldsymbol{\mu})'\right) \\
&= \frac{1}{n^2} E\left(\sum_{j=1}^{n}\sum_{l=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_l - \boldsymbol{\mu})'\right) \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n}\sum_{l=1}^{n} E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_l - \boldsymbol{\mu})'\right) \\
&= \frac{1}{n^2}\left(\sum_{j=1}^{n} E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'\right)
\end{aligned}$$

where the last equality follows from the assumption the $\mathbf{X}_j$ constitute a random sample and are hence independent. Now, since $\Sigma = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$ for any $\mathbf{X}$, we have

$$\text{Cov}(\bar{\mathbf{X}}) = \frac{n}{n^2}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}/n$$

To prove that $E(\mathbf{S}_n) = ((n-1)/n)\boldsymbol{\Sigma}$, first note that $(X_{ji} - \bar{X}_i)(X_{jk} - \bar{X}_k)$ is the (i,k)th element of $(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'$. Then

$$\begin{aligned}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})' &= \sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})\mathbf{X}_j' - \left(\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})\right)\bar{\mathbf{X}}' \\ &= \sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\bar{\mathbf{X}}\bar{\mathbf{X}}\end{aligned}$$

The expected value is then

$$E\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\bar{\mathbf{X}}\bar{\mathbf{X}}'\right) = \sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j') - nE(\bar{\mathbf{X}}\bar{\mathbf{X}}')$$

It can be shown (see textbook) that for any random vector $\mathbf{V}$ with mean $\boldsymbol{\mu}_V$ and covariance $\boldsymbol{\Sigma}_V$, $E(\mathbf{V}\mathbf{V}') = \boldsymbol{\Sigma}_v + \boldsymbol{\mu}_v\boldsymbol{\mu}_v'$. So,

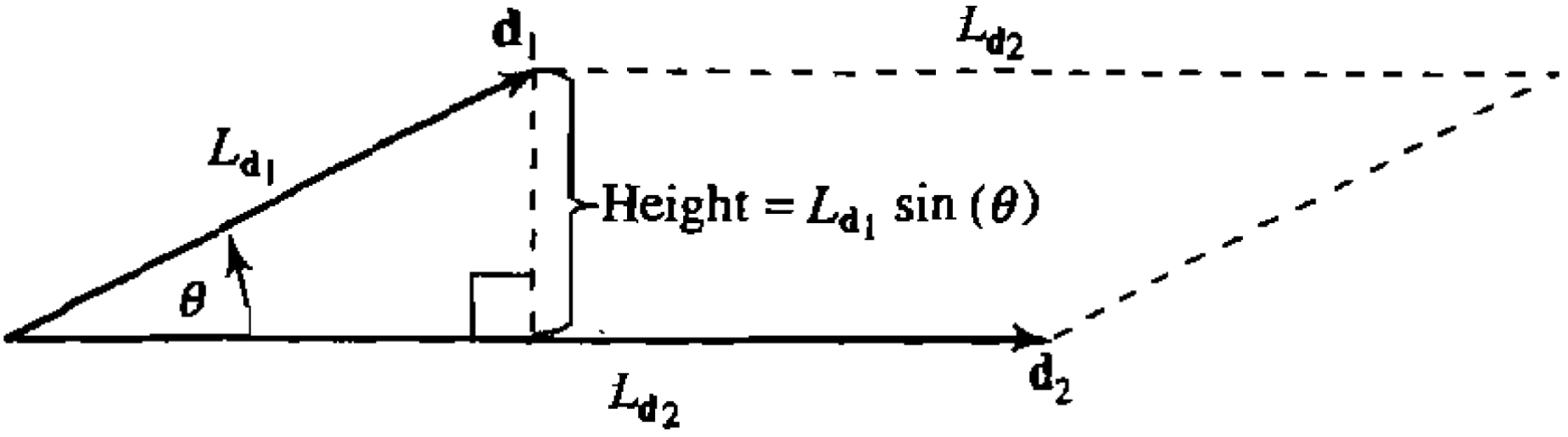
$$\sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j') - nE(\bar{\mathbf{X}}\bar{\mathbf{X}}') = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}' - n\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right) = (n-1)\boldsymbol{\Sigma}$$

and therefore

$$\begin{aligned}E(\mathbf{S}_n) &= E\left\{\frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})'\right\} = \frac{1}{n}\left\{\sum_{j=1}^{n}E(\mathbf{X}_j\mathbf{X}_j') - nE(\bar{\mathbf{X}}\bar{\mathbf{X}}')\right\} \\ &= \frac{n-1}{n}\boldsymbol{\Sigma}\end{aligned}$$

Generalized Variance

Consider the bivariate ($p = 2$) scenario, and let $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$, $i = 1, 2$, be the deviation vectors. The figure demonstrates things geometrically.



Generalized Sample Variance: Continuing from the previous slide, we have that the area of the trapezoid formed by the $\mathbf{d}_i$, $i = 1, 2$, equals $|L_{\mathbf{d}_1} \sin(\theta)| L_{\mathbf{d}_2}$. We can rewrite this as $L_{\mathbf{d}_1} L_{\mathbf{d}_2} \sqrt{1 - \cos^2(\theta)}$. Furthermore, we know that $L_{\mathbf{d}_i} = \sqrt{(n-1)s_{ii}}$, $i = 1, 2$, and $\cos(\theta) = r_{12}$. So

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

Now note that the determinant of $\mathbf{S}$ equals

$$\begin{aligned} |\mathbf{S}| &= \left| \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} \right| = \left| \begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}} r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}} r_{12} & s_{22} \end{bmatrix} \right| \\ &= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 = s_{11}s_{22}(1 - r_{12}^2) \end{aligned}$$

so that

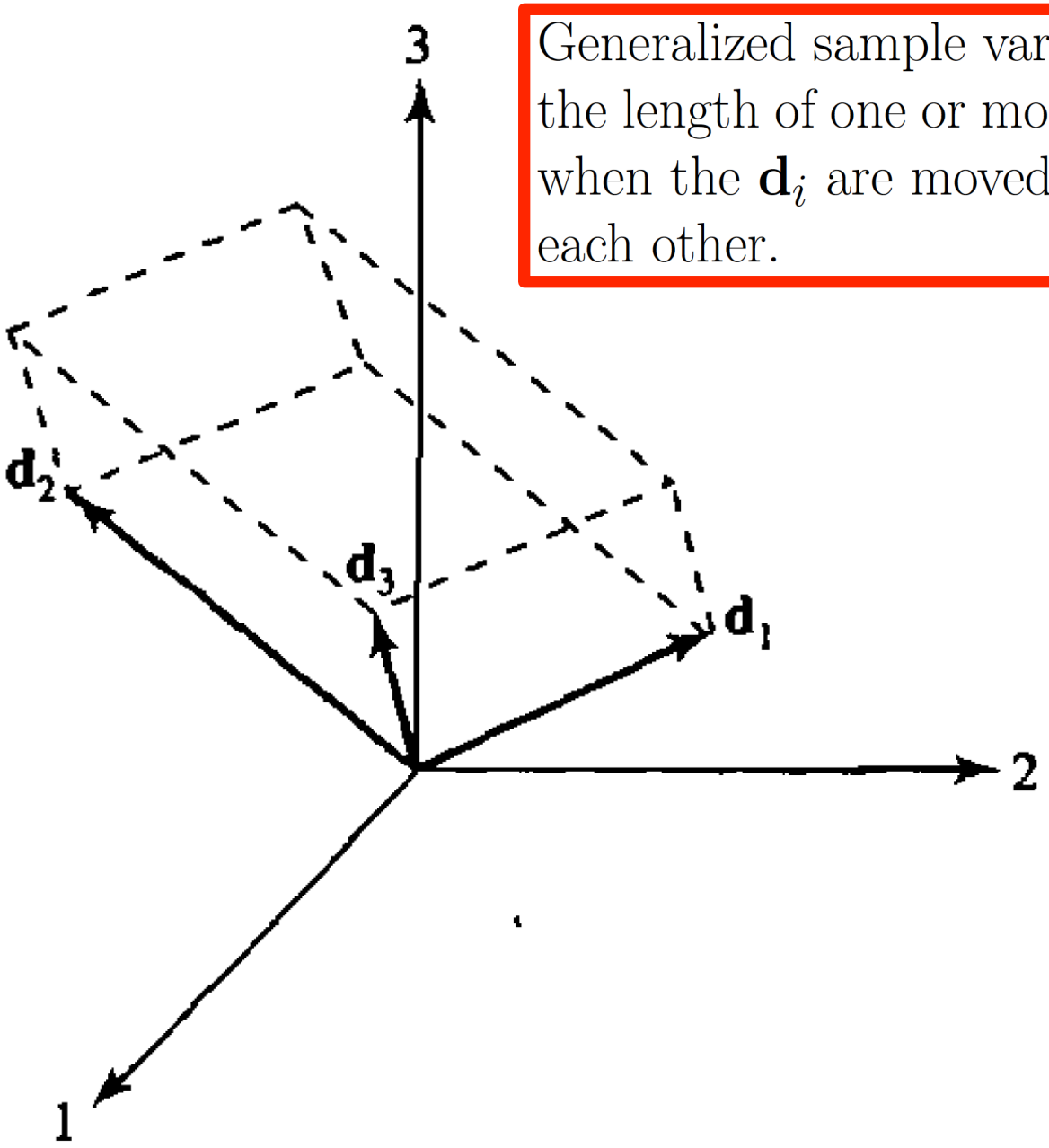
$$|\mathbf{S}| = (\text{area})^2 / (n-1)^2$$

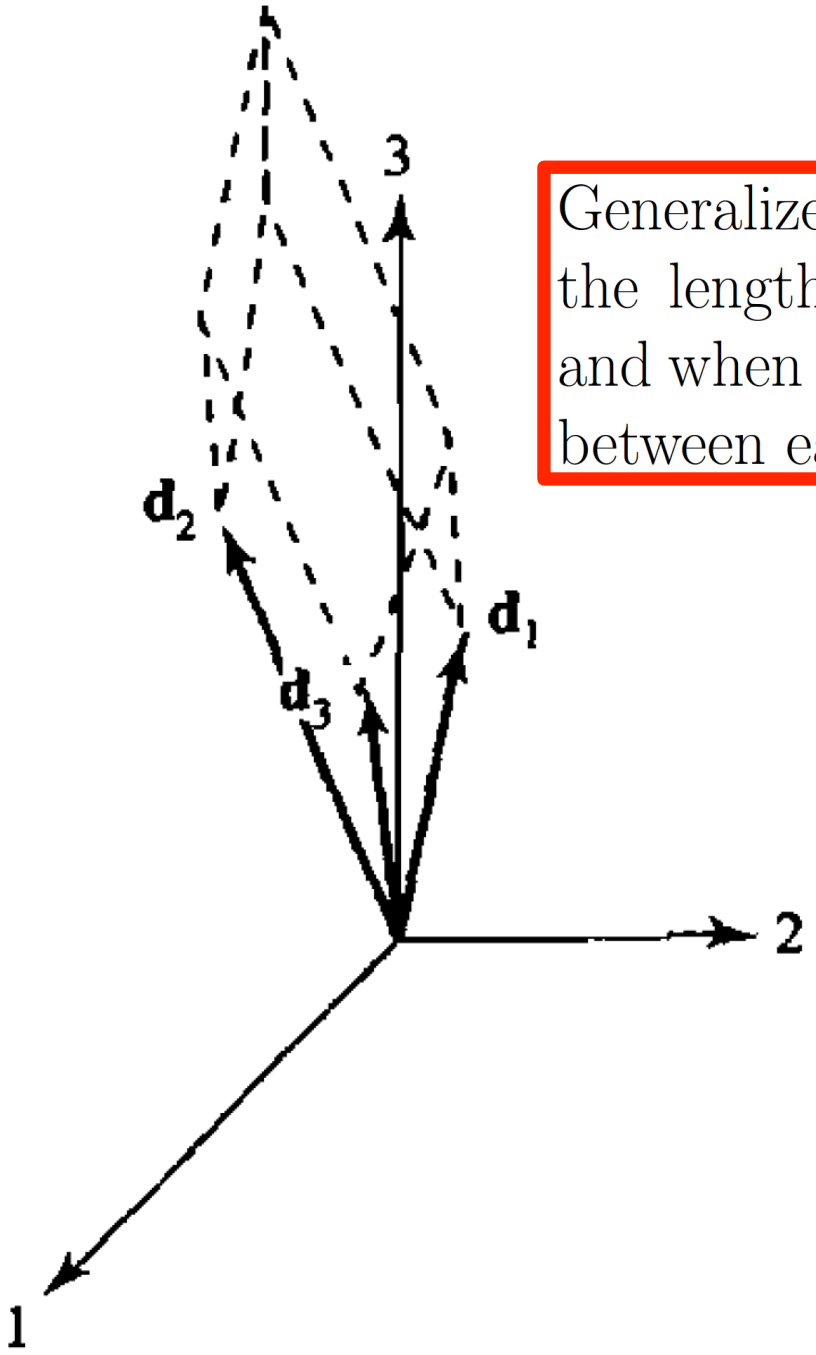
More generally, in p dimensions, it can be shown that

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

We call $|\mathbf{S}|$ the *generalized sample variance* and use it as a single number for quantifying the variability in a multivariate data set.

Generalized sample variance will increase when the length of one or more of the $\mathbf{d}_i$ increase and when the $\mathbf{d}_i$ are moved to be at right angles to each other.





Generalized sample variance will decrease when the length of one or more of the $\mathbf{d}_i$ decrease and when the $\mathbf{d}_i$ are moved to have small angles between each other.

Recall that distances can be expressed as quadratic forms. And if $\mathbf{A}$ is a positive definite matrix, then the coordinates $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ of the points a constant distance c from a fixed point $\boldsymbol{\mu}$ satisfy

$$(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

Consider $\bar{\mathbf{x}}$ as the fixed point (instead of $\boldsymbol{\mu}$), and use $\mathbf{S}^{-1}$ in place of $\mathbf{A}$. Then the vectors a fixed distance c from $\bar{\mathbf{x}}$ satisfy

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})$$

This is a hyperellipsoid centered at $\bar{\mathbf{x}}$. It can be shown that

$$\text{Volume of } \left\{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq c^2\right\} \propto |\mathbf{S}|^{1/2}c^p$$

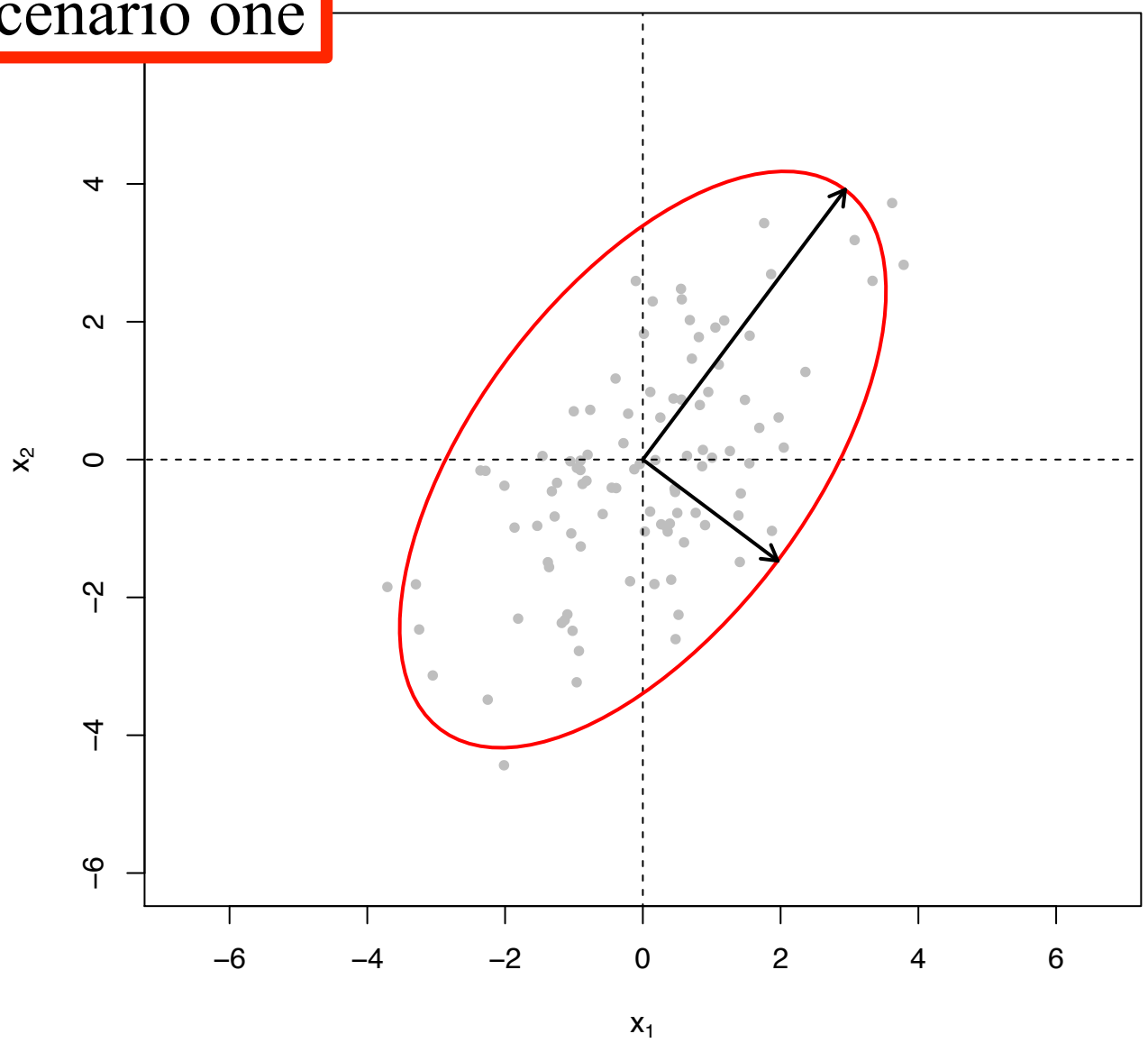
In other words, the square of the volume of the above ellipsoid is proportional to the generalized sample variance.

Example (Part 1): The generalized variance is limited in that it attempts to represent a $p \times p$ covariance matrix with a single number. In particular, it is possible to have covariance matrices that differ greatly in their correlation structure but have equal generalized variance. As an example, consider the following three sample covariance matrices

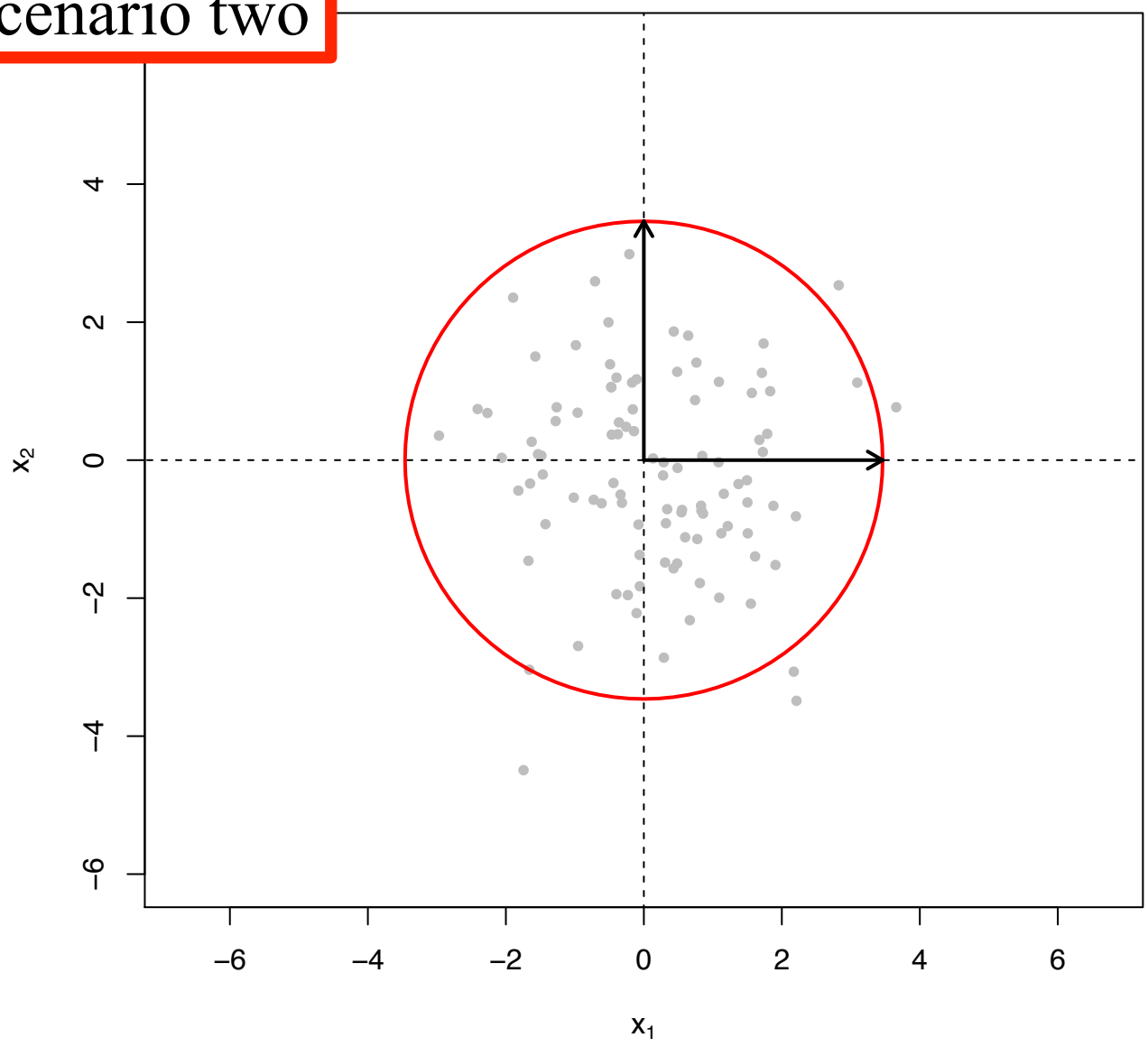
$$\mathbf{S} = \begin{bmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 2.5 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}$$

In the first scenario, the two variables are positively correlated ($r = 0.6$). In the second scenario, the variables are uncorrelated ($r = 0$). And in the third scenario, the variables are negatively correlated ($r = -0.6$). Thus, the covariances are very different with respect to correlation structure. However, the generalized variances for all three scenarios equal $|\mathbf{S}| = 4$. The scenarios are illustrated in the following figures.

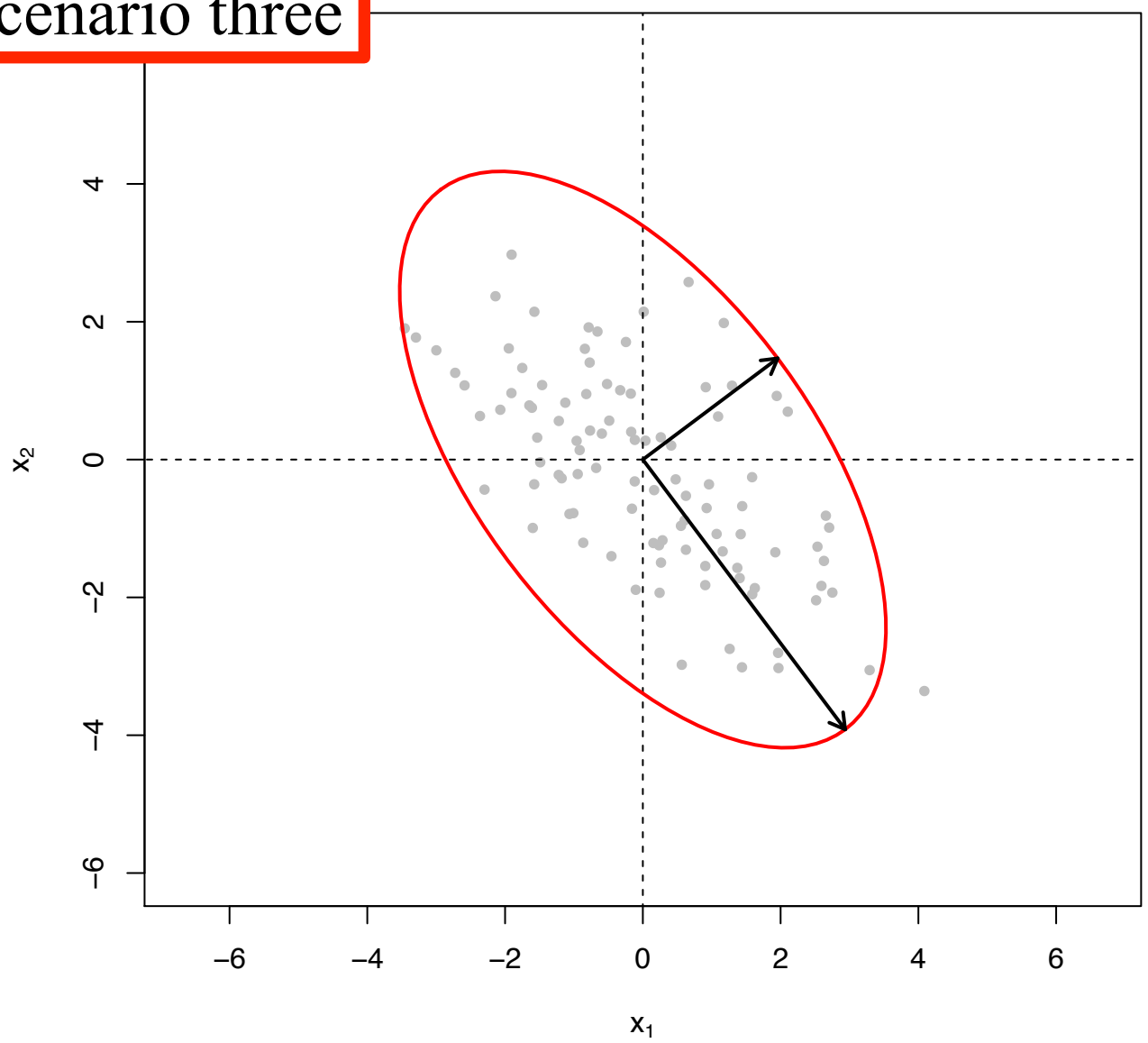
Scenario one



Scenario two



Scenario three



Example (Part 2): In the preceding figures, the red curves are ellipses corresponding to points with a constant distance c from the origin (the mean is $\mathbf{0}$ in each scenario): $(\mathbf{x}-\mathbf{0})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{0}) \leq c^2$. The angle of rotation of the ellipse θ is such that $\cos(\theta) = r$, the correlation coefficient.

Now consider one of the $\mathbf{S}$, with eigenvalues λ_1 and λ_2 and normalized eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$. We have seen that $\mathbf{S}^{-1}$ has eigenvalues $1/\lambda_1$ and $1/\lambda_2$ and the same normalized eigenvectors as $\mathbf{S}$. We also know that the points $\mathbf{x}$ for which the above distance from the origin *equals* c satisfy

$$\mathbf{x}'\mathbf{S}^{-1}\mathbf{x} = \frac{1}{\lambda_1}(\mathbf{x}'\mathbf{e}_1)^2 + \frac{1}{\lambda_2}(\mathbf{x}'\mathbf{e}_2)^2 = c^2$$

from which we derived that the region of constant distance is an ellipse with axes equal to $\mathbf{e}_1$ and $\mathbf{e}_2$ and half-lengths $c\sqrt{\lambda_1}$ and $c\sqrt{\lambda_2}$.

Finally, it turns out (we'll see why soon) that the choice $c^2 = 5.99$ corresponds to an ellipse within which we expect 95% of observations to lie. In the preceding figures, the red ellipses correspond to these inner 95% regions, and the vectors are the scaled eigenvalues.

Example (Part 2): In the preceding figures, the red curves are ellipses corresponding to points with a constant distance c from the origin (the mean is $\mathbf{0}$ in each scenario): $(\mathbf{x}-\mathbf{0})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{0}) \leq c^2$. The angle of rotation of the ellipse θ is such that $\cos(\theta) = r$, the correlation coefficient.

Now consider one of the $\mathbf{S}$, with eigenvalues λ_1 and λ_2 and normalized eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$. We have seen that $\mathbf{S}^{-1}$ has eigenvalues $1/\lambda_1$ and $1/\lambda_2$ and the same normalized eigenvectors as $\mathbf{S}$. We also know that the points $\mathbf{x}$ for which the above distance from the origin *equals* c satisfy

$$\mathbf{x}'\mathbf{S}^{-1}\mathbf{x} = \frac{1}{\lambda_1}\left(\mathbf{x}'\mathbf{e}_1\right)^2 + \frac{1}{\lambda_2}\left(\mathbf{x}'\mathbf{e}_2\right)^2 = c^2$$

from which we derived that the region of constant distance is an ellipse with axes equal to $\mathbf{e}_1$ and $\mathbf{e}_2$ and half-lengths $c\sqrt{\lambda_1}$ and $c\sqrt{\lambda_2}$.

Finally, it turns out (we'll see why soon) that the choice $c^2 = 5.99$ corresponds to an ellipse within which we expect 95% of observati[illegible]he preceding figures, the red ellipses correspond to these inne[illegible]nd the vectors are the scaled eigenvalues.

Another complication with using the generalized variance is that it can equal zero. Denote the $n \times p$ matrix with deviation vectors in the columns as

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

$$= \underset{(n\times p)}{\mathbf{X}} - \underset{(n\times 1)(1\times p)}{\mathbf{1}\ \bar{\mathbf{x}}'}$$

It turns out that the generalized variance corresponding to these deviation vectors will equal zero if and only if the columns of $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ are linearly dependent.

To prove the preceding result, suppose that the columns of $\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}'$ are linearly dependent, in which case there is some $\mathbf{a} \neq \mathbf{0}$ such that $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$. Noting that $(n-1)\mathbf{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$, we then have that

$$(n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$$

so that the columns of $\mathbf{S}$ are also linearly dependent. It turns out that the determinant of a square matrix whose columns are linearly dependent equals zero. Similarly, suppose that $|\mathbf{S}| = 0$. A zero determinant also implies linear dependence of the columns of $\mathbf{S}$, based on which we can write

$$\mathbf{a}'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = L^2_{(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}} = \mathbf{0}$$

Then, since $\mathbf{a}'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}$ is a sum of squared differences, the length can only equal zero if all of the differences equal zero; i.e., we require that $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$, or linear dependence of the columns of $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$.

Example (Part 1): Note that if $\mathbf{a}'\mathbf{x}_j = c$ for all j and $\mathbf{a} \neq \mathbf{0}$, then $\mathbf{a}'(\mathbf{x}_j - \bar{\mathbf{x}}) = 0$, so that

$$(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$$

Certainly, $c = 0$ satisfies the above condition, in which case the columns of $\mathbf{X}$ themselves are linearly dependent. However, the result holds more generally. For example, suppose we have data on stock portfolio composition for $n = 4$ randomly-selected portfolio managers. Suppose each portfolio is made up entirely of stocks, and let x_{j1} be the proportion of stocks in manager js portfolio that are large cap. Define x_{j2} and x_{j3} similarly, for mid- and small-cap stocks. Thus, $x_{j1} + x_{j2} + x_{j3} = \mathbf{a}'\mathbf{x}_j = 1$ for all j, with $\mathbf{a}' = \mathbf{1}'$. Here is a hypothetical realization:

$$\mathbf{X} = \begin{bmatrix} 0.25 & 0.25 & 0.50 \\ 0.15 & 0.65 & 0.20 \\ 0.40 & 0.30 & 0.30 \\ 0.70 & 0.05 & 0.25 \end{bmatrix}$$

Example (Part 2): We have $\bar{\mathbf{x}}' = [0.3750, 0.3125, 0.3125]$ and

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}' = \begin{bmatrix} -0.1250 & -0.0625 & 0.1875 \\ -0.2250 & 0.3375 & -0.1125 \\ 0.0250 & -0.0125 & -0.0125 \\ 0.3250 & -0.2625 & -0.0625 \end{bmatrix}$$

Note that the third deviation vector equals the negative sum of the first two deviation vectors, so the columns of $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ are linearly dependent. We can calculate

$$\begin{aligned} \mathbf{S} &= \frac{1}{3}\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'\right)'\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'\right) \\ &= \frac{1}{3}\begin{bmatrix} 0.1725 & -0.1538 & -0.0188 \\ -0.1538 & 0.1869 & -0.0331 \\ -0.0188 & -0.0331 & 0.0519 \end{bmatrix} \end{aligned}$$

And $|\mathbf{S}| = 0$.

Example (Part 3): Note also that, if $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$, then

$$(n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{0} = \mathbf{0}$$

so the columns of $\mathbf{S}$ are also linearly dependent. Furthermore, since $\mathbf{S}\mathbf{a} = \mathbf{0} = 0\mathbf{a}$, $\mathbf{a}$ is an eigenvector of $\mathbf{S}$ with associated eigenvalue 0. Computing the eigenvalues and eigenvectors of $\mathbf{S}$ will therefore provide information on the linear combination responsible for linear dependence. In our example, the third eigenvalue is 0, with a corresponding normalized eigenvector of $\mathbf{e}_3' = (-1/\sqrt{3})\mathbf{1}'$. Ignoring the normalization constant, the linear combination suggested by $\mathbf{e}_3$ is the sum (as we would expect).

Summary Comments:

- The generalized variance is a one-number summary of the variability in a multivariate data set.
- Covariance matrices that differ greatly in correlation structure can have equal generalized variances.
- Generalized variances equal zero if any of the following (equivalent) conditions hold:
  - There is a linear combination of the variables that is constant for all items. Special case: one variable is a linear combination of the others.
  - The deviation vectors are linearly dependent.
  - The covariance matrix has a zero eigenvalue.
- A generalized variance of zero means that one or more of the variables can be expressed in terms of the others. Any such variables should be dropped from the analysis if possible.

Summary Comments:

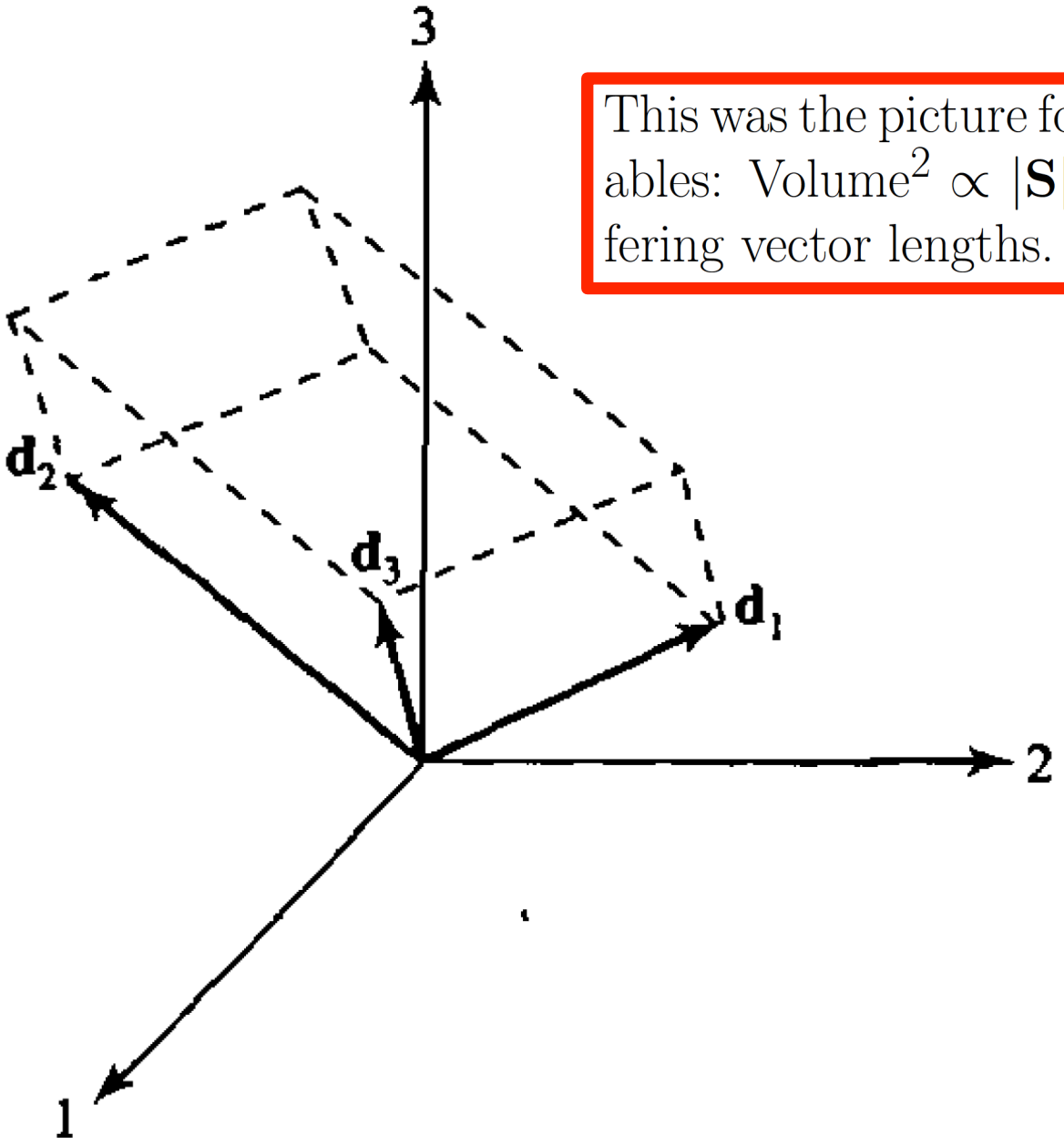
- The generalized variance is a one-number summary of the variability in a multivariate data set.
- Covariance matrices that differ greatly in correlation structure can have equal generalized variances.
- Generalized variances equal zero if any of the following (equivalent) conditions hold:
  - There is a linear combination of the variables that is constant for all i… …e other…
  - The…
  - The…

One last thing: $|\mathbf{S}| = 0$ if $n \leq p$.
And $n \leq p$ a lot these days.

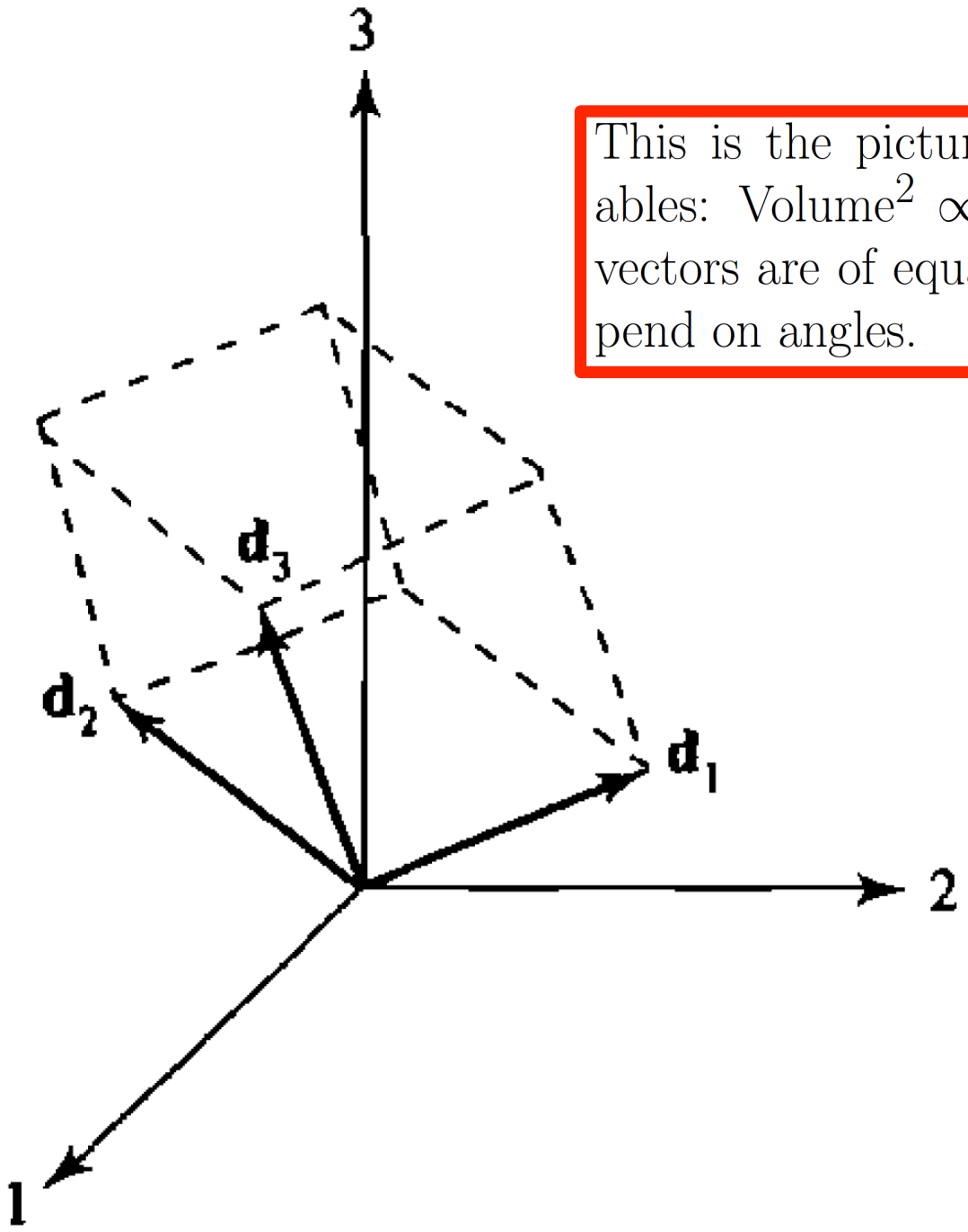
- A generalized variance of zero means that one or more of the variables can be expressed in terms of the others. Any such variables should be dropped from the analysis if possible.

An alternative to $|\mathbf{S}|$ is $|\mathbf{R}|$, where $\mathbf{R}$ is the sample correlation matrix. This corresponds to computing the generalized variance on the standardized variables, replacing x_{jk} with $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$, or the kth column with $(\mathbf{y}_k - \bar{x}_k\mathbf{1})/\sqrt{s_{kk}}$. Each of the standardized vectors have the same length $(\sqrt{n-1})$, so the magnitude of $|\mathbf{R}|$ only depends on the angles between them. In particular, $|\mathbf{R}|$ will be large when the vectors are nearly perpendicular and will be small when two or more of the vectors are in almost the same direction. Furthermore, the cosine of the angle between the ith and kth standardized vectors equals the sample correlation coefficient r_{ik}. So, $|\mathbf{R}|$ will be large if all of the r_{ik} are close to zero (corresponding to the vectors being nearly perpendicular to one another). Similarly, $|\mathbf{R}|$ will be small if one or more of the r_{ik} are close to ± 1.

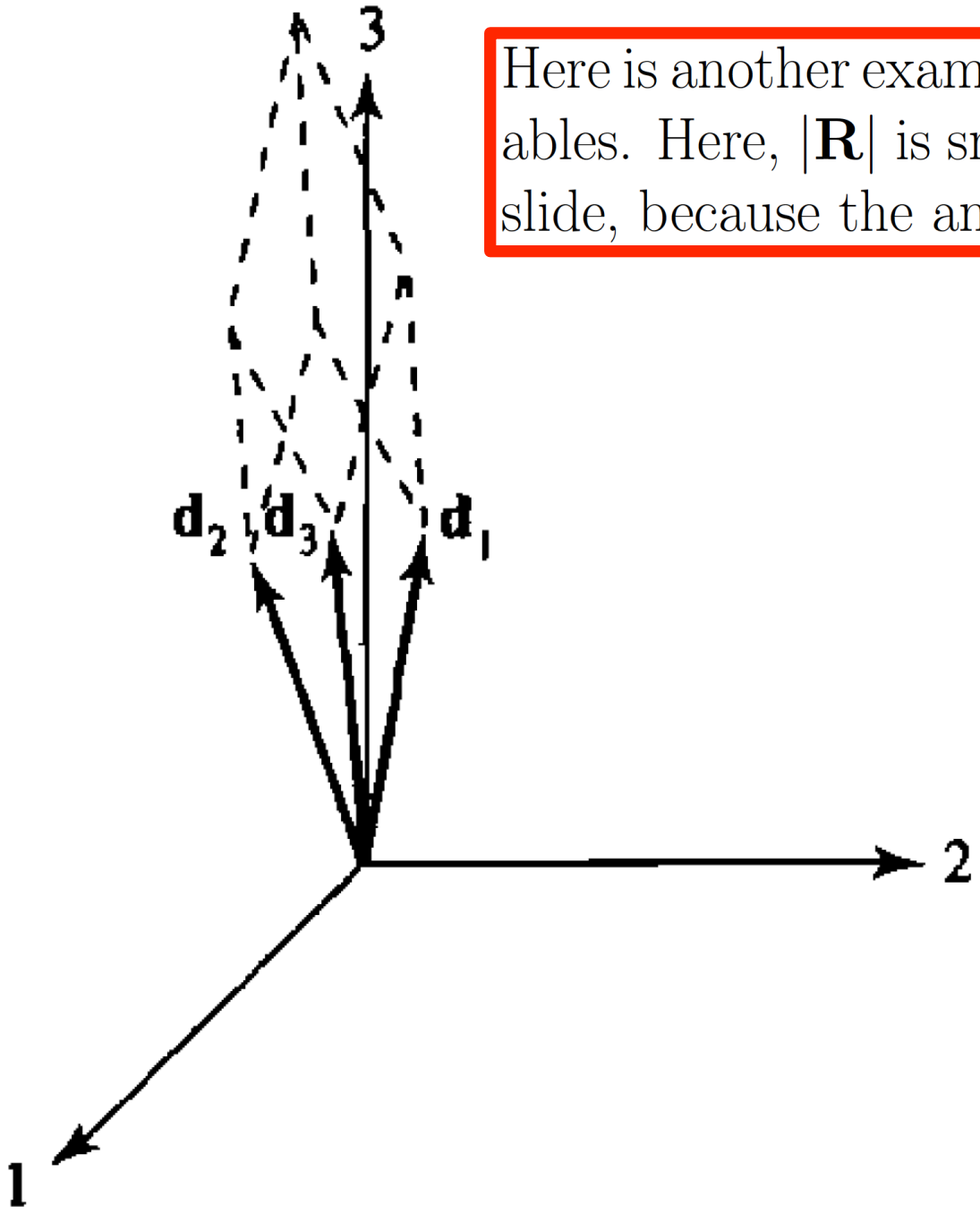
As with $|\mathbf{S}|$, we can interpret $|\mathbf{R}|$ as a volume, now corresponding to that delineated by the standardized variables: $|\mathbf{R}| = (n-1)^{-p}(\text{volume})^2$.



This was the picture for the unstandardized variables: Volume^2 $\propto |\mathbf{S}|$. Volume affected by differing vector lengths.



This is the picture for the standardized variables: Volume^2 $\propto |\mathbf{R}|$. Note that each of the vectors are of equal length, so volume only depend on angles.



Here is another example using standardized variables. Here, $|\mathbf{R}|$ is smaller than on the previous slide, because the angles are small.

Sample Statistics as Matrix Operations

$\bar{\mathbf{X}}$ and $\mathbf{S}$: First, note that $\bar{\mathbf{x}} = (1/n)\mathbf{X}'\mathbf{1}$. We have already used the fact that the $n \times p$ matrix of means is

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

and the $n \times p$ matrix of deviations is

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

We can then write

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

$\bar{\mathbf{X}}$ **and S**: First, note that $\bar{\mathbf{x}} = (1/n)\mathbf{X}'\mathbf{1}$. We have already used the fact that the $n \times p$ matrix of means is

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

and the $n \times p$ matrix of deviations is

Note that

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' - \frac{1}{n}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

We can then write

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

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

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

$\bar{\mathbf{X}}$ **and S**: First, note that $\bar{\mathbf{x}} = (1/n)\mathbf{X}'\mathbf{1}$. We have already used the fact that the $n \times p$ matrix of means

Define $\mathbf{D}^{1/2}$ as the *sample standard deviation matrix*, the diagonal matrix with (k,k)th element equal to $\sqrt{s_{kk}}$. Then, $\mathbf{D}^{-1/2}$ is the diagonal matrix with (k,k)th element equal to $1/\sqrt{s_{kk}}$. Then we can write the sample correlation matrix as

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

and the sample covariance matrix as

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

and the $n \times p$ m

$$\begin{bmatrix} x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

We can then write

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

Linear Combinations of the Components of X: Consider a single random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ and a vector of constants $\mathbf{b}' = [b_1, b_2, \ldots, b_p]$, so that $\mathbf{b}'\mathbf{X} = b_1X_2 + b_2X_2 + \cdots + b_pX_p$ is a linear combination. Suppose we have a random sample of n realizations of $\mathbf{X}$: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and that we are interested in the linear combinations defined by $\mathbf{b}$. Then

$$\text{The sample mean of the } \mathbf{b}'\mathbf{x}_j = \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}'\mathbf{x}_j = \mathbf{b}'\bar{\mathbf{x}}$$

Similarly,

$$\text{The sample variance of the } \mathbf{b}'\mathbf{x}_j = \frac{1}{n-1}\sum_{j=1}^{n} \left(\mathbf{b}'\mathbf{x}_j - \mathbf{b}'\bar{\mathbf{x}}\right)^2$$

$$= \mathbf{b}' \left[\frac{1}{n-1}\sum_{j=1}^{n} \left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'\right] \mathbf{b} = \mathbf{b}'\mathbf{S}\mathbf{b}$$

Also, with a second vector of constants $\mathbf{c}' = [c_1, c_2, \ldots, c_p]$,

$$\text{The sample covariance of the } \mathbf{b}'\mathbf{x}_j \text{ and } \mathbf{c}'\mathbf{x}_j = \mathbf{b}'\mathbf{S}\mathbf{c}$$

Linear Combinations of the Components of X: Consider a single random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ and a vector of constants $\mathbf{b}' = [b_1, b_2, \ldots, b_p]$, so that $\mathbf{b}'\mathbf{X} = b_1X_2 + b_2X_2 + \cdots + b_pX_p$ is a linear combination. Suppose we have a random sample of n realizations of $\mathbf{X}$: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and that we are interested in the linear combinations defined by $\mathbf{b}$. Then

Because, for two p-vectors $\mathbf{u}$ and $\mathbf{v}$, $(\mathbf{u}'\mathbf{v})^2 = \mathbf{u}'\mathbf{v}\mathbf{v}'\mathbf{u}$.

Similarly,

The sample variance of the $\mathbf{b}'\mathbf{x}_j = \dfrac{1}{n-1}\sum_{j=1}^{n}(\mathbf{b}'\mathbf{x}_j - \mathbf{b}'\bar{\mathbf{x}})^2$

$$= \mathbf{b}'\left[\frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right]\mathbf{b} = \mathbf{b}'\mathbf{S}\mathbf{b}$$

Also, with a second vector of constants $\mathbf{c}' = [c_1, c_2, \ldots, c_p]$,

The sample covariance of the $\mathbf{b}'\mathbf{x}_j$ and $\mathbf{c}'\mathbf{x}_j = \mathbf{b}'\mathbf{S}\mathbf{c}$

Linear Combinations of the Components of X: Consider a single random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ and a vector of constants $\mathbf{b}' = [b_1, b_2, \ldots, b_p]$, so that $\mathbf{b}'\mathbf{X} = b_1X_2 + b_2X_2 + \cdots + b_pX_p$ is a linear combination. Suppose we have a random sample of n realizations of $\mathbf{X}$: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and that we are interested in the linear combinations defined by $\mathbf{b}$. Then

$$\text{The sample mean of the } \mathbf{b}'\mathbf{x}_j = \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}'\mathbf{x}_j = \mathbf{b}'\bar{\mathbf{x}}$$

Similarly,

$$\text{The sample variance of the } \mathbf{b}'\mathbf{x}_j = \frac{1}{n-1}\sum_{j=1}^{n} \left(\mathbf{b}'\mathbf{x}_j - \mathbf{b}'\bar{\mathbf{x}}\right)^2$$

$$= \mathbf{b}'\left[\frac{1}{n-1}\sum_{j=1}^{n} \left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'\right]\mathbf{b} = \mathbf{b}'\mathbf{S}\mathbf{b}$$

Also, with a second vector of constants $\mathbf{c}' = [c_1, c_2, \ldots, c_p]$

$$\text{The sample covariance of the } \mathbf{b}'\mathbf{x}_j \text{ and } \mathbf{c}'\mathbf{x}_j =$$