Geometría de la Muestra y Muestreo Aleatorio

Valores Esperados de la Media Muestral y la Matriz de Covarianza
Varianza Generalizada

Santiago Alférez

Agosto de 2020

MACC
Universidad del Rosario

Contenidos

Valores Esperados de la Media Muestral y la Matriz de Covarianza

Valores esperados de la media y la covarianza muestrales

Consideraciones de los valores esperados

- Si las n componentes no son independientes o las distribuciones marginales no son idénticas, la influencia de las medidas individuales (coordenadas) sobre la localización es asimétrica.
- Entonces, podríamos considerar usar una distancia en la cuál las coordenadas fueran ponderadas de forma desigual, cómo en la distancia estadística o la forma cuadrática.
- Se puede ver cómo $\overline{\mathbf{X}}$ y $\mathbf{S}_n$ se comportan como estimadores puntuales del vector de media poblacional correspondiente $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$.

Valores esperados de la media y la covarianza muestrales

Estimador para μ

Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ una muestra aleatoria con una distribución conjunta que tiene vector media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$. Entonces, $\overline{\mathbf{X}}$ es un estimador insesgado de $\boldsymbol{\mu}$ y su matriz de covarianza es:

$$\frac{1}{n}\boldsymbol{\Sigma}.$$

$$E(\overline{\mathbf{X}}) = \boldsymbol{\mu} \qquad \text{(vector de media poblacional)}$$

$$\mathrm{Cov}(\overline{\mathbf{X}}) = \tfrac{1}{n}\boldsymbol{\Sigma} \qquad \left(\begin{array}{c}\text{matriz de covarianza poblacional}\\ \text{dividida por el tamaño de la muestra}\end{array}\right)$$

Valores esperados de la media y la covarianza muestrales

Estimador para $\boldsymbol{\Sigma}$

Para la matriz de covarianza $\mathbf{S}_n$,

$$E\left(\mathbf{S}_n\right) = \frac{n-1}{n}\boldsymbol{\Sigma} = \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma} \quad \text{o,} \quad E\left(\frac{n}{n-1}\mathbf{S}_n\right) = \boldsymbol{\Sigma}$$

Entonces, $[n/(n-1)]\mathbf{S}_n$ es un estimador insesgado de $\boldsymbol{\Sigma}$, mientras que $\mathbf{S}_n$ es un estimador sesgado con $sesgo = E\left(\mathbf{S}_n\right) - \boldsymbol{\Sigma} = -(1/n)\boldsymbol{\Sigma}$

Matriz de covarianza muestral insesgada

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

Puesto que $\mathbf{S}$ se usa generalmente para pruebas multivariables, $\mathbf{S}_n$ será sustituida por $\mathbf{S}$.

Varianza Generalizada

Varianza Generalizada

Matriz de covarianzas (y varianzas)

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

Contiene p varianzas y $\frac{1}{2}p(p-1)$ covarianzas.

Varianza muestral generalizada

Es deseable un sólo número para expresar la variación descrita por $\mathbf{S}$:

$$\text{Generalized sample variance} = |\mathbf{S}|$$

Varianza Generalizada

Ejemplo

Para la siguiente matriz de datos:

$$X = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$$

calcule la varianza muestral generalizada.

Ejercicio

Para la siguiente matriz de datos:

$$X = \begin{bmatrix} 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$$

calcule la varianza muestral generalizada.

Varianza muestral generalizada

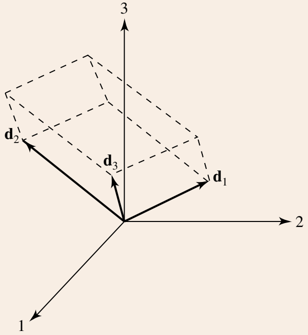
Interpretación geométrica de la varianza generalizada

La varianza muestral generalizada, para un conjunto de datos fijo, es proporcional al cuadrado del volumen generado por los p vectores de desviación $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$ para $i = 1, \ldots, p$ en el espacio n-dimensional. Es decir,

$$\text{Varianza muestral generalizada} = |\mathbf{S}| = (n-1)^{-p}\ (\text{volumen})^{\,2}.$$

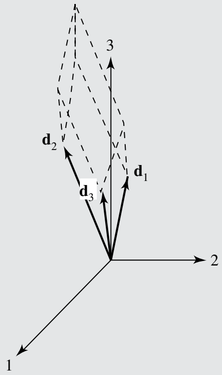
Varianza muestral generalizada

Varianza generalizada grande



- La varianza generalizada aumenta si la longitud de cualquier $\mathbf{d}_i$ se incrementa.
- Se incrementa si los vectores residuales se mueven hasta que están a ángulos rectos unos de otros.

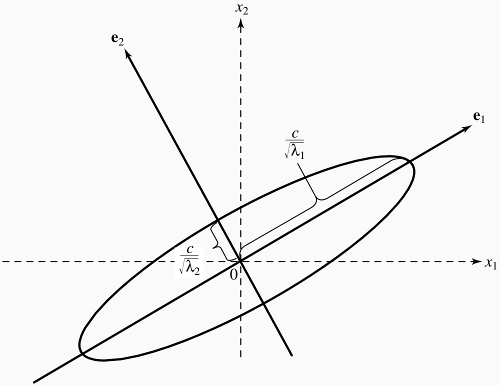
Varianza generalizada pequeña



- El volumen disminuye si uno de los $s_i i$ es pequeño o si uno de los vectores de desviación se encuentra cerca al (hiper) plano formado por los otros.

Recordando la interpretación geométrica de la distancia

- La distancia es determinada a partir de una forma cuadrática definida positiva $\mathbf{x}'\mathbf{A}\mathbf{x}$.
- El cuadrado de la distancia desde $\mathbf{x}$ a un punto fijo arbitrario $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p]$ es dado por la expresión $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})$
- los puntos a la distancia c se encuentran en una elipse cuyos ejes están dados por los vectores propios de $\mathbf{A}$ con longitudes proporcionales a los recíprocos de las raíces cuadradas de los valores propios.



Interpretación geométrica de la varianza generalizada

- La varianza generalizada se puede interpretar en el espacio p-dimensional de los datos.
- La interpretación más intuitiva se refiere a la dispersión alrededor del punto medio muestral $\overline{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$.
- Si $\overline{\mathbf{x}}$ es un punto fijo $\boldsymbol{\mu}$ y $\mathbf{S}^{-1}$ hace el papel de $\mathbf{A}$ (la matriz de los coeficientes de la distancia). Las coordenadas $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ de los puntos a una distancia constante c de $\overline{\mathbf{x}}$ satisfacen:

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

- Cuando $p = 1, (\mathbf{x} - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) = (x_1 - \bar{x}_1)^2 / s_{11}$ es la distancia al cuadrado de x_1 a $\bar{x}_1$ en unidades de desviación estándar.

Varianza muestral generalizada

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

Interpretación geométrica de la varianza generalizada

- La ecuación anterior define un hiperelipsoide (una elipse con $p = 2$) centrada en $\overline{\mathbf{x}}$.
- Se puede mostrar (mediante cálculo integral) que el volumen de éste hiperelipsoide está relacionado al $|\mathbf{S}|$.
- Volumen de $\left\{\mathbf{x} : (\mathbf{x} - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) \leq c^2\right\} = k_p|\mathbf{S}|^{1/2}c^p$
- (Volumen del elipsoide) $^2 =$ (constante) (varianza muestral generalizada)
- Un volumen grande corresponde a una varianza generalizada grande.

Ejemplo de la Varianza Muestral Generalizada (Interpretación Geométrica)

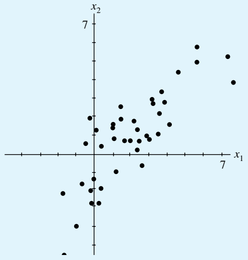
Varianza muestral generalizada

Ejemplo: interpretación de la varianza generalizada

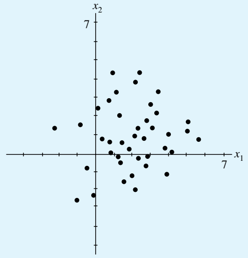
Se tienen tres datasets, cada uno con la misma media $\overline{\mathbf{x}} = [2, 1]$ y las siguientes matrices de covarianza (y la correlación) calculada:

$$\mathbf{S}_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, r_1 = .8 \quad \mathbf{S}_2 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, r_2 = 0 \quad \mathbf{S}_3 = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, r_3 = -.8$$

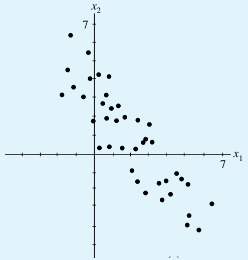
1



2



3



Varianza muestral generalizada: continuando con el ejemplo

Desarrollo del ejemplo

- Cada matriz de covarianza tiene información sobre la variabilidad de las variables y también información para calcular el coeficiente de correlación.
- Entonces, $\mathbf{S}$ captura la orientación y el tamaño del patrón de dispersión.

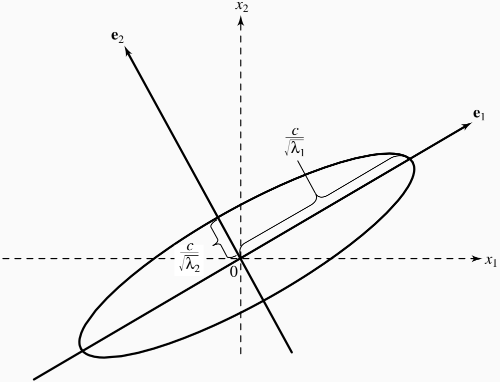
1

Los valores propios y vectores propios obtenidos de $\mathbf{S}$ describen la forma del diagrama de dispersión. Se pueden determinar los valores propios (y luego los vectores propios) de $\mathbf{S}_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ mediante la ecuación característica, obteniéndose

$$\lambda_1 = 9, \mathbf{e}_1' = [1/\sqrt{2}, 1/\sqrt{2}] \quad \text{y} \quad \lambda_2 = 1, \mathbf{e}_2' = [1/\sqrt{2}, -1/\sqrt{2}]$$

Recordando la interpretación geométrica de la distancia

- La distancia es determinada a partir de una forma cuadrática definida positiva $\mathbf{x}'\mathbf{A}\mathbf{x}$.
- El cuadrado de la distancia desde $\mathbf{x}$ a un punto fijo arbitrario $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p]$ es dado por la expresión $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})$
- los puntos a la distancia c se encuentran en una elipse cuyos ejes están dados por los vectores propios de $\mathbf{A}$ con longitudes proporcionales a los recíprocos de las raíces cuadradas de los valores propios.



Varianza muestral generalizada: continuando con el ejemplo

1

$\mathbf{S}_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ $\lambda_1 = 9, \mathbf{e}_1' = [1/\sqrt{2}, 1/\sqrt{2}]$; $\lambda_2 = 1, \mathbf{e}_2' = [1/\sqrt{2}, -1/\sqrt{2}]$

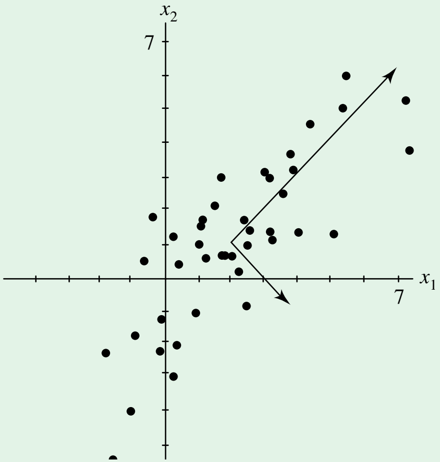
- La elipse centrada en la media $\overline{\mathbf{x}} = [2, 1]$ (para los tres casos) es:
$$(\mathbf{x} - \overline{\mathbf{x}})' \mathrm{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \leq c^2$$
- Podemos describir esta elipse mediante una forma cuadrática $(\mathbf{x} - \boldsymbol{\mu})' \mathbf{A} (\mathbf{x} - \boldsymbol{\mu})$ con $\mathbf{A} = \mathbf{S}^{-1}$ y $\mu = \overline{\mathbf{x}}$.
- Si $\mathbf{S}\mathbf{e} = \lambda \mathbf{e}$, entonces multiplicar a la izquierda por $\mathbf{S}^{-1}$ da $\mathbf{S}^{-1}\mathbf{S}\mathbf{e} = \lambda \mathbf{S}^{-1}\mathbf{e}$, o $\mathbf{S}^{-1}\mathbf{e} = \lambda^{-1}\mathbf{e}$.
- Por lo tanto, $(\lambda, \mathbf{e})$ es un valor/vector propio para $\mathbf{S}$, entonces $\left(\lambda^{-1}, \mathbf{e}\right)$ es un valor/vector propio para $\mathbf{S}^{-1}$.
- Por lo tanto, usando los valores propios de $\mathbf{S}$, sabemos que la elipse extiende $c\sqrt{\lambda_i}$ en la dirección de $\mathbf{e}_i$ desde $\overline{\mathbf{x}}$.

Varianza muestral generalizada: continuando con el ejemplo

1

- Si $p = 2$ dimensiones, escoger $c^2 = 5.99$ producirá una elipse que contiene aproximadamente el 95% de las observaciones.
- Los vectores $3\sqrt{5.99}\mathbf{e}_1$ y $\sqrt{5.99}\mathbf{e}_2$ se encuentran en dirección de los ejes de la elipse y sus longitudes son comparables al tamaño del patrón en cada dirección.

1

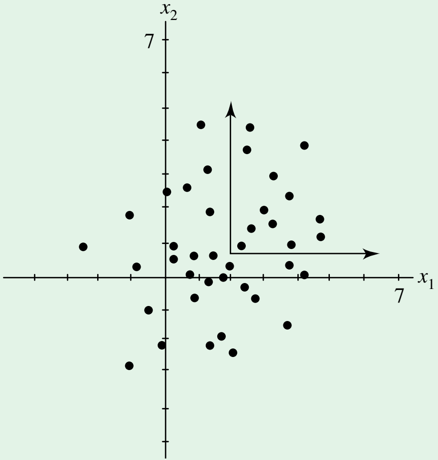


Varianza muestral generalizada: continuando con el ejemplo

2

- Para $\mathbf{S_2} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$ podemos encontrar que el conjunto de valores/vectores propios son: $\lambda_1 = 3, \mathbf{e}_1' = [1, 0]$ y $\lambda_2 = 3, \mathbf{e}_2' = [0, 1]$.
- Entonces los vectores $\sqrt{3}\sqrt{5.99}\mathbf{e}_1$ y $\sqrt{3}\sqrt{5.99}\mathbf{e}_2$ se encuentran en la dirección de los ejes de la elipse.

2

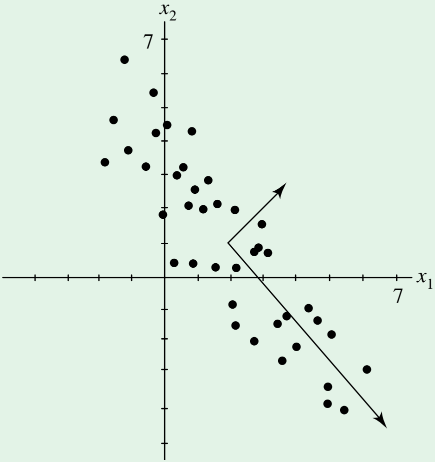


Varianza muestral generalizada: continuando con el ejemplo

3

- Para $\mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$ podemos encontrar que el conjunto de valores/vectores propios son:
  $\lambda_1 = 9, \mathbf{e}_1' = [1/\sqrt{2}, \quad -1/\sqrt{2}]$
  y
  $\lambda_2 = 1, \mathbf{e}_2' = \begin{bmatrix} 1/\sqrt{2}, & 1/\sqrt{2} \end{bmatrix}$.
- Entonces los vectores $\sqrt{3}\sqrt{5.99}\mathbf{e}_1$ y $\sqrt{5.99}\mathbf{e}_2$ se encuentran en la dirección de los ejes de la elipse.

3



Varianza muestral generalizada

Conclusiones sobre la varianza generalizada

- Es 2 dimensiones es fácil graficar los ejes de la elipse (centrada en la media). En mayores dimensiones es más complejo, pero el mismo procedimiento, que usa los vectores propios, funciona para encontrar los ejes.
- La varianza generalizada ($|\mathbf{S}|$) puede producir el mismo valor para patrones diferentes, puesto que no contiene ninguna información acerca de la orientación de los patrones.
- $|\mathbf{S}|$ puede expresarse mediante el producto de los valores propios de $\mathbf{S}$: $\lambda_1\lambda_2\cdots\lambda_p$. Además, la elipse centrada en la media es descrita por $\mathbf{S}^{-1}$, cuyos ejes son proporcionales a la raíz cuadrada de los λ_i's.
- Entonces, los valores propios suministran información acerca de la variabilidad en todas las direcciones en el espacio p-dimensional de los datos.