Sample Geometry

Edps/Soc 584, Psych 594

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Spring 2017

Outline

- Motivation
- Variable Space
- Observation Space
  - What it is.
  - Mean
  - Covariance
- Global summary statistic for $\boldsymbol{S}$

Reading:

- Johnson & Wichern, Chapter 3

Motivation

- The sample $\bar{\boldsymbol{x}}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ have geometric interpretations.
- The geometric interpretation shows how $\bar{\boldsymbol{x}}$, $\boldsymbol{S}$, and $\boldsymbol{R}$ are related to $\boldsymbol{X}_{n\times p}$.
- 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., $\bar{\boldsymbol{x}}$, $\boldsymbol{S}$, and $\boldsymbol{R}$) 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.

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

- (2) The joint distribution of all p variables is the same for all cases (individuals, rows). We are drawing multivariate observations from the same unchanging population.

The Variable Space

- Each row of $\boldsymbol{X}$ is a point in "p-space" or variable space.
- The variables (columns of $\boldsymbol{X}$) define the axes.
- Consider the n points in the p dimensional space.

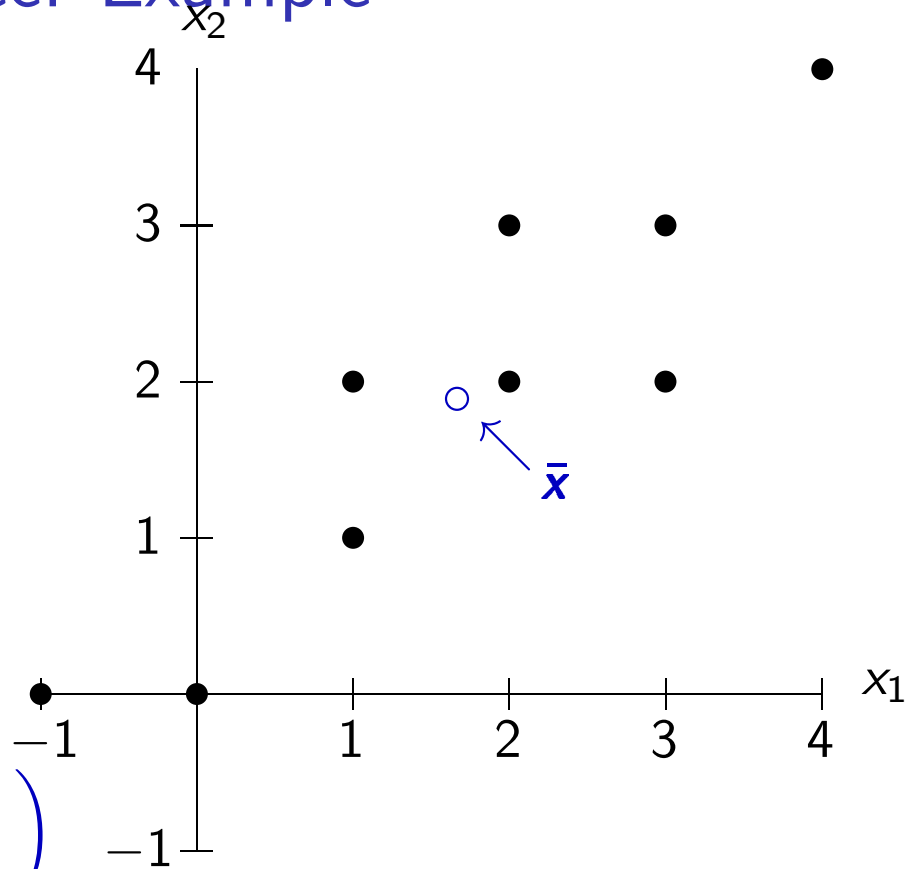
The center of the point "cloud" is $\bar{\boldsymbol{x}}$.

The variability and covariability is measured by $\boldsymbol{S}$.

The Variable Space: Example

The data

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



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

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

The Observation Space

Important alternative geometric representation of $\boldsymbol{X}$ 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 $\boldsymbol{X}$ is a vector in the n-space.

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

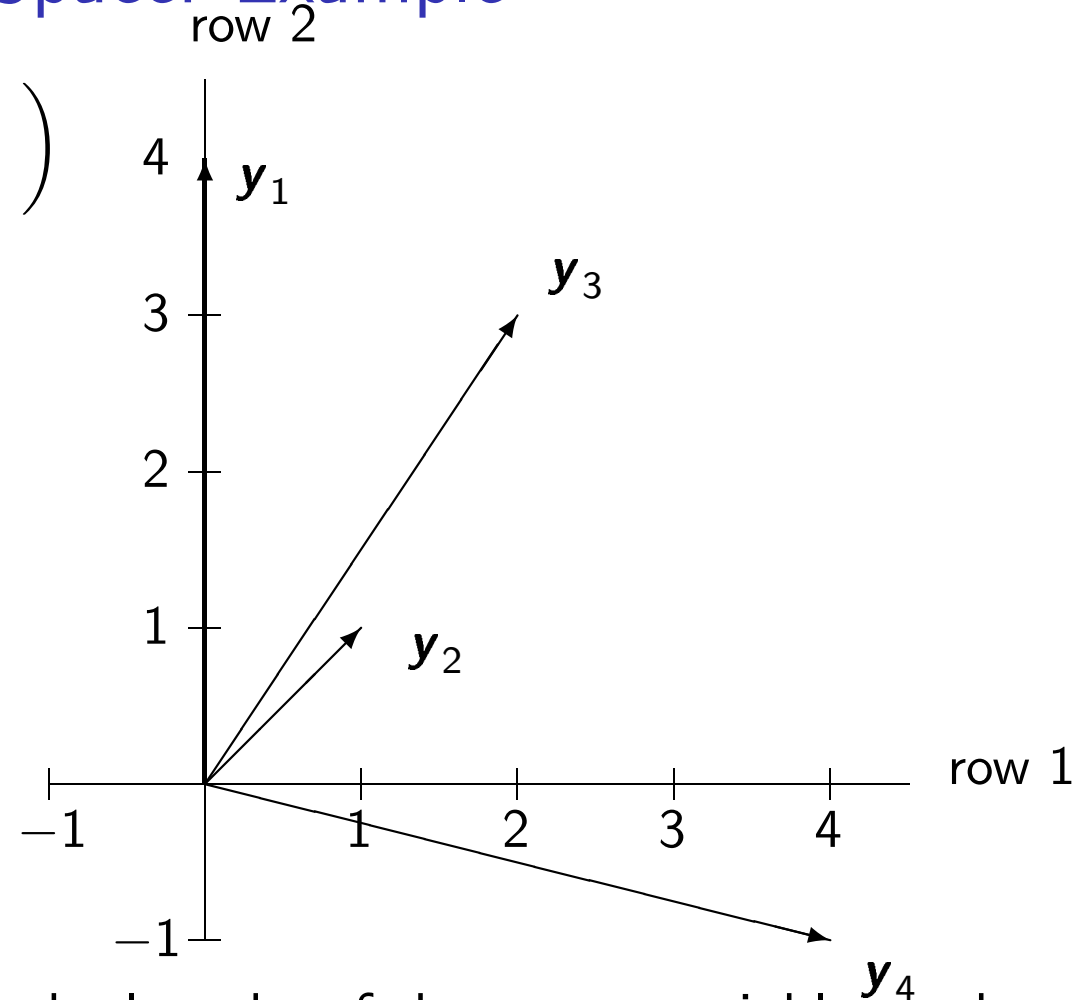
where

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

The Observation Space: Example

The data

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



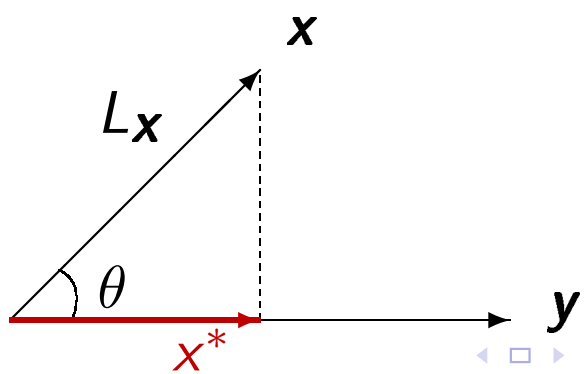
In this space, we talk about the lengths of the vector variables and about the angle between them.

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 $\boldsymbol{x}$ onto $\boldsymbol{y}$ equals

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



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 $\boldsymbol{y}_k' = (x_{1k}, x_{2k}, \ldots, x_{nk})$ on the the vector $\mathbf{1}_n =$

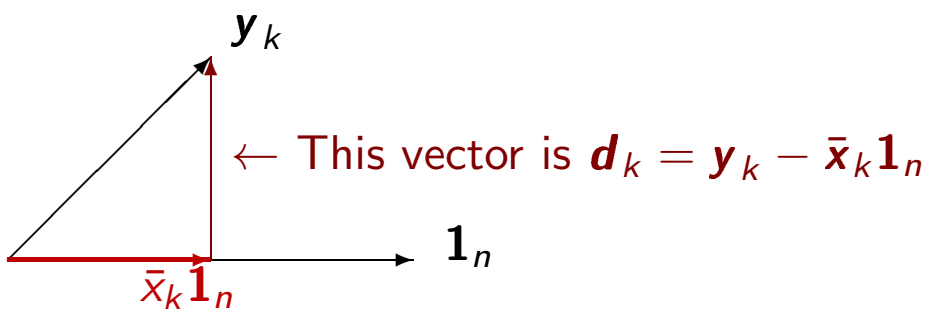
$$\left(\frac{\boldsymbol{y}_k'\mathbf{1}_n}{\mathbf{1}_n'\mathbf{1}_n}\right)\mathbf{1} = \frac{(x_{1k}, x_{2k}, \ldots, x_{nk}) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}{\underbrace{(1+1+\cdots+1)}_{n}}\mathbf{1}_n$$

$$= \frac{\sum_{j=1}^{n} x_{jk}}{n}\mathbf{1}_n = \bar{x}_k\mathbf{1}_n$$

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 $\boldsymbol{y}_k$ onto the vector $\mathbf{1}_n$.

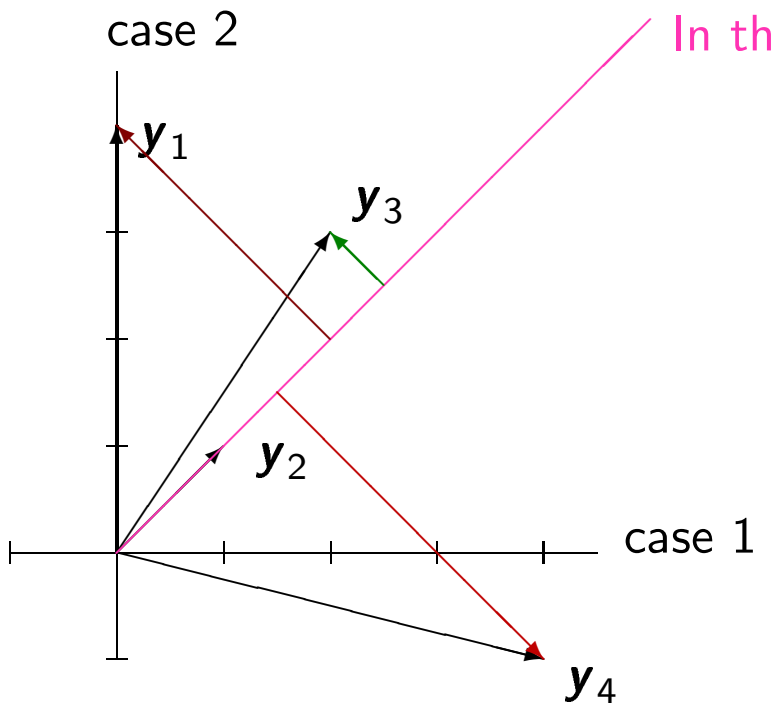
- Projection of $\boldsymbol{y}_k$ onto $\mathbf{1}_n$:



- $\boldsymbol{d}_k$ is the Deviation vector.
- We have the decomposition of $\boldsymbol{y}_k$ into two parts:

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

Back to our little example



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

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

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

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

2-dimensional example continued

For our 2-case example with 4 variables

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

Note:

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

What is the rank of this correlation matrix?

Deviation Vectors: Variances & Covariance

- Sample variance: The squared lengths of deviation vectors,

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

are sums of squared deviation from the mean , so

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

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

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

are sums of cross products , so

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

Deviation Vectors: Correlation

The angle between two deviation vectors equals

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

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

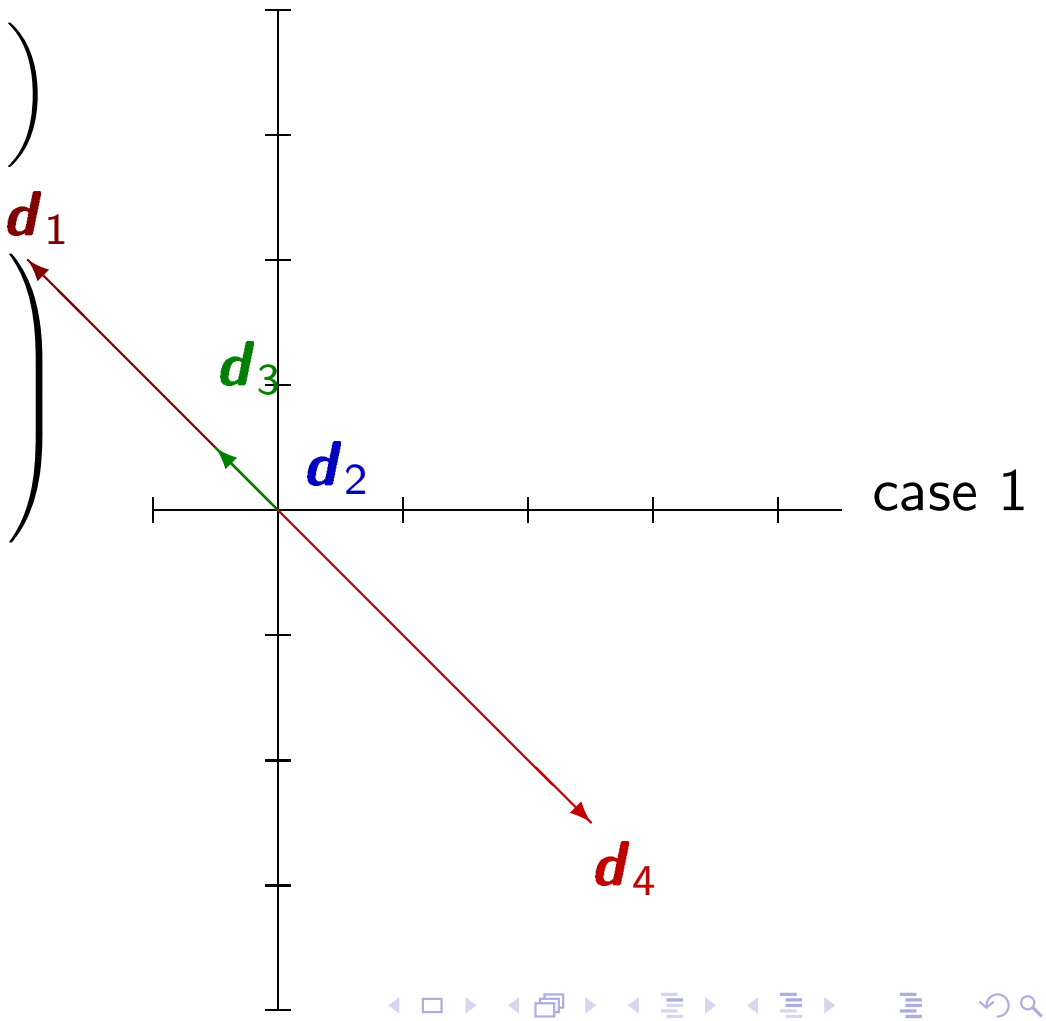
- If two deviation vectors have the same orientation (same direction), $\theta = 0^\circ \longrightarrow \cos(\theta) = 1 = r_{ik}$.
- If two deviation vectors have the opposite orientation, $\theta = 180^\circ \longrightarrow \cos(\theta) = -1 = r_{ik}$.
- If two deviation vectors are perpendicular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

Back to simple example (pages 12-13)

Correlation: Simple Example

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

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



3-Dimensional Example

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

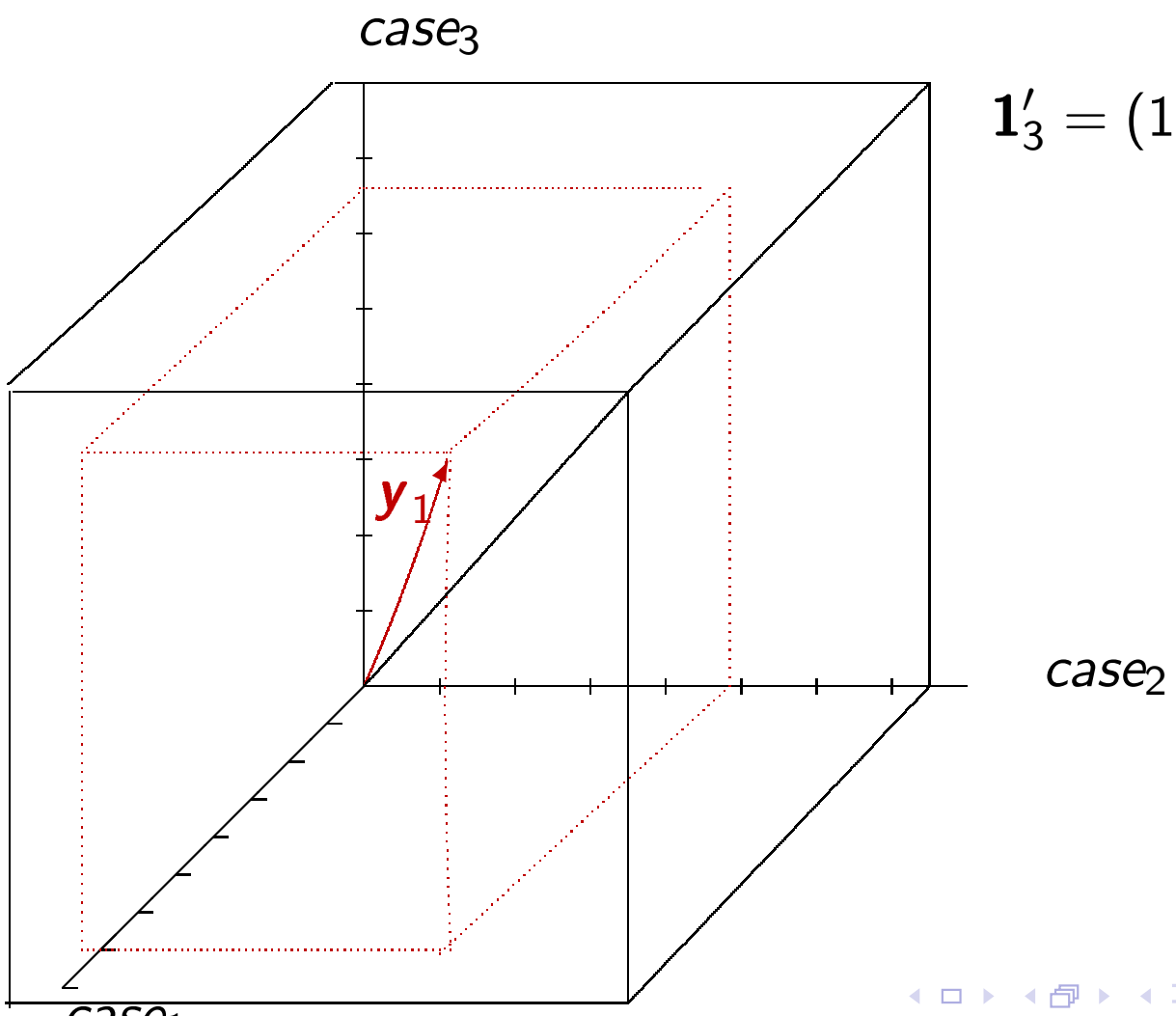
$$\boldsymbol{X} = \begin{pmatrix} 700 & 650 & 750 \\ 500 & 550 & 640 \\ 660 & 750 & 650 \end{pmatrix}$$

So

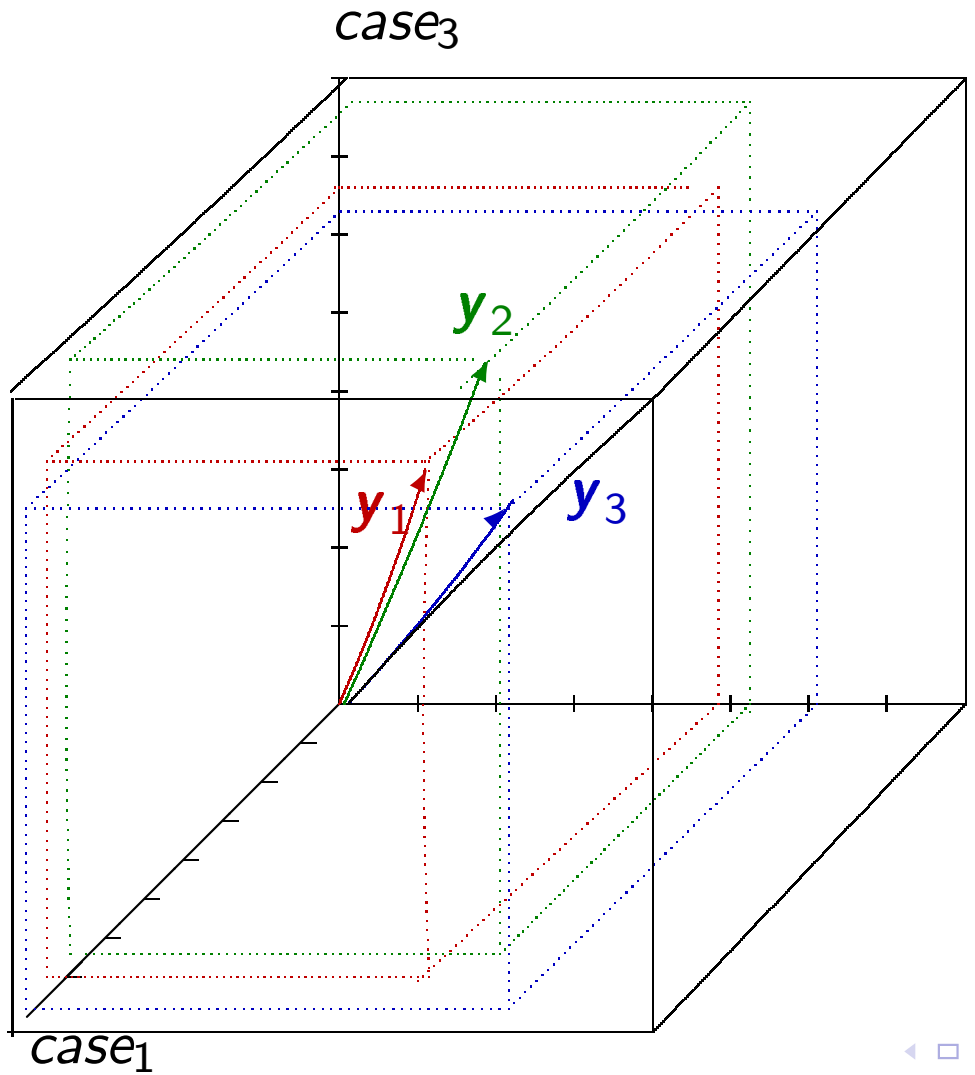
$$\boldsymbol{y}_1 = \begin{pmatrix} 700 \\ 500 \\ 660 \end{pmatrix} \qquad \boldsymbol{y}_2 = \begin{pmatrix} 650 \\ 550 \\ 750 \end{pmatrix} \qquad \boldsymbol{y}_3 = \begin{pmatrix} 750 \\ 640 \\ 650 \end{pmatrix}$$

$$\bar{x}_1 = \frac{1860}{3} = 620 \qquad \bar{x}_2 = \frac{1950}{3} = 650 \qquad \bar{x}_3 = \frac{2040}{3} = 680$$

3-Dimensional Example: Variable 1



3-Dimensional Example: Graph of Variables



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

3-Dimensional Example

Means are the multiples of $\mathbf{1}$ that give projection of $\boldsymbol{y}_i$ onto

Using the first variable for illustration

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

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

- $\boldsymbol{d}_1$ is perpendicular to $\bar{x}_1$ as it should be:

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

Deviation Vectors

Notes (continued):

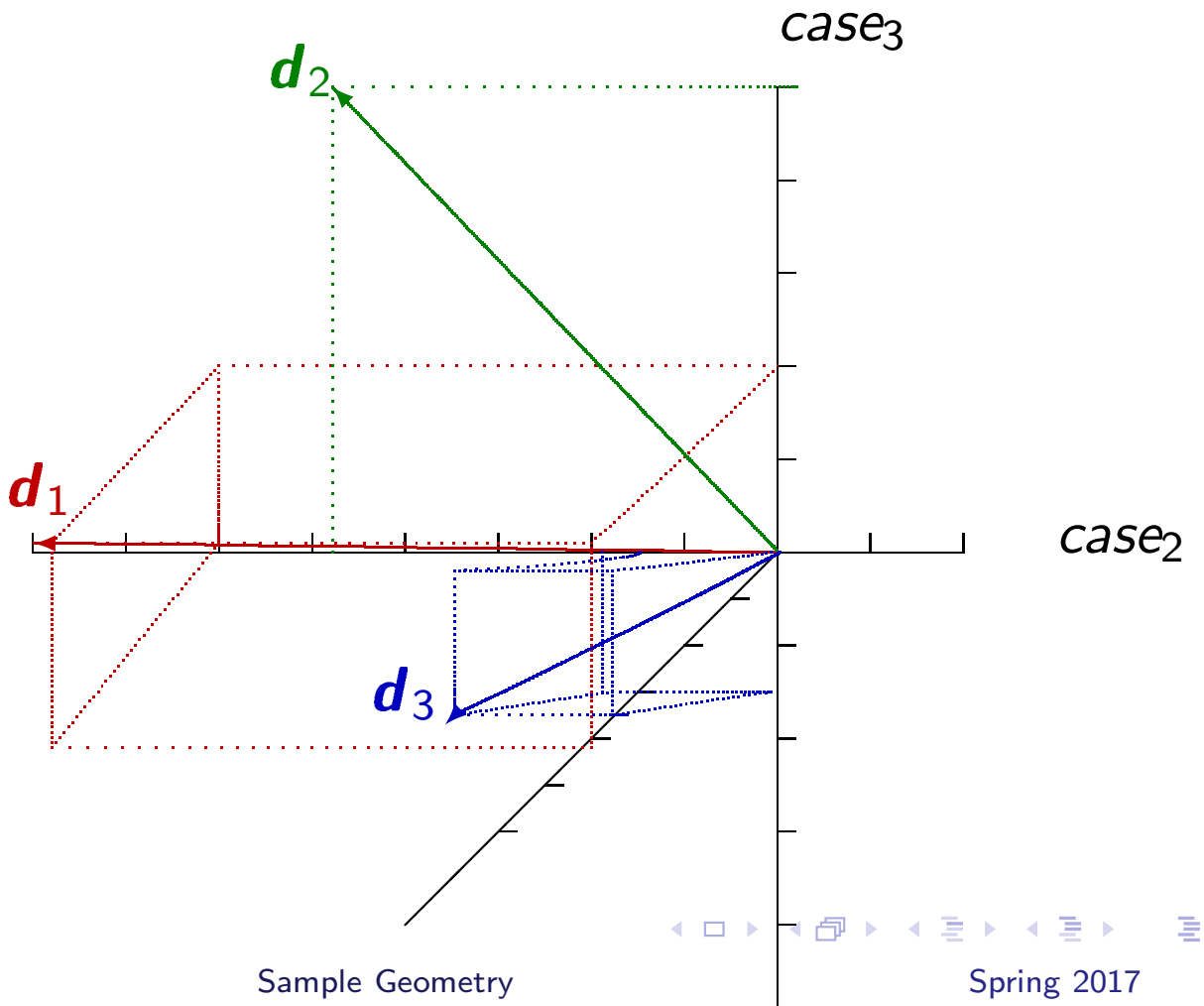
- $\boldsymbol{y}_1$ is decomposed into two parts:

$$\boldsymbol{y}_1 = \bar{x}_1 \mathbf{1}_n + \boldsymbol{d}_1$$

- If $\boldsymbol{d}_i = \mathbf{0}$, then $\boldsymbol{y}_i = \bar{x}_i \mathbf{1}$.
- All of them

$$\boldsymbol{d}_1 = \begin{pmatrix} 80 \\ -120 \\ 40 \end{pmatrix} \qquad \boldsymbol{d}_2 = \begin{pmatrix} 0 \\ -100 \\ 100 \end{pmatrix} \qquad \boldsymbol{d}_3 = \begin{pmatrix} 70 \\ -40 \\ -30 \end{pmatrix}$$

Graph of 3 deviation vectors



Variances

Variance are proportional to the lengths of deviation vectors

Variable 1:

$$L_{\boldsymbol{d}_1} = \sqrt{80^2 + 120^2 + 40^2} = \sqrt{22400} = 149.67 = \sqrt{ns_{11}} = \sqrt{3s_{11}}$$

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

Variable 2:

$$L_{\boldsymbol{d}_2} = \sqrt{0^2 + 100^2 + 100^2} = 100\sqrt{2} = 141.42 = \sqrt{3s_{22}}$$

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

Variable 3:

$$L_{\boldsymbol{d}_3} = \sqrt{70^2 + 40^2 + 30^2} = \sqrt{7400} = 86.02 = \sqrt{3s_{33}}$$

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

Correlations

Correlations are the cosines of the angles between deviation vectors

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

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

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

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

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

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

Summary: Observation Space

- The axes of the observation space are defined by cases (individuals) and variables are represented as vectors.
- The projection of the column $\boldsymbol{y}_k$ of the data matrix $\boldsymbol{X}_{n\times p}$ onto the equal angular vector $\mathbf{1}_n$ is the vector $\bar{x}_k\mathbf{1}$.
- The information contained in $\boldsymbol{S}$ is obtained from the deviation vectors: $\boldsymbol{d}_k = \boldsymbol{y}_k - \bar{x}_k\mathbf{1}_n = \{(x_{jk} - \bar{x}_k)\}$

  The variance: $L^2_{\boldsymbol{d}_k} = \boldsymbol{d}'_k\boldsymbol{d}_k = ns_{kk}$

  The covariance: $\boldsymbol{d}'_i\boldsymbol{d}_k = ns_{ik}$
- The sample correlation coefficient r_{ik} is the cosine of the angle between $\boldsymbol{d}_i$ and $\boldsymbol{d}_k$.

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.

DATA

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

where

$$\boldsymbol{X}_j' = (X_{j1}, X_{j2}, \ldots, X_{jp})$$

Each $\boldsymbol{X}_j'$ 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 $\boldsymbol{X}_j$,

$$f(\boldsymbol{x}_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$$

Important Result

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

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

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

Estimating Covariance Matrix

To estimate $\boldsymbol{\Sigma}$, first note

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

So

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

is an unbiased estimator of $\boldsymbol{\Sigma}$.

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

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,need p variances and $p(p-1)/2$ covariances.
- It would be nice to have a single statistic that summarizes the information in $\boldsymbol{S}$ (i.e., reflects all the variances and covariance).

Generalized Sample Variance

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

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

$$\begin{aligned} |\boldsymbol{A}| &= a_{11} && \text{for } k = 1 \\ |\boldsymbol{A}| &= a_{11}a_{22} - a_{12}a_{21} && \text{for } k = 2 \\ |\boldsymbol{A}| &= \sum_{j=1}^{k} a_{ij}|A_{ij}|(-1)^{i+j} && \text{for } k > 1 \end{aligned}$$

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

Example: GSV

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

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

What does this mean?

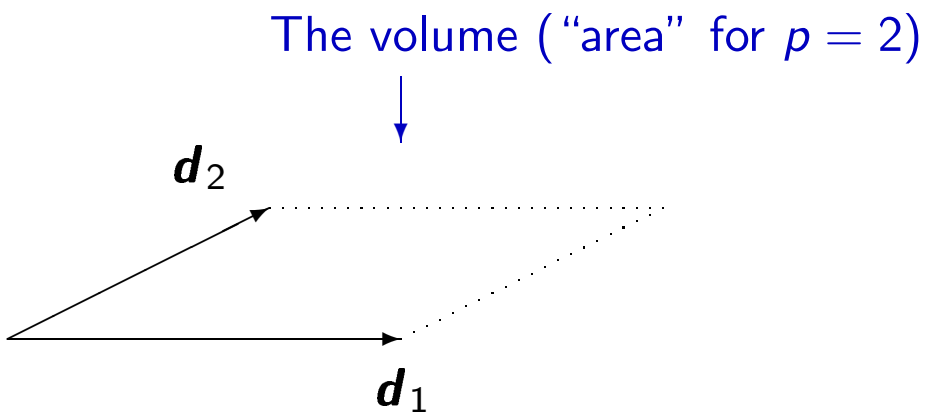
Two ways to interpret it (geometrically):

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

GSV: n-space interpretation

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

e.g., $p = 2$ and $n =$ "whatever"



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

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

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

Implications for Size of *GSV*

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

Does the *GSV* ($|\boldsymbol{S}|$) increase or decrease when

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

$$r_{ik} = \cos(\theta_{ik}) = \frac{\boldsymbol{d}_i \boldsymbol{d}_k}{L_{\boldsymbol{d}_i} L_{\boldsymbol{d}_k}}$$

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

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

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

For this, we need some more definitions/concepts:

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

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

  in standardized coordinates; that is,

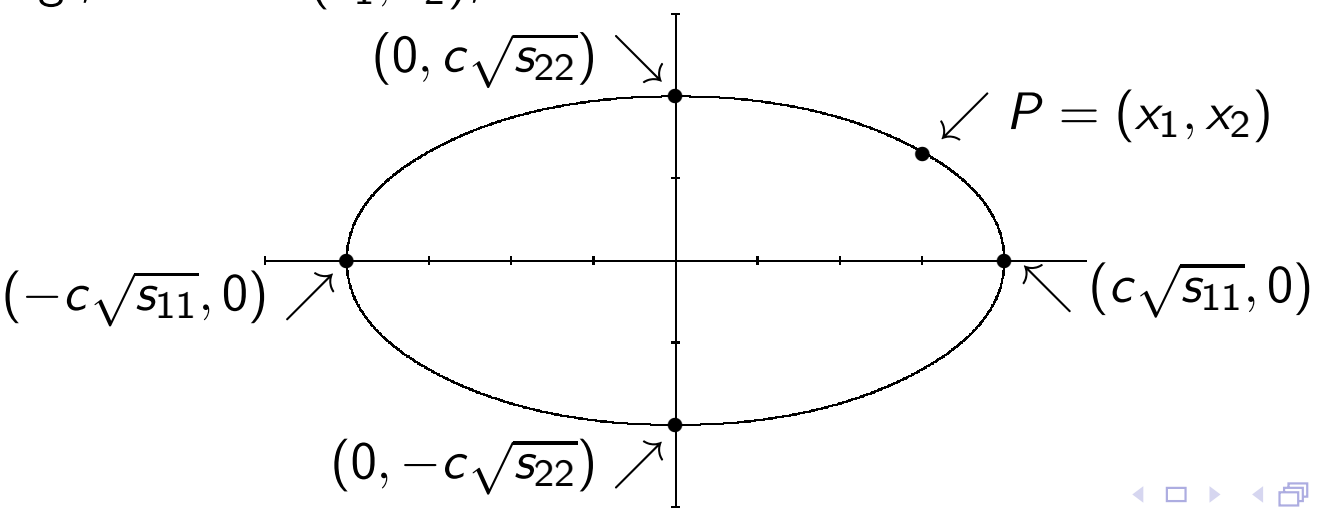
$$\begin{aligned} D(0,P) &= \sqrt{x_1^{*2} + x_2^{*2} + \cdots + x_p^{*2}} \\ &= \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} + \cdots, + \frac{x_p^2}{s_{pp}}} \end{aligned}$$

Constant Statistical Distance

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

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

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 $P = (x_1, x_2)$,



Statistical Distance between two Points

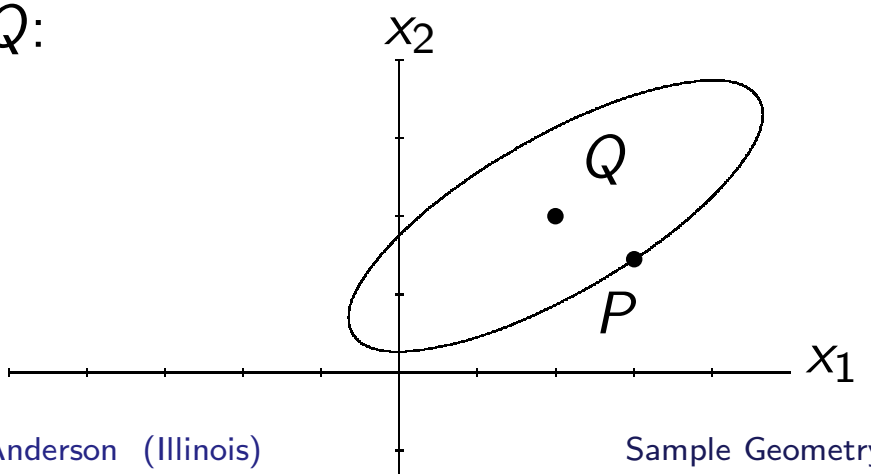
The statistical distance of a point $P = (x_1, x_2)$ from point $Q = (y_1, y_2)$ is

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

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:



Statistical Distance between two Points

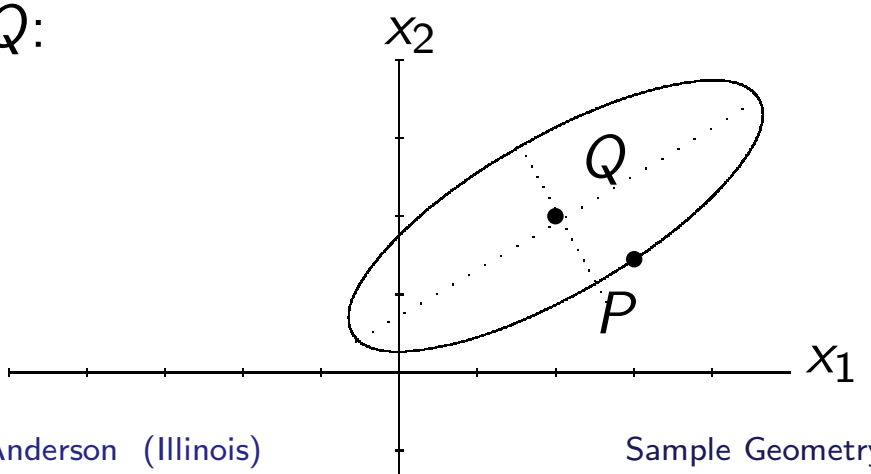
The statistical distance of a point $P = (x_1, x_2)$ from point $Q = (y_1, y_2)$ is

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

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:



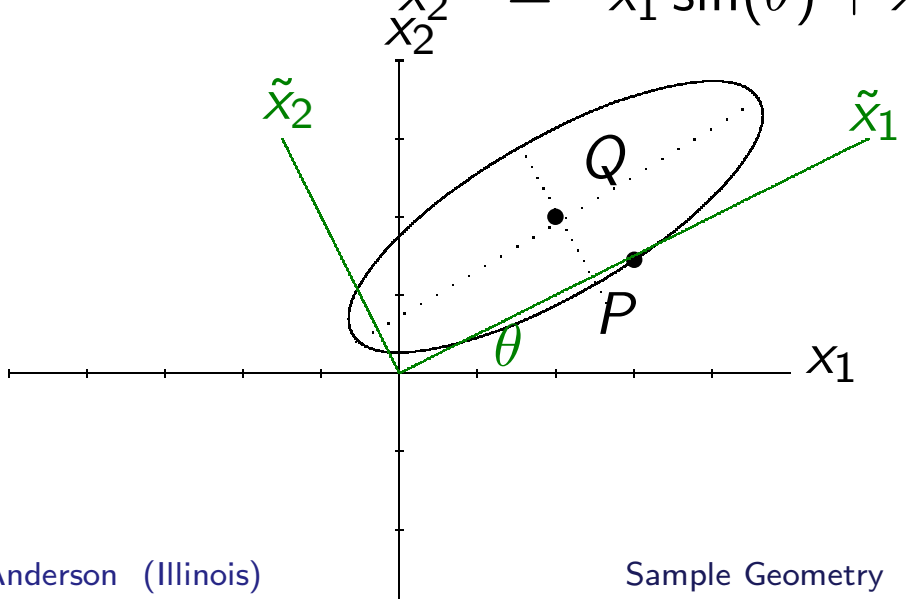
Rotation of axes

When variable are correlated, we have a rotation of the axes through and angle θ:

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

Namely,

$$\begin{aligned} \tilde{x}_1 &= x_1 \cos(\theta) + x_2 \sin(\theta) \\ \tilde{x}_2 &= x_1 \sin(\theta) + x_2 \cos(\theta) \end{aligned}$$



Back to Interpretation of *GSV*

The following is an equation of an ellipsoid

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

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

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

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

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

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

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

GSV Interpretation

The Volume of this ellipsoid is

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

Note: $|\boldsymbol{S}| = GSV$

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 $GSV = 0$

Consider matrix of deviations,

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

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

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

Implications & Limitations of *GSV*

If the

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

$$n \leq p$$

then $|\boldsymbol{S}| = GSV = 0$ 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*.

Different Patterns S, Same GSV

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

eg.,

$$\boldsymbol{S}_0 = \begin{pmatrix} 3.0 & 0.0 \\ 0.0 & 3.0 \end{pmatrix} \longrightarrow |\boldsymbol{S}_0| = 9.00$$

$$\boldsymbol{S}_1 = \begin{pmatrix} 5.0 & 4.0 \\ 4.0 & 5.0 \end{pmatrix} \longrightarrow |\boldsymbol{S}_1| = 9.00$$

$$\boldsymbol{S}_2 = \begin{pmatrix} 5.0 & -4.0 \\ -4.0 & 5.0 \end{pmatrix} \longrightarrow |\boldsymbol{S}_2| = 9.00$$

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

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 be increasing s_{ii}'s.

larger parellepiped (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., $\boldsymbol{S}^*$) is the sample correlation matrix $\boldsymbol{S}^* = \boldsymbol{R}$.

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

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

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

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

GSV using *R*

$|\boldsymbol{R}|$ focuses only on angles between the $\boldsymbol{d}_i^*$'s.

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

e.g.,

$$\boldsymbol{R}_0 = \begin{pmatrix} 1.0 & .0 \\ .0 & 1.0 \end{pmatrix} \longrightarrow |\boldsymbol{R}_0| = 1.00$$

$$\boldsymbol{R}_1 = \begin{pmatrix} 1.0 & .8 \\ .8 & 1.0 \end{pmatrix} \longrightarrow |\boldsymbol{R}_1| = 0.36$$

$$\boldsymbol{R}_2 = \begin{pmatrix} 1.0 & -.8 \\ -.8 & 1.0 \end{pmatrix} \longrightarrow |\boldsymbol{R}_2| = \underline{\quad\quad}$$

$$\boldsymbol{R}_3 = \begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix} \longrightarrow |\boldsymbol{R}_3| = 0.00$$

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:

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

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

e.g.,

$$|\boldsymbol{S}_1| = 9 = (5)(5)0.36 = 9$$

and

$$|\boldsymbol{S}_0| = (3)(3)|\boldsymbol{R}| = (3)(3)(1) = 9$$

Another Global Statistic

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

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

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