Independencia/correlación condicional [distr. normal], predicción encadenada, ejemplos

© 2022, Antonio Sala Piqueras. Universitat Politècnica de València. Todos los derechos reservados.

Presentación en vídeo: http://personales.upv.es/asala/YT/V/condnoco.html

Este código ejecutó sin errores en Matlab `R2022a`

Tabla de Contenidos

Independencia Condicional 1
Variables normales multidimensionales condicionalmente no-correladas 2
Ejemplo 1 3
Ejemplo 2 3
2.1: Incorreladas, pero no "condicionalmente" 4
2.2: correladas pero condicionalmente incorreladas 5
Redes bayesianas asociadas 5
Conclusiones 7

Independencia Condicional

Supongamos que tenemos tres variables aleatorias a, b, c.

- Se dice que b y c son "condicionalmente independientes dado $a = a_m$" si

$$p(b|c = c_m, a = a_m) = p(b|a = a_m)$$

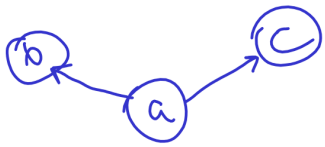
o

$$p(c|b = b_m, a = a_m) = p(c|a = a_m).$$

Equivalentemente, si $p(b, c|a) = p(b|a) \cdot p(c|a)$.

- Si b y c son condicionalmente independientes dada cualquier observación posible de a, entonces diremos que son "condicionalmente independientes dado a".

Esa relación entre las variables permite expresarlas en modelos gráficos del estilo de:



donde la ausencia de "flecha" indica "condicionalmente independiente dado a".

Esas redes dibujan un "árbol" donde cada variable es condicionalmente independiente de "todo lo que no sean sus hijos" dado sus "padres".

Variables normales multidimensionales condicionalmente no-correladas

En general, independencia es difícil de probar, falta de correlación (con cierta confianza) es más fácil si se tienen "muchos" datos. No correlación equivale a no poder hacer predictores lineales útiles.

Si tenemos una matriz simétrica de varianzas-covarianzas conjunta:

$$\Sigma := \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}$$

entonces, la mejor predicción lineal de (b, c) dado a es:

$$\begin{pmatrix} \widehat{b}_{|a} \\ \widehat{c}_{|a} \end{pmatrix} = \begin{pmatrix} \Sigma_{ba} \\ \Sigma_{ca} \end{pmatrix} \Sigma_{aa}^{-1} \quad \cdot \ a$$

con una VC del error de predición dada por:

$$\Sigma_{err} := \begin{pmatrix} \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{cb} & \Sigma_{cc} \end{pmatrix} - \begin{pmatrix} \Sigma_{ba} \\ \Sigma_{ca} \end{pmatrix} \Sigma_{aa}^{-1} \begin{pmatrix} \Sigma_{ba}^T & \Sigma_{ca}^T \end{pmatrix} = \begin{pmatrix} \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ba}^T & \Sigma_{bc} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ca}^T \\ \Sigma_{cb} - \Sigma_{ca}\Sigma_{aa}^{-1}\Sigma_{ba}^T & \Sigma_{cc} - \Sigma_{ca}\Sigma_{aa}^{-1}\Sigma_{ca}^T \end{pmatrix}$$

De modo que si $\Sigma_{err\,(1,2)} = \Sigma_{bc} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ca}^T = 0$, esto es, $\Sigma_{bc} = \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ca}^T$, entonces Σ_{err} es una matriz diagonal, con lo que información extra sobre $e := c - \widehat{c}_{|a}$ no es útil para predecir *LINEALMENTE* b con mayor precisión que $\widehat{b}_{|a}$, y viceversa, información sobre b no es útil para predecir *linealmente* c mejor que $\widehat{c}_{|a}$.

En el caso de distribución normal, eso serviría para determinar "exactamente" la distribución condicional en media y varianza, por lo que las llamaremos "condicionalmente no correladas", pero, evidentemente, sólo es cierto en ciertos casos, por ejemplo con la suposición de normalidad. De lo

contrario, la verdadera distribución de probabilidad "condicional" no tiene por qué tener la media y la varianza de las fórmulas de mejor predicción lineal.

De hecho, siendo distribución normal, **también serían condicionalmente independientes.** En este material, simplemente, nos fijamos en la correlación.

Ejemplo 1

La "suma" $a = b + c$, suponiendo vza b=2, vza c=1, b y c independientes, tendría una matriz VC:

$$E[ab] = E[(b+c)b] = E[b^2] + E[bc] = E[b^2],$$

$$E[ac] = E[(b+c)c] = E[bc] + E[c^2] = E[c^2],$$

$$E[a^2] = E[(b+c)^2] = E[b^2 + c^2 + 2bc] = E[b^2] + E[c^2]$$

```
VCabc=[3 2 1;2 2 0;1 0 1]
```

```
VCabc = 3×3
     3     2     1
     2     2     0
     1     0     1
```

```
Resid=VCabc(2:3,2:3)-VCabc(2:3,1)*inv(VCabc(1,1))*VCabc(1,2:3)
```

```
Resid = 2×2
    0.6667   -0.6667
   -0.6667    0.6667
```

```
[V,D]=eig(Resid)
```

```
V = 2×2
   -0.7071   -0.7071
   -0.7071    0.7071
D = 2×2
    0.0000         0
         0    1.3333
```

El valor propio "cero" indica que sólo hay "una" variable aleatoria y que en cierta dirección, la incertidumbre es cero (claro, conocida la suma y una de las variables, la otra se puede "despejar" de forma determinista).

Ejemplo 2

```
Sigma=@(s_bc) [2 3 4;3 10 s_bc;4 s_bc 20]; %dejamos cov(b,c) como "parámetro ajustable'
S0=Sigma(0) %sin saber "a", las variables "b" y "c" son no correladas.
```

```
S0 = 3×3
     2     3     4
```

```
     3    10     0
     4     0    20
```

```
eig(Sigma(0)) %debe ser def+ para que tenga sentido estadístico
```

```
ans = 3×1
    0.2648
   10.8492
   20.8860
```

2.1: Incorreladas, pero no "condicionalmente"

Nótese que S0 podría venir de variables "b" y "c" no correladas, pero que conocido "a" ya no lo fueran:

```
S0(2:3,2:3)%sale diagonal
```

```
ans = 2×2
    10     0
     0    20
```

```
vzaerrbc=S0(2:3,2:3)-S0(2:3,1)*inv(S0(1,1))*S0(1,2:3) %NO sale diagonal
```

```
vzaerrbc = 2×2
    5.5000   -6.0000
   -6.0000   12.0000
```

```
eig(vzaerrbc)% sigue siendo positiva definida
```

```
ans = 2×1
    1.9263
   15.5737
```

Sólo saldrían "condicionalmente" no correladas si Σ_{ab} o Σ_{ac} fueran, una de las dos al menos, cero:

si b y c son independientes, obtener información de variables correladas con "b" pero no "c" (o viceversa) no mejorará la predicción... pero si hay información "a" correlada con "ambos", entonces ya saber "a" y uno de ellos permite, estimar el otro mejor.

La mejor pred de "c", dado "a" y "b", tendrá una varianza residual del error de predicción dada por:

```
S0(3,3)-S0(3,1:2)*inv(S0(1:2,1:2))*S0(1:2,3)
```

```
ans = 5.4545
```

que es mejor que la vza "20" (igual a la marginal) si sólo conociera "b", y mejor que la varianza de la mejor predicción sólo conociendo "a", que sale "12":

```
S0(3,3)-S0(3,1)*inv(S0(1,1))*S0(1,3)
```

```
ans = 12
```

Sale lo mismo que si añadiera información de "b" después de haber ya supuesto conocido "a":

```
vzaerrbc(2,2)-vzaerrbc(2,1)*inv(vzaerrbc(1,1))*vzaerrbc(1,2)
```

```
ans = 5.4545
```

2.2: correladas pero condicionalmente incorreladas

El valor de covarianza entre b y c para ser condicionalmente independientes (bueno, no correladas sería lo único garantizado en rigor) es:

```
bc_ci=S0(2,1)*inv(S0(1,1))*S0(1,3)
```

```
bc_ci = 6
```

```
Sci=Sigma(bc_ci)
```

```
Sci = 3×3
     2     3     4
     3    10     6
     4     6    20
```

```
eig(Sci) %es def+
```

```
ans = 3×1
    0.7521
    7.3408
   23.9070
```

```
Resid_given_a=Sci(2:3,2:3)-Sci(2:3,1)*inv(Sci(1,1))*Sci(1,2:3) %sale diagonal, cond. in
```

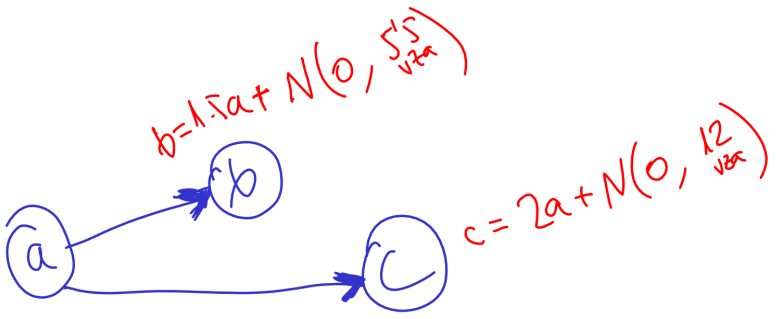
```
Resid_given_a = 2×2
    5.5000         0
         0   12.0000
```

Redes bayesianas asociadas

- Podríamos generar un modelo con:

```
bestpred_given_a=Sci(2:3,1)*inv(Sci(1,1))
```

```
bestpred_given_a = 2×1
    1.5000
    2.0000
```



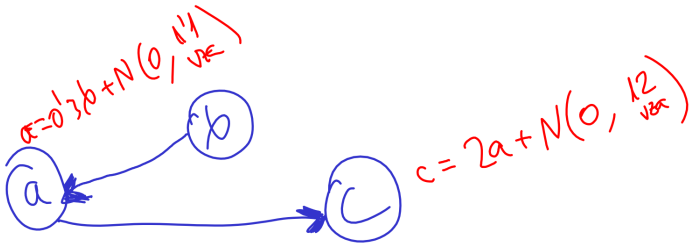
- O bien, otros modelos como:

```
bestpred_given_b=Sci([1 3],2)*inv(Sci(2,2))
```

```
bestpred_given_b = 2×1
    0.3000
    0.6000
```

```
Resid_given_b=Sci([1 3],[1 3])-Sci([1 3],2)*inv(Sci(2,2))*Sci(2,[1 3])
```

```
Resid_given_b = 2×2
    1.1000    2.2000
    2.2000   16.4000
```



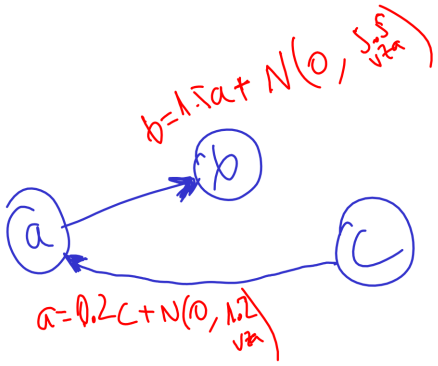
- y un tercer modelo:

```
bestpred_given_c=Sci([1 2],3)*inv(Sci(3,3))
```

```
bestpred_given_c = 2×1
    0.2000
    0.3000
```

```
Resid_given_c=Sci([1 2],[1 2])-Sci([1 2],3)*inv(Sci(3,3))*Sci(3,[1 2])
```

```
Resid_given_c = 2×2
    1.2000    1.8000
    1.8000    8.2000
```



Conclusiones

La independencia condicional parece un "refinamiento inútil" pero está detras de muchas de nuestras suposiciones sobre el mundo (estados, medidas, causa "a" común de dos cosas "b" y "c", ...):
que sabiendo "ciertas cosas" puedo olvidar el efecto de "otras" para estudiar cierta variable... que sabiendo "a" no necesito pensar en "c" para razonar sobre "b", aunque sepa que "a" influye sobre "c" (ya me ocuparé luego de eso si quiero).

-- Me permite no tener que "pensar en mis modelos teóricos" o "calcular con datos" la correlación estadística entre todos los pares de variables posibles, porque se obtiene a partir de otras.

-- Me permite pensar en "variables intermedias" o en "nodos de árbol" que explican las relaciones entre otras. Las "variables intermedias" permiten facilitar computacionalmente la "inferencia estadística" en modelos complejos (redes bayesianas), al poder "propagar" sin necesidad de formar matrices VC de "todo". No objetivo de este material. Las flechas no indican, en principio, "causalidad", pero suele "intentarse que así sea" en modelos más complejos.

Si no se pueden encontrar variables "condicionalmente independientes", no puede establecerse esa "variable intermedia" para los modelos o las predicciones y hay que manipular matrices VC más grandes en las predicciones.