
Sample Geometry and Random Sampling

Edps/Soc 584 and Psych 594

Applied Multivariate Statistics

Carolyn J. Anderson

Department of Educational Psychology



ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

© Board of Trustees, University of Illinois



Outline

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Motivation
- Variable Space
- Observation Space
 - ◆ What it is.
 - ◆ Mean
 - ◆ Covariance
- Global summary statistic for S

Reading & Resources:

- Johnson & Wichern, Chapter 3



Motivation

● Outline

Motivation

● Motivation

● Assumptions

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- The sample \bar{x} , S and R have geometric interpretations.
- The geometric interpretation shows how \bar{x} , S , and R are related to $X_{n \times p}$.
- Studying these relationships provides insight into multivariate methods.
- Introduce a summary statistic that describes the variability in the data. This is based on geometry.



Assumptions

The geometric interpretation of sample descriptive statistics (i.e., \bar{x} , S , and R) is based on two assumptions.

- (1) Each row of the data is unrelated to all of the others (i.e., independent observations).
 - ◆ Each row corresponds to a case, individual, sampling unit.
 - ◆ Each row is a multivariate observation (point) in p -dimensional space.

$$\mathbf{X}_{n \times p} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \vdots & X_{np} \end{pmatrix}$$

- (2) The joint distribution of all p variables is the same for all cases (individuals, rows). We are drawing multivariate observations from the same population and it does not change.

● Outline

Motivation

● Motivation

● Assumptions

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance



The Variable Space

- Each row of X is a point in “ p -space” or variable space.
- The variables (columns of X) define the axes.
- Consider the n points in the p dimensional space.

The center of the point “cloud” is \bar{x} .

The variability and covariability is measured by S .

● Outline

Motivation

Variable Space

● The Variable Space

● The Variable Space: Example

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

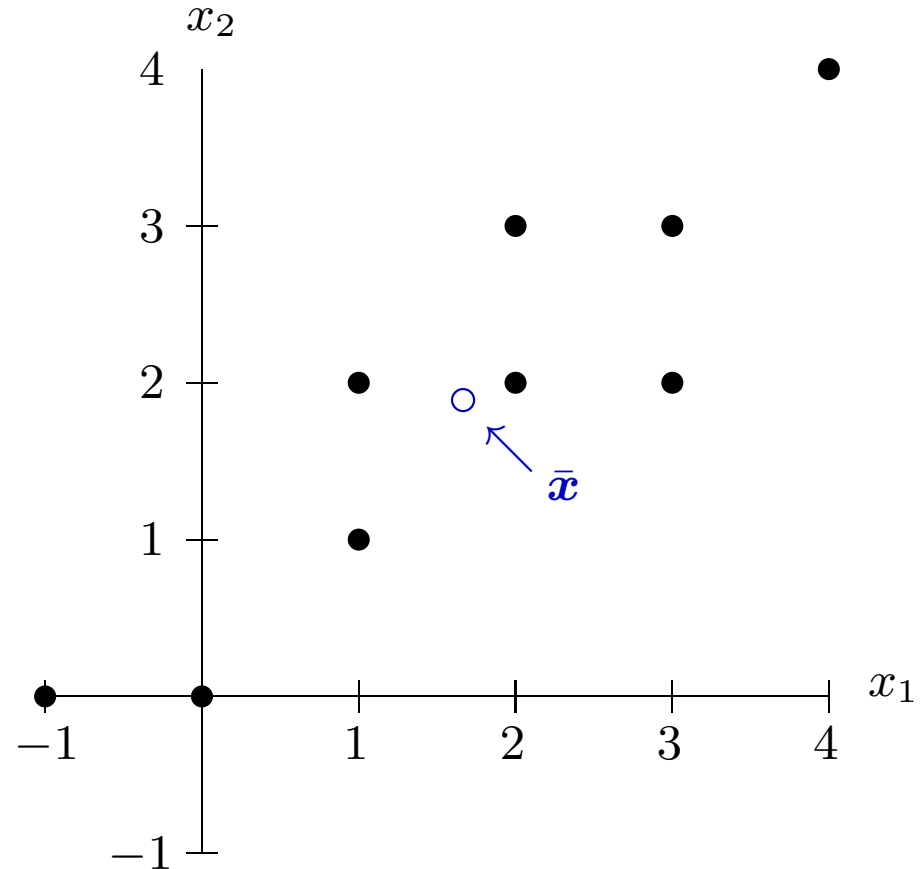
Generalized Variance



The Variable Space: Example

The data

$$\mathbf{X}_{9 \times 2} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 3 & 2 \\ -1 & 0 \\ 3 & 3 \\ 2 & 3 \\ 1 & 2 \\ 4 & 4 \end{pmatrix}$$



$$\bar{\mathbf{x}} = \frac{1}{9} \begin{pmatrix} 15 \\ 17 \end{pmatrix} = \begin{pmatrix} 1.66 \\ 1.88 \end{pmatrix}$$

What would happen to picture if multiply \mathbf{X} by 2?



The Observation Space

Important alternative geometric representation of \mathbf{X} that considers the data as p points in an n -dimensional space.

- Each observation (case, individual) defines an axis or dimension of the space.
- Each column of \mathbf{X} is a **vector** in the n -space.

$$\mathbf{X}_{n \times p} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \vdots & X_{np} \end{pmatrix} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p)$$

where

$$\mathbf{y}_k = \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{pmatrix} \quad \text{for } k = 1, \dots, p.$$

● Outline

Motivation

Variable Space

Observation Space

● The Observation Space

● The Observation Space:
Example

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected
Values

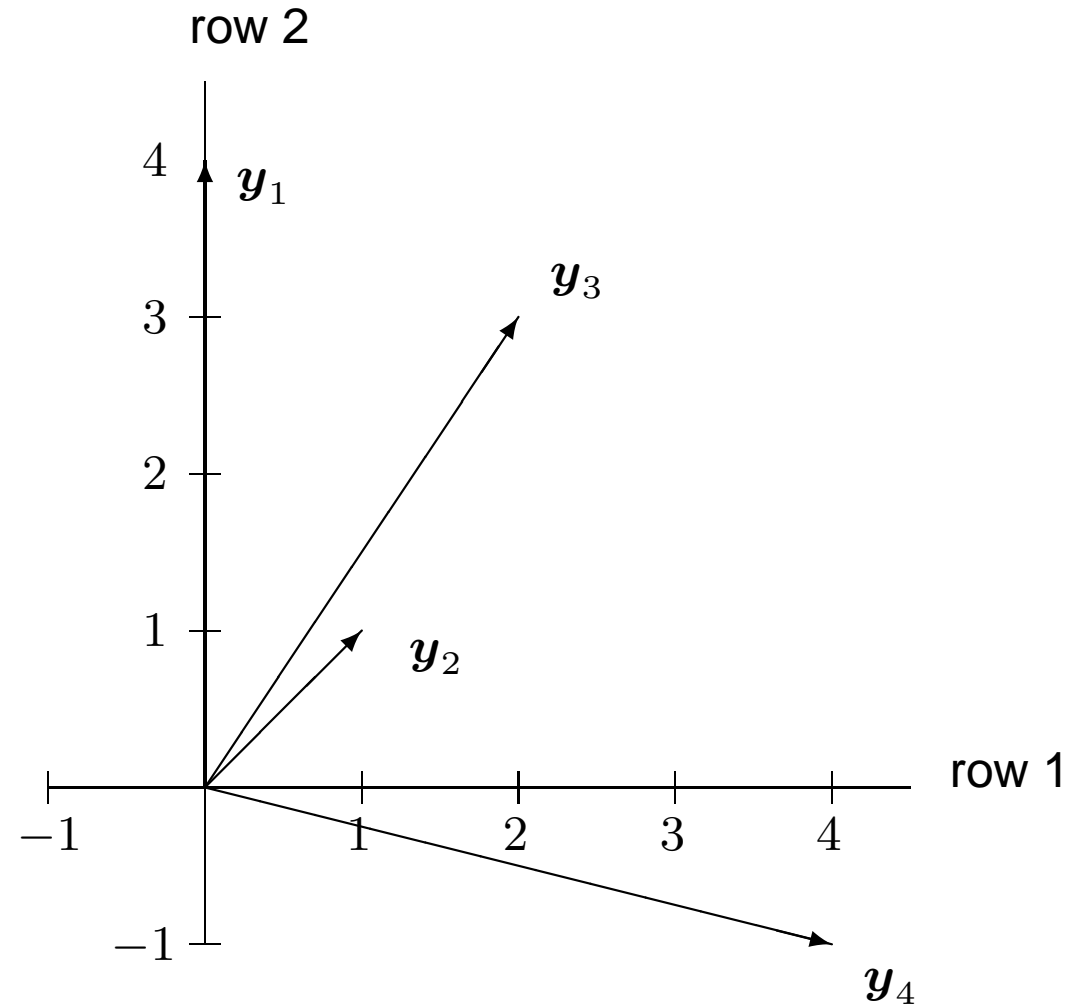
Generalized Variance



The Observation Space: Example

The data

$$\mathbf{X}_{2 \times 4} = \begin{pmatrix} 0 & 1 & 2 & 4 \\ 4 & 1 & 3 & -1 \end{pmatrix}$$



In this space, we talk about the lengths of the vector variables and about the angle between them.

● Outline

Motivation

Variable Space

Observation Space

● The Observation Space

● The Observation Space:
Example

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected
Values

Generalized Variance



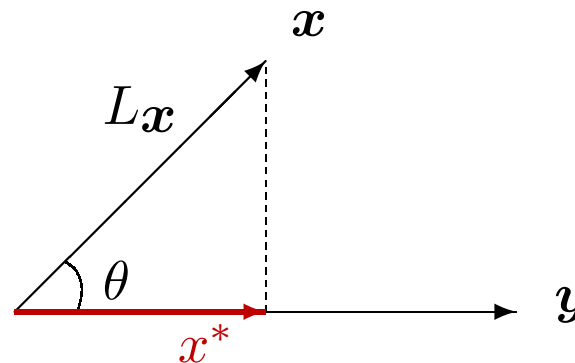
The Observation Space: The mean

■ **Question:** How does \bar{x} relate to the vectors in the observation space?

■ **Answer:** The projection of the vector variable onto $\mathbf{1}_n$.

Quick recall: The projection of x onto y equals

$$x^* = \frac{x'y}{y'y} y = \left(\frac{x'y}{Ly} \right) \left(\frac{y}{Ly} \right)$$



● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

● The Observation Space: The mean

● The Mean & Projection

● The Mean in the Observation Space

● Back to our little example

● 2-dimensional example continued

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance



The Mean & Projection

- The “equal angular vector” is an $(n \times 1)$ vector

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

- Projection of $\mathbf{y}'_k = (x_{1k}, x_{2k}, \dots, x_{nk})$ on the the vector $\mathbf{1}_n =$

$$\begin{aligned} \left(\frac{\mathbf{y}'_k \mathbf{1}_n}{\mathbf{1}'_n \mathbf{1}_n} \right) \mathbf{1} &= \frac{(x_{1k}, x_{2k}, \dots, x_{nk}) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}{\underbrace{(1 + 1 + \dots + 1)}_n} \mathbf{1}_n \\ &= \frac{\sum_{j=1}^n x_{jk}}{n} \mathbf{1}_n = \bar{x}_k \mathbf{1}_n \end{aligned}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

● The Observation Space: The mean

● The Mean & Projection

● The Mean in the Observation Space

● Back to our little example

● 2-dimensional example continued

Covariance & Correlation

Random Sample and Expected Values

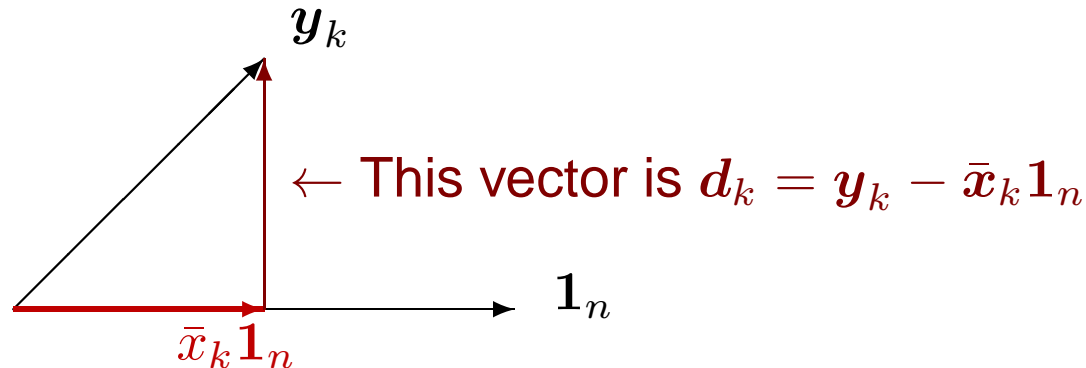
Generalized Variance



The Mean in the Observation Space

The sample mean \bar{x}_k of the k^{th} variable corresponds to the multiple of $\mathbf{1}_n$ required to give the projection of \mathbf{y}_k onto the vector $\mathbf{1}_n$.

- Projection of \mathbf{y}_k onto $\mathbf{1}_n$:



- d_k is the **Deviation vector**.
- We have the decomposition of \mathbf{y}_k into two parts:

$$\begin{aligned}\mathbf{y}_k &= \underbrace{\bar{x}_k}_{\text{mean}} \mathbf{1}_n + \underbrace{d_k}_{\text{deviation from mean}} \\ &= \underbrace{\bar{x}_k}_{\text{mean}} \mathbf{1}_n + \underbrace{\mathbf{y}_k - \bar{x}_k \mathbf{1}_n}_{\text{deviation from mean}}\end{aligned}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

● The Observation Space: The mean

● The Mean & Projection

● The Mean in the Observation Space

● Back to our little example

● 2-dimensional example continued

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance



Back to our little example

- Outline

- Motivation

- Variable Space

- Observation Space

- Obs Space: Mean

- The Observation Space: The mean

- The Mean & Projection

- The Mean in the Observation Space

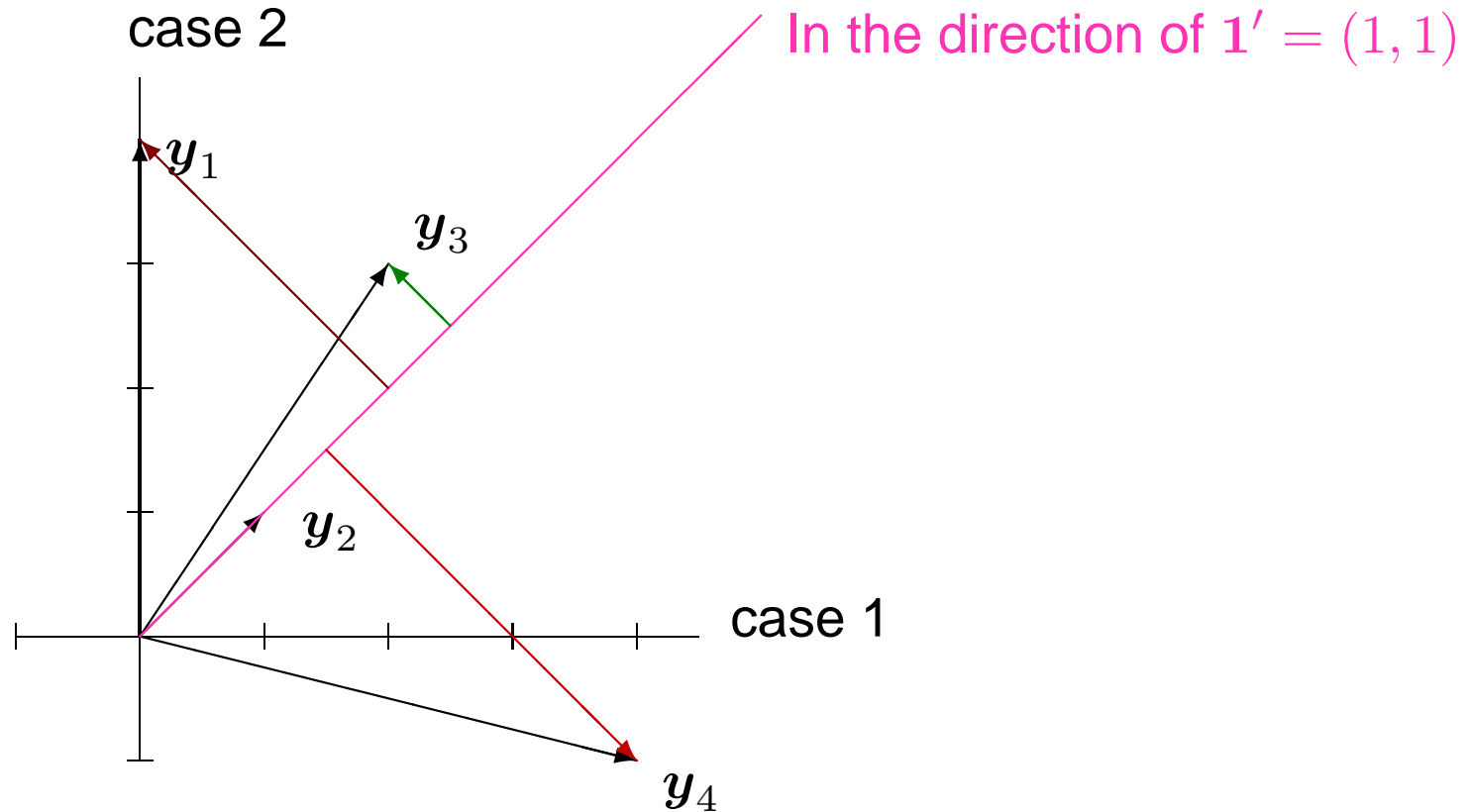
- **Back to our little example**

- 2-dimensional example continued

- Covariance & Correlation

- Random Sample and Expected Values

- Generalized Variance



$$\bar{x}_1 \mathbf{1} = ([0(1) + 4(1)]/2) \mathbf{1} = (2.0) \mathbf{1} \rightarrow \bar{x}_1 = 2$$

$$\bar{x}_2 \mathbf{1} = ([1(1) + 1(1)]/2) \mathbf{1} = (1.0) \mathbf{1} \rightarrow \bar{x}_2 = 1$$

$$\bar{x}_3 \mathbf{1} = ([2(1) + 3(1)]/2) \mathbf{1} = (2.5) \mathbf{1} \rightarrow \bar{x}_3 = 2.5$$

$$\bar{x}_4 \mathbf{1} = ([4(1) - 1(1)]/2) \mathbf{1} = (1.5) \mathbf{1} \rightarrow \bar{x}_4 = 1.5$$



2-dimensional example continued

For our 2-case example with 4 variables

$$\mathbf{X}_{2 \times 4} = \begin{pmatrix} 0 & 1 & 2 & 4 \\ 4 & 1 & 3 & -1 \end{pmatrix}$$

Note:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ -1 & 0 & -1 & 1 \end{pmatrix}$$

What is the rank of this correlation matrix?

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

● The Observation Space: The mean

● The Mean & Projection

● The Mean in the Observation Space

● Back to our little example

● 2-dimensional example continued

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance



Deviation Vectors: Variances & Covariance

- **Sample variance:** The squared lengths of deviation vectors,

$$L^2 \mathbf{d}_k = \mathbf{d}'_k \mathbf{d}_k = \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2,$$

are sums of squared deviation from the mean, so

$$s_{kk} = \frac{1}{n} L^2 \mathbf{d}_k = \frac{1}{n} \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2.$$

- **Sample covariance:** The inner products between two deviation vectors,

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

are sums of cross products, so

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

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation

Space

Random Sample and Expected Values

Generalized Variance



Deviation Vectors: Correlation

The angle between two deviation vectors equals

$$\begin{aligned}
\cos(\theta_{ik}) &= \frac{\mathbf{d}'_i \mathbf{d}_k}{L_{\mathbf{d}_i} L_{\mathbf{d}_k}} \\
&= \frac{\sum_{j=1}^n (x_{jk} - \bar{x}_k)(x_{ji} - \bar{x}_i)}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}} \\
&= r_{ik}
\end{aligned}$$

The cosine of the angle between two deviation vectors is the **sample correlation coefficient** of the variables.

- If two deviation vectors have the same orientation (same direction), $\theta = 0^\circ \longrightarrow \cos(\theta) = 1 = r_{ik}$.
- If two deviation vectors have the opposite orientation, $\theta = 180^\circ \longrightarrow \cos(\theta) = -1 = r_{ik}$.
- If two deviation vectors are perpendicular (orthogonal), $\theta = 90^\circ \longrightarrow \cos(\theta) = 0 = r_{ik}$.

[Back to simple example \(pages 12-13\)](#)

- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
 - Deviation Vectors: Variances & Covariance
 - **Deviation Vectors: Correlation**
 - Correlation: Simple Example
 - 3-Dimensional Example
 - 3-Dimensional Example: Variable 1
 - 3-Dimensional Example: Graph of Variables
 - 3-Dimensional Example: Deviation Vectors
 - Graph of 3 deviation vectors
 - Variances
 - Correlations
 - Summary: Observation Space
- Random Sample and Expected Values
- Generalized Variance

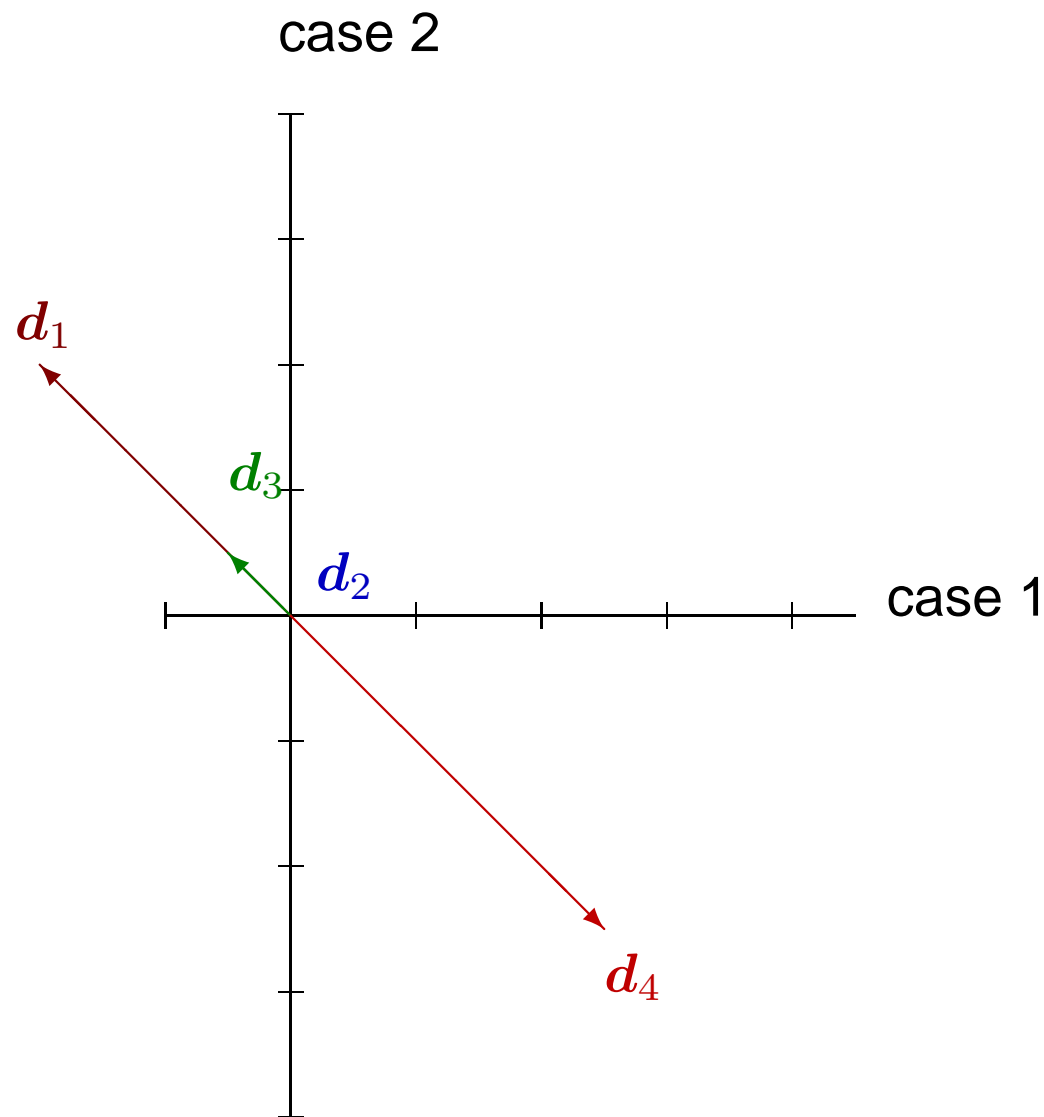


Correlation: Simple Example

- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
 - Deviation Vectors: Variances & Covariance
 - Deviation Vectors: Correlation
 - **Correlation: Simple Example**
 - 3-Dimensional Example
 - 3-Dimensional Example: Variable 1
 - 3-Dimensional Example: Graph of Variables
 - 3-Dimensional Example
 - Deviation Vectors
 - Graph of 3 deviation vectors
 - Variances
 - Correlations
 - Summary: Observation Space
- Random Sample and Expected Values
- Generalized Variance

$$X_{2 \times 4} = \begin{pmatrix} 0 & 1 & 2 & 4 \\ 4 & 1 & 3 & -1 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ -1 & 0 & -1 & 1 \end{pmatrix}$$





3-Dimensional Example

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● **3-Dimensional Example**

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation

Space

Random Sample and Expected

Values

Generalized Variance

Suppose that $p = 3$, $n = 3$ and the data are

$$\mathbf{X} = \begin{pmatrix} 700 & 650 & 750 \\ 500 & 550 & 640 \\ 660 & 750 & 650 \end{pmatrix}$$

So

$$\mathbf{y}_1 = \begin{pmatrix} 700 \\ 500 \\ 660 \end{pmatrix}$$

$$\mathbf{y}_2 = \begin{pmatrix} 650 \\ 550 \\ 750 \end{pmatrix}$$

$$\mathbf{y}_3 = \begin{pmatrix} 750 \\ 640 \\ 650 \end{pmatrix}$$

$$\bar{x}_1 = \frac{1860}{3} = 620$$

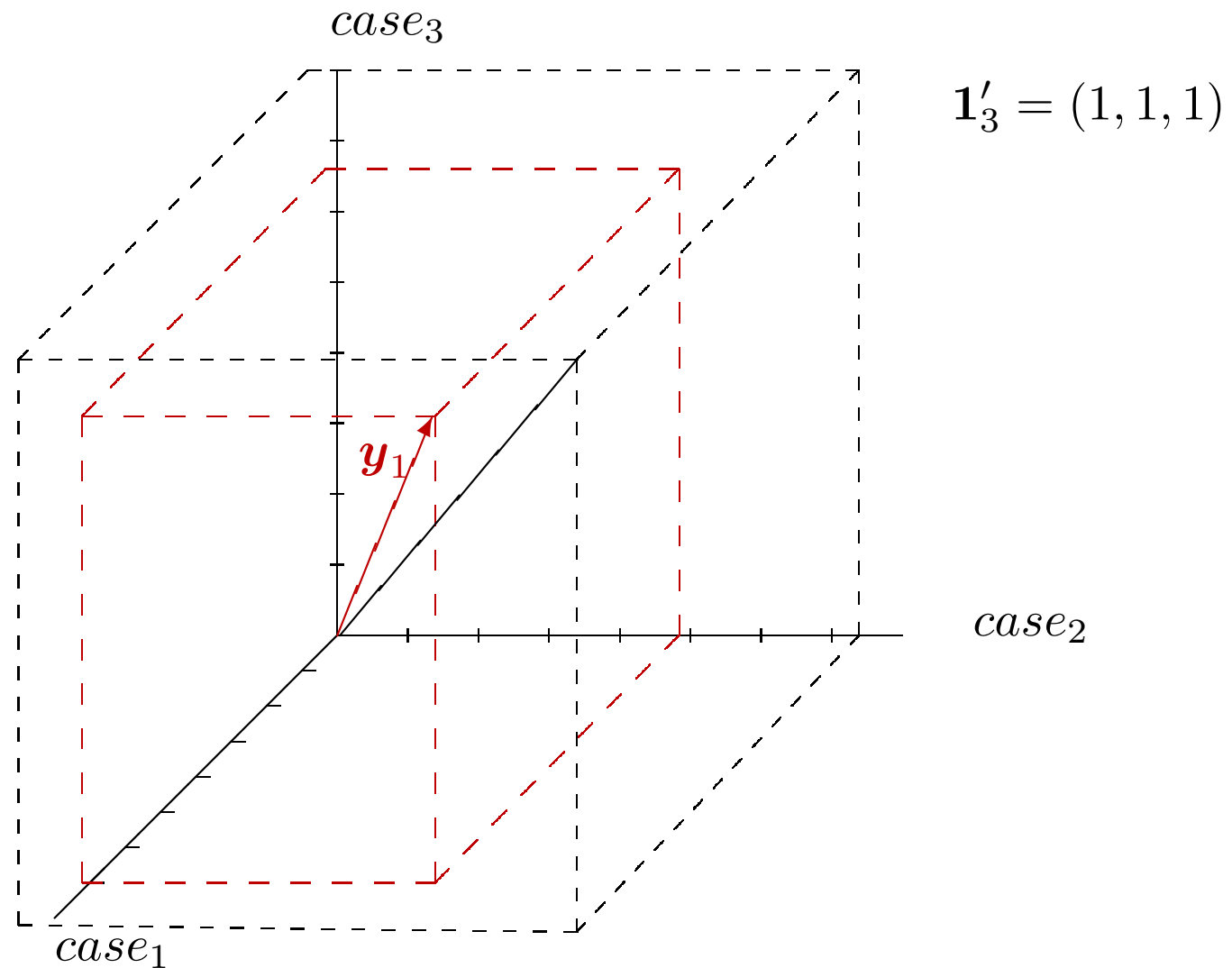
$$\bar{x}_2 = \frac{1950}{3} = 650$$

$$\bar{x}_3 = \frac{2040}{3} = 680$$



3-Dimensional Example: Variable 1

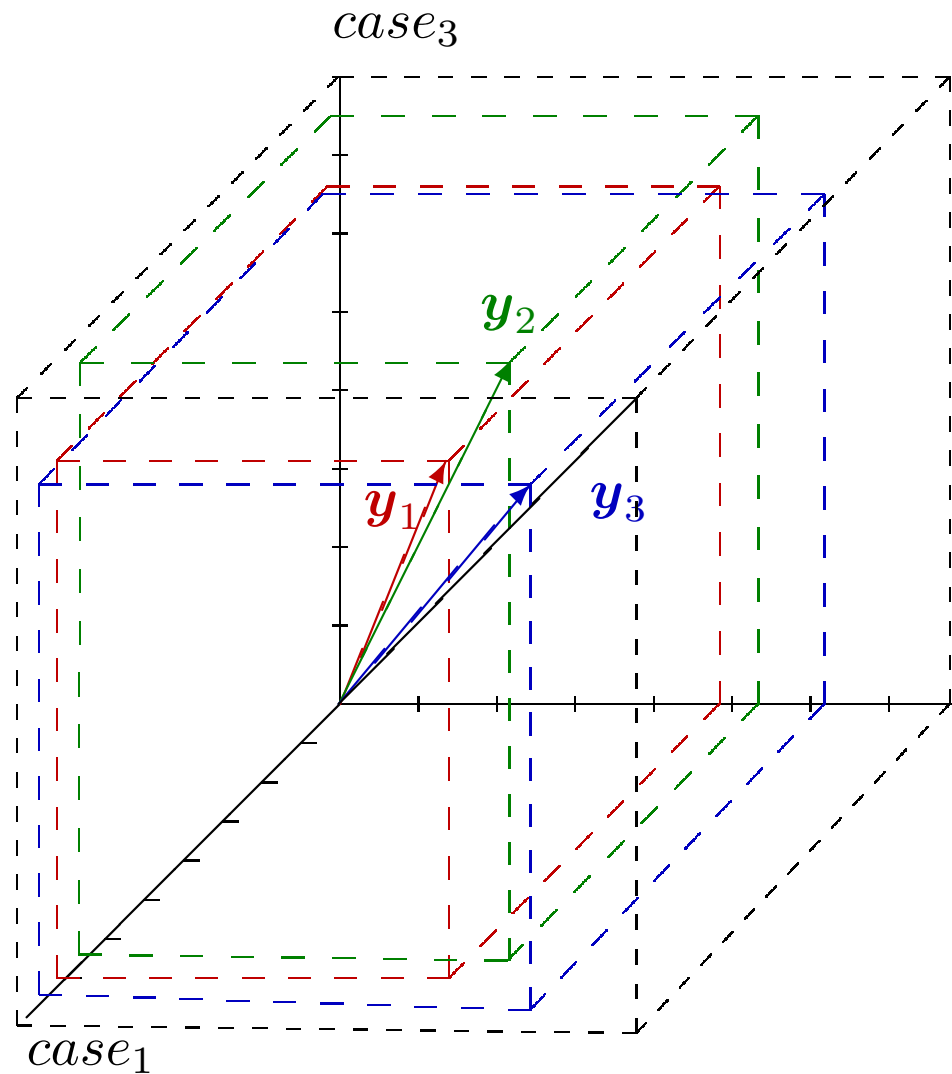
- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
 - Deviation Vectors: Variances & Covariance
 - Deviation Vectors: Correlation
 - Correlation: Simple Example
 - 3-Dimensional Example
 - 3-Dimensional Example: Variable 1
 - 3-Dimensional Example: Graph of Variables
 - 3-Dimensional Example: Deviation Vectors
 - Graph of 3 deviation vectors
 - Variances
 - Correlations
 - Summary: Observation Space
- Random Sample and Expected Values
- Generalized Variance





3-Dimensional Example: Graph of Variables

- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
 - Deviation Vectors: Variances & Covariance
 - Deviation Vectors: Correlation
 - Correlation: Simple Example
 - 3-Dimensional Example
 - 3-Dimensional Example: Variable 1
 - **3-Dimensional Example: Graph of Variables**
 - 3-Dimensional Example
 - Deviation Vectors
 - Graph of 3 deviation vectors
 - Variances
 - Correlations
 - Summary: Observation Space
- Random Sample and Expected Values
- Generalized Variance



$$\mathbf{1}'_3 = (1, 1, 1)$$

case₂



3-Dimensional Example

Means are the multiples of 1 that give projection of y_i onto

Using the first variable for illustration

$$\frac{\mathbf{y}'_1 \mathbf{1}}{\mathbf{1}' \mathbf{1}} \mathbf{1} = \bar{x}_1 \mathbf{1} = 620 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 620 \\ 620 \\ 620 \end{pmatrix}$$

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1 \mathbf{1} = \begin{pmatrix} 700 \\ 500 \\ 660 \end{pmatrix} - \begin{pmatrix} 620 \\ 620 \\ 620 \end{pmatrix} = \begin{pmatrix} 80 \\ -120 \\ 40 \end{pmatrix}$$

Notes:

■ \mathbf{d}_1 is perpendicular to \bar{x}_1 as it should be:

$$\mathbf{d}'_1 \bar{x}_1 \mathbf{1} = (80, -120, 40) \begin{pmatrix} 620 \\ 620 \\ 620 \end{pmatrix} = 620(80 - 120 + 40) = 0$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation

Space

Random Sample and Expected Values

Generalized Variance



Deviation Vectors

Notes (continued):

- y_1 is decomposed into two parts:

$$y_1 = \bar{x}_1 \mathbf{1}_n + d_1$$

- If $d_i = 0$, then $y_i = \bar{x}_i \mathbf{1}$.

- All of them

$$d_1 = \begin{pmatrix} 80 \\ -120 \\ 40 \end{pmatrix}$$

$$d_2 = \begin{pmatrix} 0 \\ -100 \\ 100 \end{pmatrix}$$

$$d_3 = \begin{pmatrix} 70 \\ -40 \\ -30 \end{pmatrix}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation

Space

Random Sample and Expected

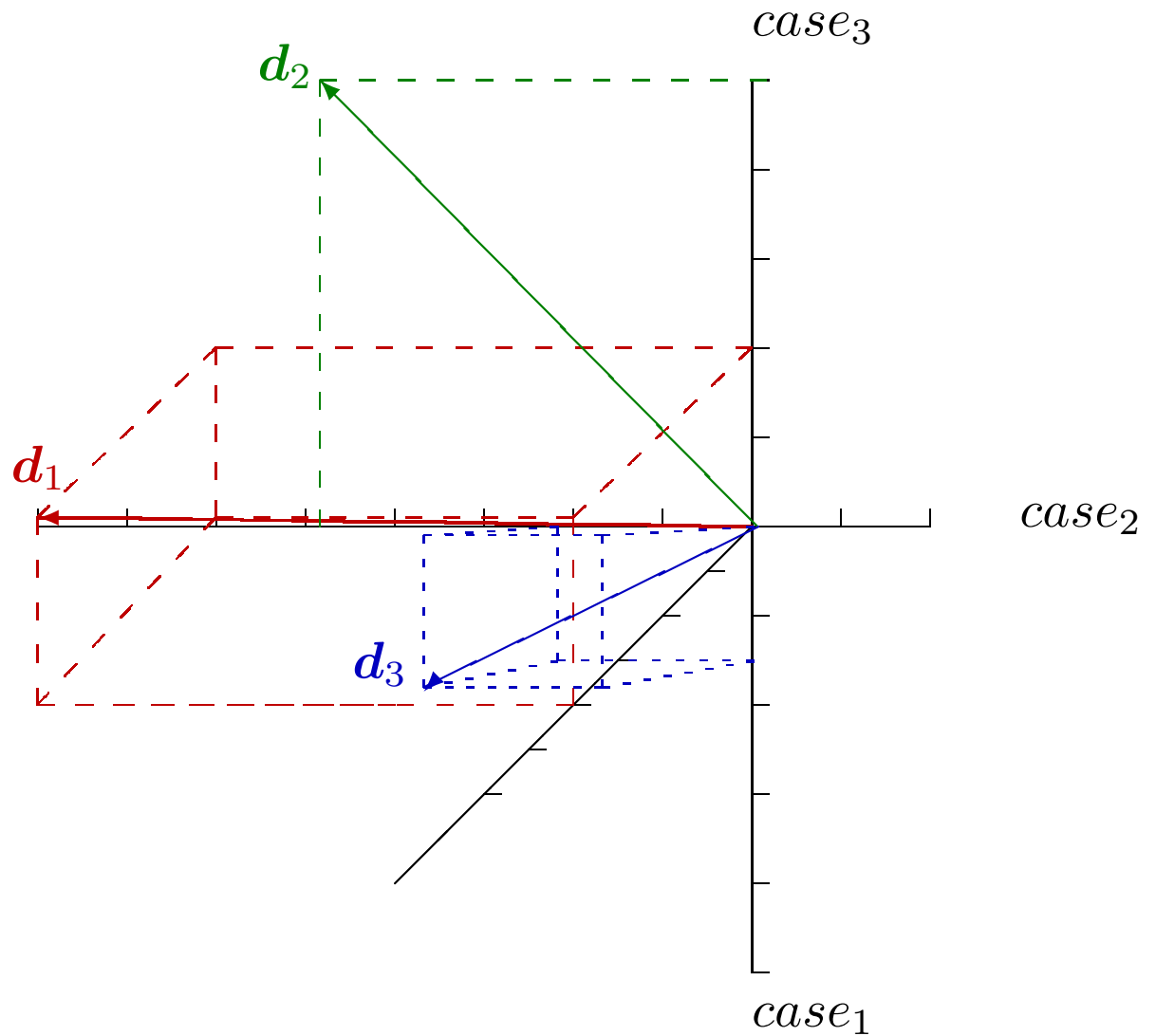
Values

Generalized Variance



Graph of 3 deviation vectors

- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
 - Deviation Vectors: Variances & Covariance
 - Deviation Vectors: Correlation
 - Correlation: Simple Example
 - 3-Dimensional Example
 - 3-Dimensional Example: Variable 1
 - 3-Dimensional Example: Graph of Variables
 - 3-Dimensional Example: Deviation Vectors
 - **Graph of 3 deviation vectors**
 - Variances
 - Correlations
 - Summary: Observation Space
- Random Sample and Expected Values
- Generalized Variance





Variances

Variance are proportional to the lengths of deviation vectors

Variable 1:

$$L_{d_1} = \sqrt{80^2 + 120^2 + 40^2} = \sqrt{22400} = 149.67 = \sqrt{ns_{11}} = \sqrt{3s_{11}}$$

So $s_{11} = 22400/3 = 7466.67$ and $\sqrt{s_{11}} = 86.41$

Variable 2:

$$L_{d_2} = \sqrt{0^2 + 100^2 + 100^2} = 100\sqrt{2} = 141.42 = \sqrt{3s_{22}}$$

So $s_{22} = 20000/3 = 6666.67$ and $\sqrt{s_{22}} = 81.65$

Variable 3:

$$L_{d_3} = \sqrt{70^2 + 40^2 + 30^2} = \sqrt{7400} = 86.02 = \sqrt{3s_{33}}$$

So $s_{33} = 7400/3 = 2466.67$ and $\sqrt{s_{33}} = 49.67$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances

& Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● **Variances**

● Correlations

● Summary: Observation

Space

Random Sample and Expected

Values

Generalized Variance



Correlations

Correlations are the cosines of the angles between deviation vectors

$$r_{12} = \frac{\mathbf{d}'_1 \mathbf{d}_2}{L_{\mathbf{d}_1} L_{\mathbf{d}_2}} = \frac{80(0) - 120(-100) + 40(100)}{(149.67)(141.42)} = \frac{16000}{21166} = .756$$

and the angle between \mathbf{d}_1 and $\mathbf{d}_2 = \cos^{-1}(.756) = 40.89^\circ$.

$$r_{13} = \frac{\mathbf{d}'_1 \mathbf{d}_3}{L_{\mathbf{d}_1} L_{\mathbf{d}_3}} = \frac{80(70) - 120(-40) + 40(-30)}{(149.67)(86.02)} = \frac{9200}{12874} = .715$$

and the angle between \mathbf{d}_1 and $\mathbf{d}_3 = \cos^{-1}(.715) = 44.39^\circ$.

$$r_{23} = \frac{\mathbf{d}'_2 \mathbf{d}_3}{L_{\mathbf{d}_2} L_{\mathbf{d}_3}} = \frac{0(70) - 100(-40) + 100(-30)}{(141.42)(86.02)} = \frac{1000}{12165.52} = .082$$

and the angle between \mathbf{d}_2 and $\mathbf{d}_3 = \cos^{-1}(.082) = 85.30^\circ$.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation

Space

Random Sample and Expected

Values

Generalized Variance



Summary: Observation Space

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

● Deviation Vectors: Variances & Covariance

● Deviation Vectors: Correlation

● Correlation: Simple Example

● 3-Dimensional Example

● 3-Dimensional Example:

Variable 1

● 3-Dimensional Example:

Graph of Variables

● 3-Dimensional Example

● Deviation Vectors

● Graph of 3 deviation vectors

● Variances

● Correlations

● Summary: Observation Space

Random Sample and Expected Values

Generalized Variance

- The axes of the observation space are defined by cases (individuals) and variables are represented as vectors.
- The projection of the column \mathbf{y}_k of the data matrix $\mathbf{X}_{n \times p}$ onto the equal angular vector $\mathbf{1}_n$ is the vector $\bar{x}_k \mathbf{1}$.
- The information contained in \mathbf{S} is obtained from the deviation vectors: $\mathbf{d}_k = \mathbf{y}_k - \bar{x}_k \mathbf{1}_n = \{ (x_{jk} - \bar{x}_k) \}$
The variance: $L^2_{\mathbf{d}_k} = \mathbf{d}'_k \mathbf{d}_k = n s_{kk}$
The covariance: $\mathbf{d}'_i \mathbf{d}_k = n s_{ik}$
- The sample correlation coefficient r_{ik} is the cosine of the angle between \mathbf{d}_i and \mathbf{d}_k .



Random Sample and Expected Values

We sample from a population to learn about a phenomenon of interest

We'll think about data (n cases) as a random sample of cases from some large population (real or hypothetical). For every case in the sample, we measure p variables.

So

- The measurements of the p variables for a **single case** will usually be **correlated or dependent**. Multivariate analysis constitutes techniques designed to account for and/or study these correlations/dependencies.
- The measurements from **different cases** must be **independent**. This assumption can be violated a number of ways (e.g., collect observations over time, poor measurement or experimental procedures, repeated observations of same case). Lack of independence can be a big problem.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

● Random Sample and Expected Values

● DATA

● Important Result

● Estimating Covariance Matrix

Generalized Variance



DATA

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

● Random Sample and Expected Values

● DATA

● Important Result

● Estimating Covariance Matrix

Generalized Variance

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \vdots \\ \mathbf{X}'_n \end{pmatrix}$$

where

$$\mathbf{X}'_j = (X_{j1}, X_{j2}, \dots, X_{jp})$$

Each \mathbf{X}_j is a random vector containing p measurements.

Distances between the n points in p -space are determined by the **joint probability function** governing each and every \mathbf{X}_j ,

$$f(\mathbf{x}_j) = f(x_{j1}, x_{j2}, \dots, x_{jp})$$



Important Result

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

● Random Sample and Expected Values

● DATA

● **Important Result**

● Estimating Covariance Matrix

Generalized Variance

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a random sample from the joint distribution $f(\mathbf{x}_j)$, which has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

- $\bar{\mathbf{X}}$ is an unbiased estimator of $\boldsymbol{\mu}$
- $\bar{\mathbf{X}}$ has covariance matrix $(1/n)\boldsymbol{\Sigma}$

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



Estimating Covariance Matrix

To estimate Σ , first note

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

So

$$\frac{n}{n-1} \mathbf{S}_n = \mathbf{S} = \frac{1}{n-1} \sum_{j=1}^n (\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})'$$

is an unbiased estimator of Σ .

- \mathbf{S} w/o a subscript has divisor $(n-1)$ and is unbiased.
- \mathbf{S}_n w/ a subscript has divisor n , biased, but is the MLE, $\hat{\Sigma}$.
- The $(i, k)^{th}$ element of the \mathbf{S} is an unbiased estimator of σ_{ik} :

$$s_{ik} = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

- $\sqrt{s_{ik}}$ is **not** an unbiased estimator of $\sqrt{\sigma_{ik}}$.
- r_{ik} is **not** an unbiased estimator of ρ_{ik} .
- The amount of bias is small for “large” n .

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

● Random Sample and

Expected Values

● DATA

● Important Result

● Estimating Covariance Matrix

Generalized Variance



Generalized Variance

The Sample covariance matrix describes the variation and covariation in and between the p variables,

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

where $s_{ik} = (1/(n - 1)) \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$.

- With just one variable, we usually only need just one statistics (the sample variance) to describe the variability in the data.
- With p variable we need p variances and $p(p - 1)/2$ covariances.
- It might be nice to have a single statistic that summarizes the information in S . This statistic should reflect all the variances and covariance.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

● Generalized Variance

● Generalized Sample Variance

● Example: GSV

● GSV: n -space interpretation

● Implications for Size of GSV

● Variable-space Interpretation of GSV

● Constant Statistical Distance

● Statistical Distance between two Points

● Rotation of axes

● Back to Interpretation of GSV

● GSV Interpretation

● When does $GSV = 0$

● Implications & Limitations of GSV

● Different patterns S , Same

● Sample Geometry and Random Sampling

● The GSV : variances and



Generalized Sample Variance

$$\begin{aligned}
 GSV &= \text{Generalized Sample Variance} \\
 &= \text{determinant of } S \\
 &= |S|
 \end{aligned}$$

Recall that the determinant of an $A_{k \times k}$ matrix is a scalar

$$|A| = a_{11} \quad \text{for } k = 1$$

$$|A| = a_{11}a_{22} - a_{12}a_{21} \quad \text{for } k = 2$$

$$|A| = \sum_{j=1}^k a_{ij} |A_{ij}| (-1)^{i+j} \quad \text{for } k > 1$$

where A_{ij} is the $(k-1) \times (k-1)$ sub-matrix of A where the i^{th} row and j^{th} column of A have been deleted.

- Outline

- Motivation

- Variable Space

- Observation Space

- Obs Space: Mean

- Covariance & Correlation

- Random Sample and Expected Values

- Generalized Variance

- Generalized Variance

- **Generalized Sample Variance**

- Example: GSV

- GSV: n -space interpretation

- Implications for Size of GSV

- Variable-space Interpretation of GSV

- Constant Statistical Distance

- Statistical Distance between two Points

- Rotation of axes

- Back to Interpretation of GSV

- GSV Interpretation

- When does $GSV = 0$

- Implications & Limitations of GSV

- Different patterns S , Same

- Example: Geometry and Random Sampling

- The GSV : variances and



Example: GSV

$$S = \begin{pmatrix} 3 & -1 & 2 \\ -1 & 4 & 3 \\ 2 & 3 & 9 \end{pmatrix}$$

$$\begin{aligned} GSV &= 3(4(9) - 3(3)) - (-1)(-1(9) - 2(3)) + 2(-1(3) - 4(2)) \\ &= 3(27) + 1(-15) + 2(-11) \\ &= 44 \end{aligned}$$

What does this mean?

Two ways to interpret it (geometrically):

- Using observation space (n -dimensional)
- Using variable space (p -dimensional)

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV
- Different patterns S , Same GSV
- Example: Geometry and Random Sampling
- The GSV : variances and

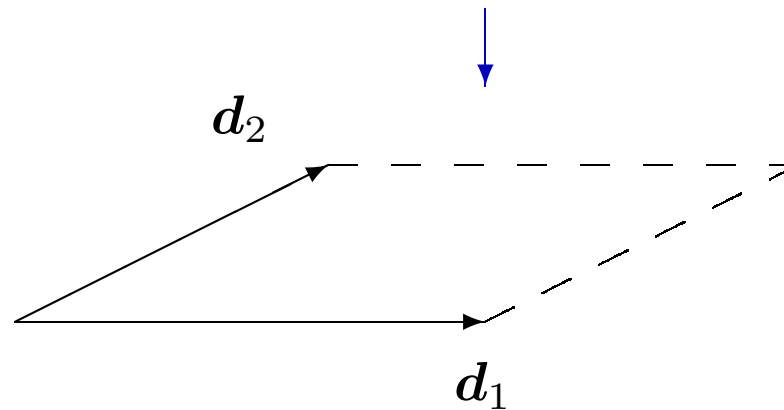


GSV: n -space interpretation

The $GSV = |\mathbf{S}|$ is related to the volume of the parallelepiped (geometric figure) defined by the p deviation vectors.

e.g., $p = 2$ and $n =$ “whatever”

The volume (“area” for $p = 2$)



Remember $d_i = y_i - \bar{x}_i \mathbf{1}$ (y_i is column vector of matrix \mathbf{X})
Specifically,

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

or volume = $\sqrt{|\mathbf{S}|} (n - p)^{p/2}$.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV

● **GSV: n -space interpretation**

- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV

GSV

- Different patterns \mathbf{S} , Same \mathbf{S}
- Example: Geometry and Random Sampling
- The GSV : variances and



Implications for Size of GSV

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

Does the GSV ($|S|$) increase or decrease when

- n decreases?
- The length of any $d_i = y_i = \bar{x}_i \mathbf{1}$ increases?
- p decreases?
- The angles (θ) between d_i 's get smaller (i.e., r_{ik} 's get closer to 1)?

$$r_{ik} = \cos(\theta_{ik}) = \frac{d_i d_k}{L_{d_i} L_{d_k}}$$

Draw an example where $\theta_{ik} = 90^\circ$ versus θ_{ik} very small.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation

● Implications for Size of GSV

● Variable-space Interpretation of GSV

- Constant Statistical Distance
- Statistical Distance between two Points

- Rotation of axes
- Back to Interpretation of GSV

- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of

GSV

- Different patterns S , Same
- Example: Geometry and Random Sampling
- The GSV : variances and



Variable-space Interpretation of GSV

In the variable space, we consider the spread of the n points in the p -dimensional space around the sample mean

$$\bar{x}' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

For this, we need some more definitions/concepts:

- The statistical distance of a point $P = (x_1, x_2, \dots, x_p)$ from the origin is the distance of the point

$$P^* = (x_1^*, x_2^*, \dots, x_p^*) = \left(\frac{x_1}{\sqrt{s_{11}}}, \frac{x_2}{\sqrt{s_{22}}}, \dots, \frac{x_p}{\sqrt{s_{pp}}} \right)$$

in standardized coordinates; that is,

$$D(0, P) = \sqrt{x_1^{*2} + x_2^{*2} + \dots + x_p^{*2}}$$

$$= \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \dots + \frac{x_p^2}{s_{pp}}}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV : n -space interpretation
- Implications for Size of GSV

● Variable-space Interpretation of GSV

- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV

GSV

- Different patterns S , Same Sample Geometry and Random Sampling
- The GSV : variances and



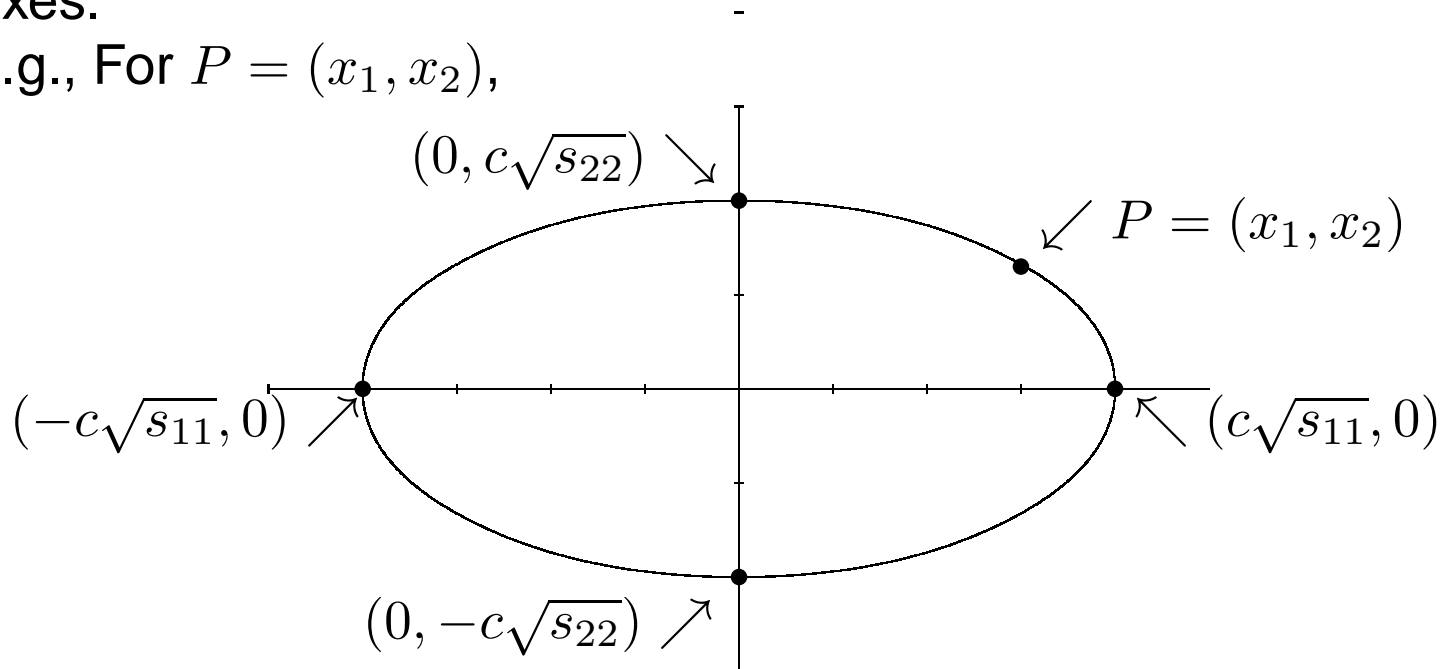
Constant Statistical Distance

All points that have coordinates (x_1, x_2, \dots, x_p) and are a constant (squared) statistical distance from the origin must satisfy

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \dots + \frac{x_p^2}{s_{pp}} = c^2.$$

This is the equation for an ellipsoid with center at the origin (zero) and major and minor axes coinciding with the coordinate axes.

e.g., For $P = (x_1, x_2)$,



- Outline

- Motivation

- Variable Space

- Observation Space

- Obs Space: Mean

- Covariance & Correlation

- Random Sample and Expected Values

- Generalized Variance

 - Generalized Variance
 - Generalized Sample Variance
 - Example: GSV
 - GSV: n -space interpretation
 - Implications for Size of GSV
 - Variable-space Interpretation of GSV
 - **Constant Statistical Distance**
 - Statistical Distance between two Points
 - Rotation of axes
 - Back to Interpretation of GSV
 - GSV Interpretation
 - When does $GSV = 0$
 - Implications & Limitations of GSV

- Different patterns S , Same Sample Geometry and Random Sampling
- The GSV : variances and



Statistical Distance between two Points

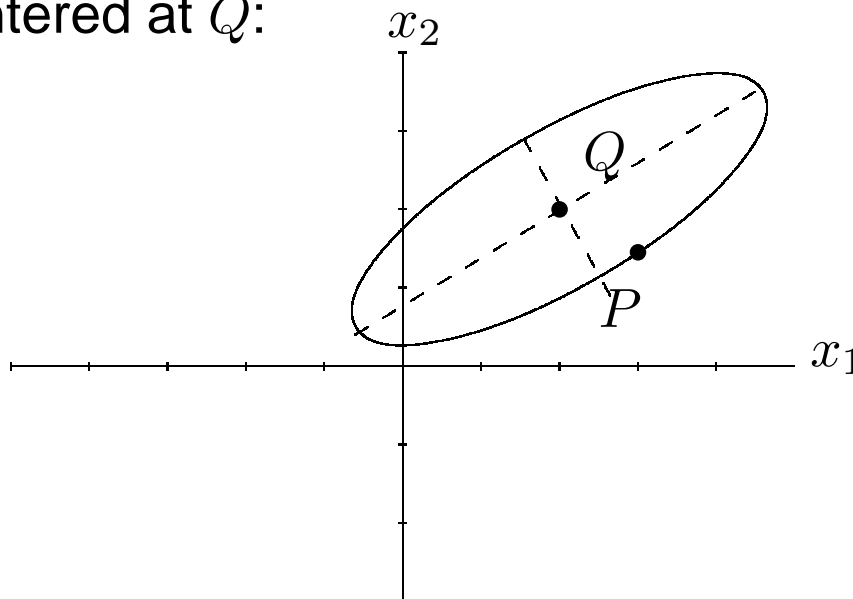
The statistical distance of a point $P = (x_1, x_2)$ from point $Q = (y_1, y_2)$ is

$$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

where a_{11} , a_{12} , and a_{22} are some constants.

For now, we'll suppose that the variables are correlated (i.e., the variable vectors are not at 90° angles).

The square distance $d(P, Q)^2$ is the equation of an ellipse centered at Q :



● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of

GSV

● Different patterns S , Same

Example Geometry and Random Sampling

● The GSV : variances and



Rotation of axes

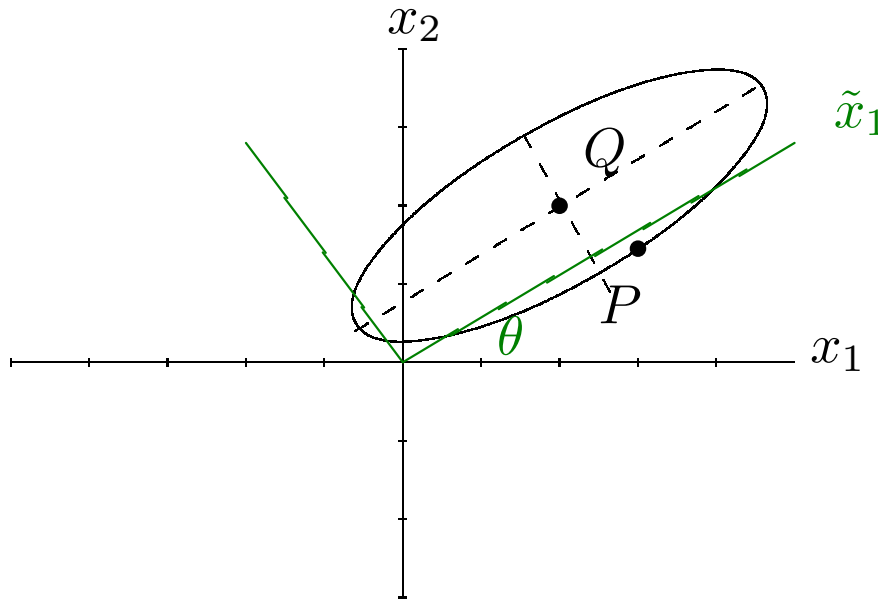
When variables are correlated, we have a rotation of the axes through an angle θ :

$$x_1 \longrightarrow \tilde{x}_1 \quad \text{and} \quad x_2 \longrightarrow \tilde{x}_2$$

Namely,

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$

$$\tilde{x}_2 = x_1 \sin(\theta) + x_2 \cos(\theta)$$



- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
- Random Sample and Expected Values
- Generalized Variance
 - Generalized Variance
 - Generalized Sample Variance
 - Example: GSV \tilde{x}_2
 - GSV: n -space interpretation
 - Implications for Size of GSV
 - Variable-space Interpretation of GSV
 - Constant Statistical Distance
 - Statistical Distance between two Points
 - **Rotation of axes**
 - Back to Interpretation of GSV
 - GSV Interpretation
 - When does $GSV = 0$
 - Implications & Limitations of GSV
- Example: Geometry and Random Sampling
- The GSV : variances and



Back to Interpretation of GSV

The following is an equation of an ellipsoid

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

where x is like point P and \bar{x} is like point Q . (A quadratic form)

The above is the (squared) statistical distance of points x from the point \bar{x} in the p -dimensional space that are a constant distance c from \bar{x} . e.g.,

$$p = 1 : \frac{(x_1 - \bar{x}_1)^2}{s_{11}} = c^2$$

$$p = 2 : a_{11}(x_1 - \bar{x}_1)^2 + 2a_{12}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) + a_{22}(x_2 - \bar{x}_2)^2 = c^2$$

where the a_{11} , a_{12} and a_{22} depend on

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \longrightarrow S^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV : n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV

● GSV Interpretation

● When does $GSV = 0$

● Implications & Limitations of

GSV

● Different patterns S , Same

● Example: Geometry and Random Sampling

● The GSV : variances and

GSV Interpretation

- Outline

- Motivation

- Variable Space

- Observation Space

- Obs Space: Mean

- Covariance & Correlation

- Random Sample and Expected Values

- Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: *GSV*
- *GSV*: n -space interpretation
- Implications for Size of *GSV*
- Variable-space Interpretation of *GSV*
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of *GSV*
- *GSV* Interpretation
- When does $GSV = 0$
- Implications & Limitations of

- GSV*

- Different patterns S , Same
- Example: Geometry and Random Sampling
- The *GSV*: variances and

The Volume of this ellipsoid is

$$\text{volume}\{x : (x - \bar{x})' S^{-1} (x - \bar{x}) \leq c^2\} = (\text{constant}) |S|^{1/2} c^p$$

Note: $|S| = GSV$

So the *GSV* is proportional to the volume of the ellipsoid where the ellipsoid represents statistical distances of observations from the vector of means.



When does $G_{SV} = 0$

Consider matrix of deviations,

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

- $G_{SV} = 0$ means that at least one column of $(\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}})$ can be expressed as a linear combination of the others.
- One (or more) of the deviation vectors d_j lies in the (hyper)plane defined by the others. (Remember of simple example).

The G_{SV} is zero if and only if at least one deviation vector lies in the (hyper)plane formed by all linear combinations of the others; that is, when the columns of the matrix $\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}$ are linearly dependent.

- Outline
- Motivation
- Variable Space
- Observation Space
- Obs Space: Mean
- Covariance & Correlation
- Random Sample and Expected Values
- Generalized Variance
 - Generalized Variance
 - Generalized Sample Variance
 - Example: GSV
 - GSV: n -space interpretation
 - Implications for Size of G_{SV}
 - Variable-space Interpretation of G_{SV}
 - Constant Statistical Distance
 - Statistical Distance between two Points
 - Rotation of axes
 - Back to Interpretation of G_{SV}
 - G_{SV} Interpretation
 - When does $G_{SV} = 0$
 - Implications & Limitations of G_{SV}
- Different patterns S , Same G_{SV}
- Example: Geometry and Random Sampling
- The G_{SV} : variances and



Implications & Limitations of GSV

If the

$$(\text{sample size}) \leq (\text{number of variables})$$

$$n \leq p$$

then $|S| = GSV = 0$ for all samples.

e.g., Example where $p = 4$ and $n = 2$

More on the meaning & interpretation of GSV that we'll elaborate on:

- The GSV represents both variability and covariability.
- You can have very different patterns of variability and association (correlational structure) but have the same value for GSV .

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV : n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV
- Different patterns S , Same GSV
- Example: Geometry and Random Sampling
- The GSV : variances and

Different Patterns S , Same GSV

Problem with $GSV = |S|$ is that different patterns of variability and covariability can give the same value for GSV .

eg.,

$$S_0 = \begin{pmatrix} 3.0 & 0.0 \\ 0.0 & 3.0 \end{pmatrix} \longrightarrow |S_0| = 9.00$$

$$S_1 = \begin{pmatrix} 5.0 & 4.0 \\ 4.0 & 5.0 \end{pmatrix} \longrightarrow |S_1| = 9.00$$

$$S_2 = \begin{pmatrix} 5.0 & -4.0 \\ -4.0 & 5.0 \end{pmatrix} \longrightarrow |S_2| = 9.00$$

Part of the problem is that GSV reflects both variability and covariability.

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV : n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of

GSV

● Different patterns S , Same

GSV

● The GSV : variances and

Sampling



The GSV : variances and covariances

Even if the strength of the linear relationship between variables is constant (i.e., covariance) the GSV could be larger simply by increasing s_{ii} 's.

larger parallelepiped (ellipsoid) \rightarrow larger volume \rightarrow larger GSV .

If you want a measure that reflects only the covariability, then first standardize all of the variables such that they have the same length (i.e., variance):

$$x_{ji}^* = \frac{x_{ji} - \bar{x}_i}{\sqrt{s_{ii}}}$$

The sample covariance matrix for x_{ji}^* 's (i.e., S^*) is the sample correlation matrix $S^* = R$.

GSV of standardized variables = $|R| = \det(R)$.

The residual/deviation vectors of the x^* 's all have length

$$L_{d_i^*} = \sqrt{n - 1}$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV : n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV

GSV

● Different patterns S , Same

Example: Geometry and Random Sampling

● The GSV : variances and



Proof that $L_{d_i^*} = \sqrt{n - 1}$

$$\begin{aligned}
L_{d_i^*}^2 &= \mathbf{d}_i^{*'} \mathbf{d}_i^* \\
&= \left(\frac{(x_{1i} - \bar{x}_i)}{\sqrt{s_{ii}}}, \frac{(x_{2i} - \bar{x}_i)}{\sqrt{s_{ii}}}, \dots, \frac{(x_{ni} - \bar{x}_i)}{\sqrt{s_{ii}}} \right) \begin{pmatrix} \frac{(x_{1i} - \bar{x}_i)}{\sqrt{s_{ii}}} \\ \frac{(x_{2i} - \bar{x}_i)}{\sqrt{s_{ii}}} \\ \vdots \\ \frac{(x_{ni} - \bar{x}_i)}{\sqrt{s_{ii}}} \end{pmatrix} \\
&= \sum_{j=1}^n \frac{(x_{ji} - \bar{x}_i)^2}{s_{ii}} \\
&= \frac{1}{s_{ii}} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \\
&= \frac{1}{\frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \\
&= n - 1
\end{aligned}$$

So $L_{d_i^*} = \sqrt{n - 1}$ which does **not** depend on i .

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV

● Different patterns S , Same

Example: Geometry and Random Sampling

- The GSV: variances and



GSV using R

$|R|$ focuses only on angles between the d_i^* 's.

- When angles = 90° , the d_i^* 's are perpendicular and $|R|$ reaches a maximum: $r_{ik} = \cos(90)/1 = 0$ and $|R| = 1$.
- When angles = 0° , the d_i^* go in the same direction and $|R|$ reaches a minimum: $r_{ik} = \cos(0)/1 = \pm 1$ and $|R| = 0$

e.g.,

$$R_0 = \begin{pmatrix} 1.0 & .0 \\ .0 & 1.0 \end{pmatrix} \longrightarrow |R_0| = 1.00$$

$$R_1 = \begin{pmatrix} 1.0 & .8 \\ .8 & 1.0 \end{pmatrix} \longrightarrow |R_1| = 0.36$$

$$R_2 = \begin{pmatrix} 1.0 & -.8 \\ -.8 & 1.0 \end{pmatrix} \longrightarrow |R_2| = \text{_____}$$

$$R_3 = \begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix} \longrightarrow |R_3| = 0.00$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV
- Different patterns S , Same Sample Geometry and Random Sampling
- The GSV : variances and



Relationship between $\det(S)$ and $\det(R)$

The two ways of computing the GSV (one on unstandardized and the other on standardized variables) are functionally related:

$$|S| = (s_{11}s_{22} \cdots s_{pp})|R| = \prod_{i=1}^p s_{ii}|R|$$

This emphasizes the fact that $|S|$ depends on the s_{ii} 's but $|R|$ doesn't.

e.g.,

$$|S_1| = 9 = (5)(5)0.36 = 9$$

and

$$|S_0| = (3)(3)|R| = (3)(3)(1) = 9$$

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV
- Different patterns S , Same GSV
- Example: Geometry and Random Sampling
- The GSV : variances and



Another Global Statistic

● Outline

Motivation

Variable Space

Observation Space

Obs Space: Mean

Covariance & Correlation

Random Sample and Expected Values

Generalized Variance

- Generalized Variance
- Generalized Sample Variance
- Example: GSV
- GSV: n -space interpretation
- Implications for Size of GSV
- Variable-space Interpretation of GSV
- Constant Statistical Distance
- Statistical Distance between two Points
- Rotation of axes
- Back to Interpretation of GSV
- GSV Interpretation
- When does $GSV = 0$
- Implications & Limitations of GSV
- Different patterns S , Same Sample Geometry and Random Sampling
- The GSV : variances and

While $GSV = |\mathbf{R}|$ focuses just on covariability, there is another measure that focuses just summarizing the variance in S :

$$\text{“Total Sample Variance”} = \sum_{i=1}^p s_{ii} = \text{trace}(\mathbf{S}) = \text{tr}(\mathbf{S})$$

We'll use this one later when we talk about principal components analysis.