Reconocimiento Estadístico de Patrones - Tarea 1

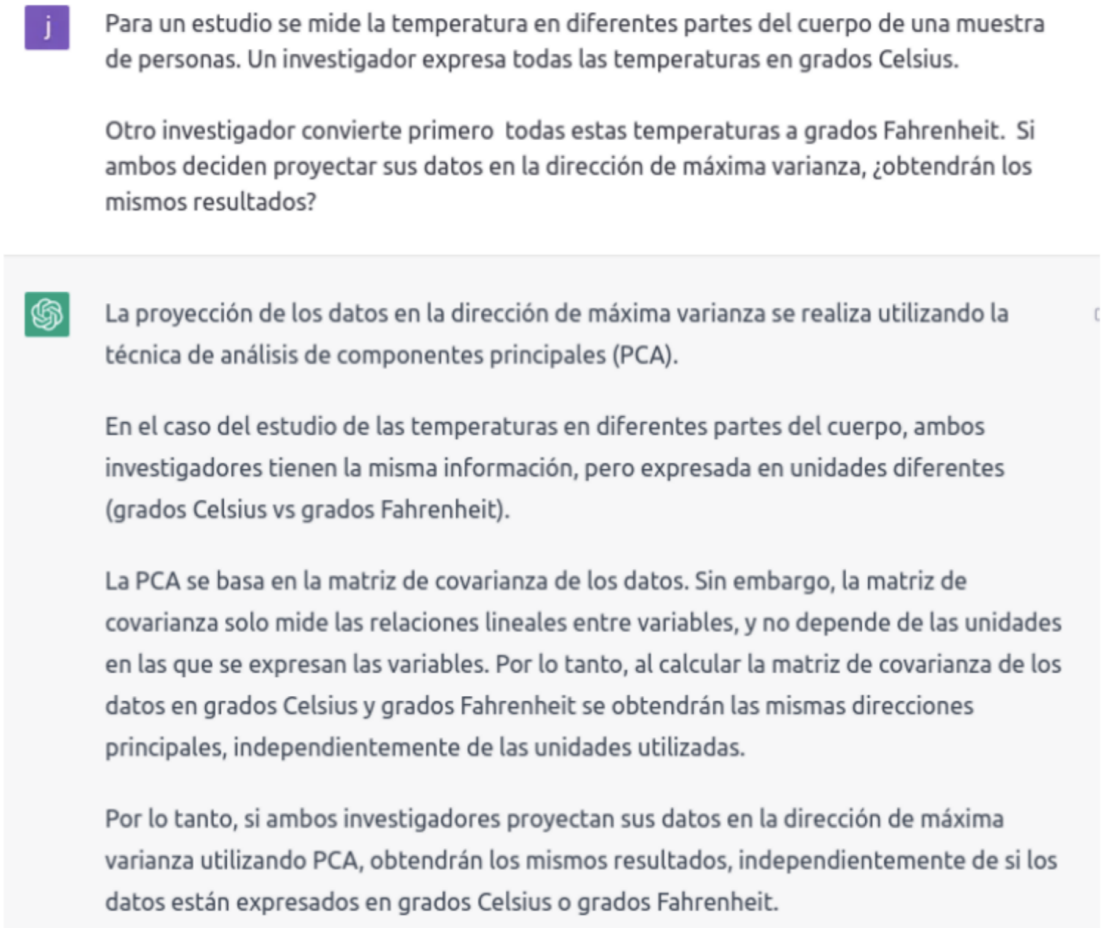
Rubén Pérez Palacios Lic. Computación Matemática, Profesor: Johan Van Horebeek

February 9, 2023

1 Ejercicios

1.1 ChatGPT

1.1.1 Problema



¿Cómo calificarías la respuesta de ChatGPT? ¿Cuál es tu respuesta? Demuéstrala formalmente.

1.1.2 Respuestas

Aunque la respuesta y la intuición es correcta, no lo es la justificación.

A pesar de que la matriz de covarianza solo mide las relaciones lineales entre variables no necesariamentre es independiente de como esten expresadas las variables. Puesto que estas pueden no ser una transformación lineal una de la otra, como es nuestro caso.

Sin embargo si se cumple en nuestro caso, veamos lo siguiente. Sean X y Y v.a. que miden la temperatura en difernetes partes del cuerpo en grados Celcius y grados Fahrenheit respectivamente. Recordemos que

$$Y = \frac{9}{5}X + 32.$$

Puesto que la varianza es invariante bajo translaciones y $Var(aX) = a^2 Var(X)$.

$$Var(Y) = Var(\frac{9}{5}X + 32) = \frac{81}{25}Var(X),$$

es decir la varianza de Y es proporcional a la varianza de X, por lo tanto maximizar la varianza de X es análogo a maximizar la varianza de Y,

$$\max_l \frac{Var(l^t X)}{||l||^2} \iff \max_l \frac{Var(l^t Y)}{||l||^2},$$

con lo que concluimos que ambos obtendran la misma dirección de máxima varianza.

1.2 Normal Estandar

1.2.1 Problema

Supongamos que $X = (X_1, X_2), Var(X_1) = Var(X_2) = 1, E[X_1] = E[X_2] = 1, y X_1 \perp X_2$. Demuetra que cualquier dirección l da máxima varianza en las proyecciones.

1.2.2 Demostración

Por definición de $Cov(X)$ obtenemos

$$Cov(X) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

por lo tanto

$$\max_l \frac{l^t Cov(X) l}{||l||^2} = \max_l \frac{l^t l}{||l||^2} = \max_l 1,$$

lo cual es independiente de l. Con lo que concluimos cualquier dirección l da máxima varianza en las proyecciones.

1.3 Demuestra que $K = -\frac{1}{2}CD^2C$

Recordemos que

- K : Es la matriz de producto punto entre variables, es decir $K_{i,j} = X_i \dot{X}_j$.
- D^2 : Es la matriz de distancias al cuadrado entre variables, es decir $D^2_{i,j} = d(X_i, X_j)^2$.
- C : Es la matriz identidad menos 1/n sobre todas las entradas, es decir $C = I - \frac{1}{n}J$.

También en clase se demostro que

$$K = -\frac{1}{2}CD^2C$$

$$K_{i,j} = -\frac{1}{2}\left(D^2_{i,j} - \frac{\sum_j D^2_{i,j}}{n} + \frac{\sum_j K_{i,j}}{n} - \frac{\sum_i D^2_{i,j}}{n} + \frac{\sum_i K_{i,j}}{n} \right)$$

$$\sum_i \sum_j D^2_{i,j} = 2n \sum_i K_{i,i}$$

por lo que nos falta demostrar que

$$CD^2C = D^2_{i,j} - \frac{1}{n}\sum_j D^2_{i,j} - \frac{1}{n}\sum_i D^2_{i,j} + \frac{1}{n^2}\sum_i \sum_j D^2_{i,j}.$$

Para ello usaremos la notación $X_{i,\cdot}$ y $X_{\cdot,j}$, para denotar al vector renglón y columna respectivamente. Ahora veamos que

$$(CD^2C)_{i,j} = (CD^2)_{i,\cdot}C_{\cdot,j} \quad \text{Por definición de multiplicación de matrices} \tag{1}$$

$$= \sum_k (CD^2)_{i,k}(C)_{k,j} \quad \text{Por definición de multiplicación de matrices} \tag{2}$$

$$= \sum_k (C_{i,\cdot}D^2_{\cdot,k})C_{k,j} \quad \text{Por definición de multiplicación de matrices} \tag{3}$$

$$= \sum_k (\sum_h C_{i,h}D^2_{h,k})C_{k,j} \quad \text{Por definición de multiplicación de matrices} \tag{4}$$

$$= \sum_k (D^2_{i,k} - \sum_h \frac{1}{n}D^2_{h,k})C_{k,j} \quad \text{Por definición de C} \tag{5}$$

$$= \sum_k (D^2_{i,k} - \frac{1}{n}\sum_h D^2_{h,k})C_{k,j} \tag{6}$$

$$= (D^2_{i,j} - \frac{1}{n}\sum_h D^2_{h,j}) + \sum_k -\frac{1}{n}(D^2_{i,k} - \frac{1}{n}\sum_h D^2_{h,k}) \quad \text{Por definición de C} \tag{7}$$

$$= D^2_{i,j} - \frac{1}{n}\sum_h D^2_{h,j} - \frac{1}{n}\sum_k (D^2_{i,k} - \frac{1}{n}\sum_h D^2_{h,k}) \tag{8}$$

$$= D^2_{i,j} - \frac{1}{n}\sum_i D^2_{i,j} - \frac{1}{n}\sum_j D^2_{i,j} + \frac{1}{n^2}\sum_i\sum_j D^2_{i,j} \tag{9}$$

$$\tag{10}$$

Por lo tanto concluimos la demostración.

1.4 Plano de direcciones PCA

1.4.1 Problema

Revisa el video sobre la maximización del cociente de Rayleigh: https://youtu.be/8TBpSUXcDww.

Haz unos pequeños cambios necesarios para demostrar que el segundo vector propio de Cov(X) es la solución del problema de maximizar el cociente bajo la restricción adicional de ser ortogonalal primer vector propio.

1.4.2 Demostración

Primero recordemos lo siguiente de la demostración. Sea A una matriz simétrica entonces

- Su SVD es $A = U\Lambda U^t$, donde $U^t = U^-1$ y Λ es una matriz diagonal con los valores propios de A.
- Definimos a $\Lambda^{1/2} := Diag(\sqrt{\mu_i})$, donde μ_i son los valores propios de A, entonces podemos ver que $\Lambda = \Lambda^{1/2}\Lambda^{1/2}$ y por lo tanto $(U\Lambda^{1/2}U^t)(U\Lambda^{1/2}U^t) = A^{1/2}A^{1/2} = A$.

- El vector dirección que maximiza la varianza es $U^t e_1 = v_1$ el primer vector propio de $Cov(X)$ con valor μ_1.
- Debido a que U es una matriz ortonormal entonces conserva ortgonalidades entonces $l_1^t l = 0$ ssi $(U^t l_1)^t (U^t l) = 0$ ssi $e_1^t y = 0$.
- Por último, si $v^t e_1 = 0$ es ssi $\sum v_i e_i = 0$ lo cúal es ssi $v_1 = 0$.

Entonces podemos ver que

$$\begin{aligned}
\max_{l: l_1^t l = 0} \frac{l^t Cov(X) l}{l^t l} &= \max_{l: l_1^t l = 1} \frac{l^t (U\Lambda^{1/2}U^t)(U\Lambda^{1/2}U^t) l}{l^t l} && \text{Puesto que } Cov(X) \text{ es simétrica} && (11)\\
&= \max_{l: l_1^t l = 0} \frac{l^t (U\Lambda^{1/2}U^t)(U\Lambda^{1/2}U^t) l}{l^t (UU^t) l} && \text{Ya que } U \text{ es una matriz unitaria} && (12)\\
&= \max_{y: y^t e_1 = 0} \frac{y \Lambda y^t}{y^t y} && \text{Con el cambio de variable } y = U^t l && (13)\\
&= \max_{y: y^t e_1 = 0} \frac{\sum_{i=1}^n \mu_i y_i^2}{\sum_{i=1}^n y_i^2} && \text{Por definición de } \Lambda && (14)\\
&= \max_{y} \frac{\sum_{i=2}^n \mu_i y_i^2}{\sum_{i=2}^n y_i^2} && \text{Ya que } y^t e_1 = 0 \text{ entonces } y_1 = 0 && (15)\\
&\leq \max_{y} \frac{\sum_{i=1}^n \mu_2 y_i^2}{\sum_{i=1}^n y_i^2} && \text{Debido a que } \mu_2 \geq \mu_i, \forall i > 2 && (16)\\
&= \max_{y} \mu_2 && && (17)\\
& && && (18)
\end{aligned}$$

Ahora veamos que si $y = (0, 1, \cdots, 0)$ entonces

$$\frac{\sum \mu_i y_i^2}{\sum y_i^2} = \frac{\mu_2}{1} = \mu_2.$$

Por lo tanto concluimos que v_2 el segundo vector propio de $Cov(X)$, es la dirección que maximiza la varianza y que además es ortgonal a v_1, cuyo máximo es μ_2.

1.5 Visualización de Datos

1.5.1 Problema

En el archivo heptatlon se pueden consultar los tiempos y el puntaje final (score) de 25 atletas que participaron en el heptatlon durante los juegos olmpicos de Seoul.

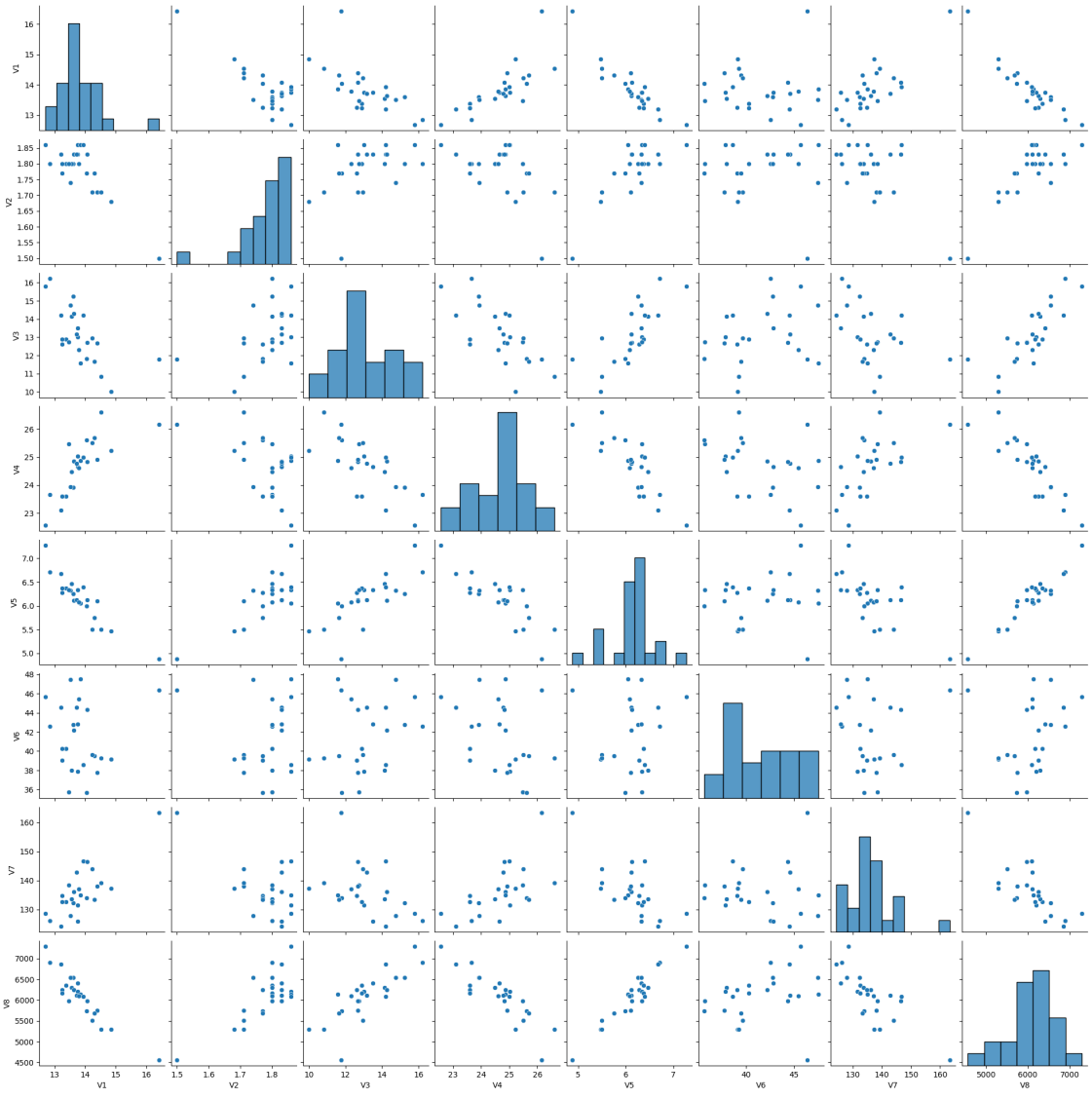
a) Busca visualizaciones informativas de estos datos multivariados.

b) Haz un análisis de componentes principales con los tiempos (sinscore). Hay una relación entre el score y las proyecciones sobre el primer CP? Puedes leer los datos con: d¡-read.table("hepatlon")

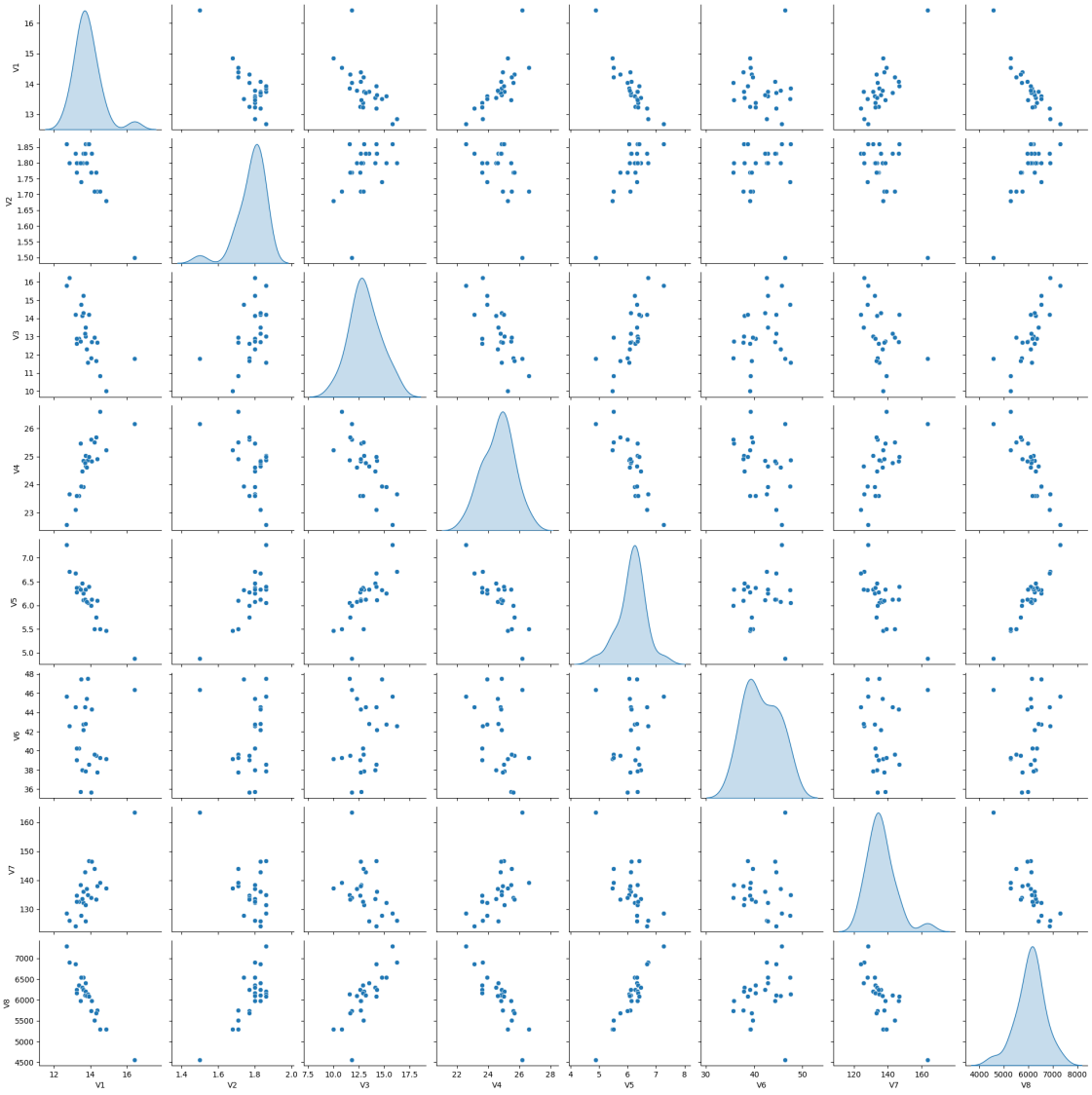
1.5.2 Solución

a) Renombramos (hurdles,highjump,shot,run200m,longjump,javelin,run800m,score) como (V1, V2, V3, V4, V5, V6, V7, V8)

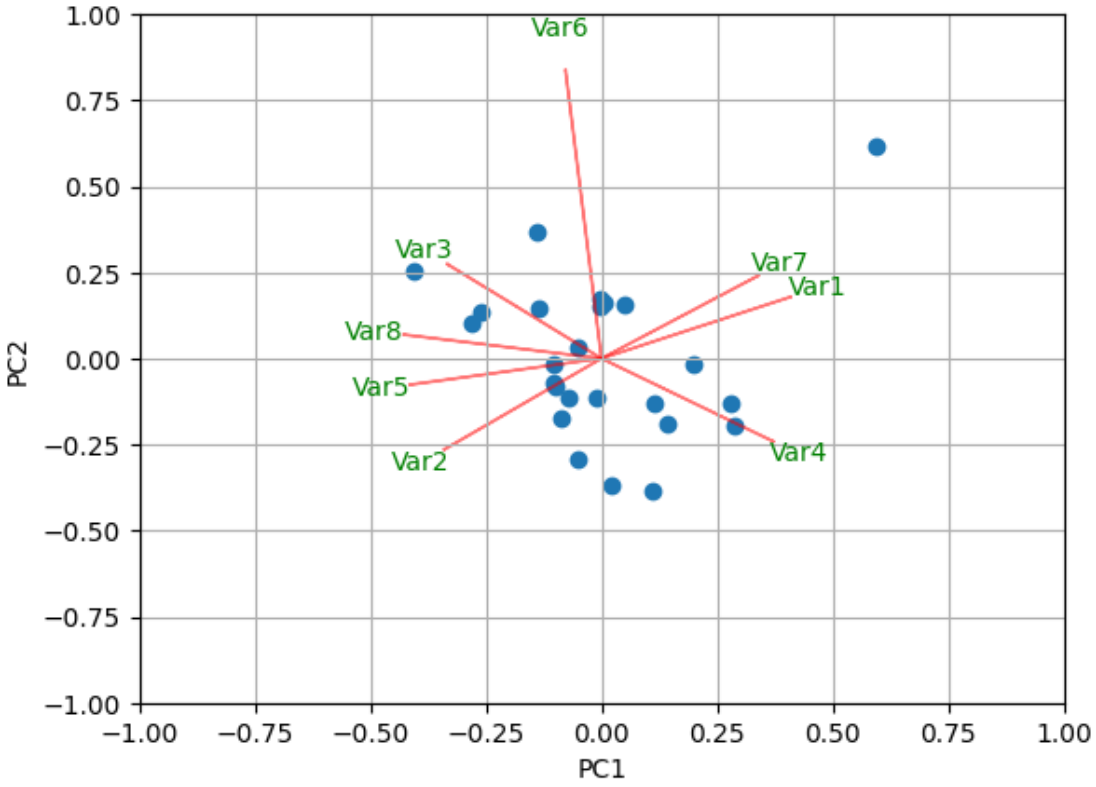
Distribuciones conjuntas a pares de (V_i,V_j) con $i \neq j$ y distirbución margina discreta V_i cuando $i = j$.



Distribuciones conjuntas a pares de (V_i,V_j) con $i \neq j$ y distirbución margina continua V_i cuando $i = j$.



Podemos ver en el biplot que v7, v1 y v4 están fuertemente relacionada con el PC1 e inversamente relacionadas v2, v5, v8, v3. En cambio la v6 está fuertemente relacionada con PC2. Así mismo podemos ver que la hay bastante correlación entre v7, v1 y v4, asi como bastante correlación entre v2, v5, v8, v3.



b)

[1.19624885e-03, -1.05204013e-04, -2.09849444e-03, 1.47499229e-03, -7.92784821e-04, -1.57865298e-03, 1.12710147e-02, -9.99930909e-01]

[1.92835321e-02, -6.05680149e-04, 8.27323011e-02, -1.73321242e-02, 1.08856631e-03, 3.27484685e-01, 9.40817259e-01, 9.91075581e-03]

[-4.42281031e-02, 5.20709703e-03, 5.23034341e-02, 2.52425088e-02, 3.33247513e-02, -9.42375189e-01, 3.24711313e-01, 4.99545357e-03]

[1.27735048e-01, -3.39503633e-02, 9.78341573e-01, 1.16494549e-01, -5.42199304e-02, 1.74323988e-02, -9.25021755e-02, -2.75215967e-03]

[2.99253782e-02, -6.54426901e-02, 1.10968349e-01, -9.90594181e-01, 8.48590375e-03, -2.83264257e-02, -1.87933943e-02, -1.82525884e-03]

[-9.77750566e-01, 1.11049607e-02, 1.19861716e-01, -1.94015933e-02, -1.64765138e-01, 4.41739357e-02, -6.02027169e-03, -1.45801494e-03]

[1.56298683e-01, 7.45409331e-02, -6.80079688e-02, -1.54259545e-02, -9.81458446e-01, -3.97783256e-02, 1.75083617e-02, 1.33740223e-03]