Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Laboratorio 1

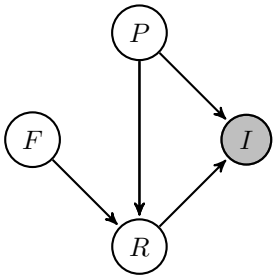
Estudiante: Daniel Minaya
Profesor: Felipe Tobar
Auxiliares: Cristóbal Alcázar
Camilo Carvajal

Fecha de entrega: 8 de septiembre

Parte 2 (Modelos gráficos)

(a) Sabemos que la cantidad de polen puede causar irritación en nariz y ojos en algunos casos, por lo que habría una dependencia entre estas dos variables. Además, la cantidad de polen también afecta a la probabilidad de sufrir Rinitis alérgica, por lo que tendríamos otra dependencia. También es sabido que si se tiene Rinitis alérgica, entonces es probable tener irritación en nariz y ojos. Por último, suponiendo que la variable de tener un familiar con Rinitis alérgica hace referencia a una enfermedad hereditaria, y no a una enfermedad que pueda ser contagiada, tendríamos que esta variable aumentaría las posibilidades de tener la Rinitas alérgica.

Definiendo las variables aleatorias $I =$ "Tener irritación en nariz y ojos", $P =$ "cantidad de polen en el ambiente", $R =$ "Tener Rinitis alérgica" y $F =$ "Tener un familiar con Rinitis alérgica", entonces, por todo lo anterior, el modelo gráfico que representa el problema estaría dado por el siguiente DAG:



(b) Usando la notación anterior, queremos encontrar $\mathbb{P}(R|I)$. Según el modelo gráfico planteado, la probabilidad conjunta está dada por:

$$\mathbb{P}(P, I, F, R) = \mathbb{P}(P) \cdot \mathbb{P}(F) \cdot \mathbb{P}(R|F, P) \cdot \mathbb{P}(I|R, P),$$

donde todas las probabilidades en el lado izquierdo son conocidas a priori. De este modo, por el teorema de Bayes tenemos:

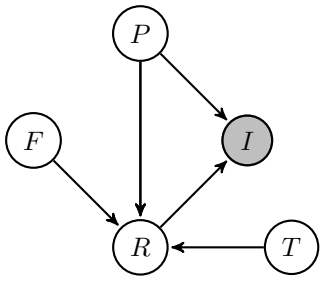
$$\mathbb{P}(R|I) = \frac{\mathbb{P}(I|R)\mathbb{P}(R)}{\mathbb{P}(I)},$$

luego, por el teorema de probabilidades tenemos:

$$\begin{aligned}
\mathbb{P}(I|R) &= \sum_k \mathbb{P}(I|R, P = k)\mathbb{P}(P = k), \\
\mathbb{P}(R) &= \sum_k \sum_j \mathbb{P}(R|F = j, P = k)\mathbb{P}(F = j)\mathbb{P}(P = k), \\
\mathbb{P}(I) &= \sum_k \sum_j \mathbb{P}(I|R = j, P = k)\mathbb{P}(R = j)\mathbb{P}(P = k),
\end{aligned}$$

donde las primeras dos igualdades están en términos de probabilidades conocidas, mientras que la última igualdad tiene el término $\mathbb{P}(R = j)$ que se conoce de la segunda igualdad, por lo que $\mathbb{P}(R|I)$ se obtendría calculando las probabilidades respectivas.

(c) En este caso debemos agregar una nueva variable al modelo gráfico, llamemósla $T =$ "Test positivo". Esta variable solo afecta a la probabilidad de tener Rinitis alérgica, por lo que el nuevo modelo gráfico estaría dado por:



Con este nuevo modelo gráfico, la nueva probabilidad conjunto sería:

$$\mathbb{P}(P, I, F, R, T) = \mathbb{P}(P) \cdot \mathbb{P}(F) \cdot \mathbb{P}(T) \cdot \mathbb{P}(R|F, P, T)\mathbb{P}(I|R, P),$$

luego la probabilidad $\mathbb{P}(R|I) = \frac{\mathbb{P}(I|R)\mathbb{P}(R)}{\mathbb{P}(I)}$ solo vería afectada su término $\mathbb{P}(R)$, el cual ahora estaría dado por

$$\mathbb{P}(R) = \sum_k \sum_j \sum_i \mathbb{P}(R|T = i, F = j, P = k)\mathbb{P}(T = i)\mathbb{P}(F = j)\mathbb{P}(P = k).$$

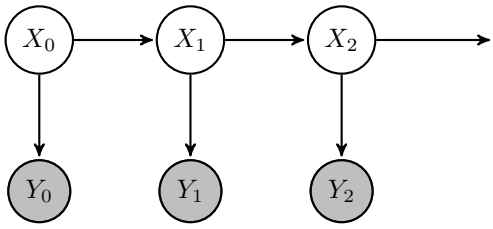
Notar que esto afectaría indirectamente a $\mathbb{P}(I)$ también, pues depende de $\mathbb{P}(R)$.

(d) En este caso utilizamos P como variable continua que sigue una distribución semi-normal, pues al tratarse de la cantidad de polen en el ambiente, ésta puede tomar solo valores positivos. En este caso, la expresión para $\mathbb{P}(R|I)$ cambia, pues pasamos P de una variable discreta a una continua, por lo que:

$$\begin{aligned}
\mathbb{P}(I|R) &= \int_0^\infty \mathbb{P}(I|R, P = x) f_P(x) dx, \\
\mathbb{P}(R) &= \int_0^\infty \sum_j \sum_i \mathbb{P}(R|T = i, F = j, P = x)\mathbb{P}(T = i)\mathbb{P}(F = j) f_P(x) dx, \\
\mathbb{P}(I) &= \int_0^\infty \sum_j \mathbb{P}(I|R = j, P = x)\mathbb{P}(R = j) f_P(x) dx,
\end{aligned}$$

donde $f_P(x)$ es la densidad de una semi-normal, es decir, $f_P(x) = \sqrt{\frac{2}{\sigma^2 \pi}} e^{-x^2/2\sigma^2}$.

(e) Definamos $X_t =$ "cantidad de polen en el tiempo t" e $Y_t =$ "irritación en nariz y ojos en el tiempo t" procesos aleatorios. Como suponemos que el futuro es independiente del paso condicionado al presente, entonces $(X_t)_t$ es una cadena de Markov, por lo que nuestro modelo gráfico ahora es un Hidden Markov Model representado por:



Queremos predecir la cantidad de polen cada 30 minutos usando como única medición el nivel de irritación en nariz y ojos, es decir, se quiere calcular $\mathbb{P}(X_{30t}|Y_{1:30t}) = \mathbb{P}(X_{30t}|Y_1, \ldots, Y_{30t})$, para lo cual se podría utilizar el filtro de Kalman.

Parte 3 (Introducción al Transporte Óptimo)

(a) Dadas dos distribuciones discretas $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ y $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$. El transporte óptimo entre α y β está dado por P^* que resuelve el siguiente problema de minimización:

$$\begin{aligned} \min_{P \in \mathcal{M}_{nm}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} C_{ij} \\ \text{s.a.} \quad & \sum_{j=1}^{m} P_{ij} = a_i, \ \forall i \in [n] \\ & \sum_{i=1}^{n} P_{ij} = b_j, \ \forall j \in [m] \\ & P_{ij} \geq 0, \ \forall i \in [n], j \in [m] \end{aligned} \tag{P}$$

donde $C_{ij} = \|x_i - y_j\|^2$. Identificando las matrices P y C como vectores en $\mathbb{R}^{nm}$ apilando las filas de la siguiente manera:

$$P_{i\bullet}^T = \begin{pmatrix} P_{i1} \\ \vdots \\ P_{im} \end{pmatrix} \in \mathbb{R}^m \Rightarrow P = \left(\begin{array}{c} P_{1\bullet}^T \\ \hline \vdots \\ \hline P_{n\bullet}^T \end{array}\right) \in \mathbb{R}^{nm}, \quad C_{i\bullet}^T = \begin{pmatrix} C_{i1} \\ \vdots \\ C_{im} \end{pmatrix} \in \mathbb{R}^m \Rightarrow C = \left(\begin{array}{c} C_{1\bullet}^T \\ \hline \vdots \\ \hline C_{n\bullet}^T \end{array}\right) \in \mathbb{R}^{nm},$$

el problema (P) se puede reescribir como el siguiente problema lineal (PL):

$$\begin{aligned} \min_{P \in \mathbb{R}^{nm}} \quad & C^T P \\ \text{s.a.} \quad & ZP = \omega \\ & P \geq 0 \end{aligned} \tag{PL}$$

donde usamos

$$Z = \left[\begin{array}{c|c|c} A_1 & \cdots & A_n \\ \hline I_m & \cdots & I_m \end{array}\right] \in \mathcal{M}_{n+m,nm}, \quad \omega = \left(\begin{array}{ccc|ccc} a_1 & \cdots & a_n & b_1 & \cdots & b_m \end{array}\right)^T \in \mathbb{R}^{n+m}$$

y $A_i \in \mathcal{M}_{nm}$ es una matriz de 0's en todas las posiciones excepto en la fila i, que tiene solo 1's e I_m es la matriz identidad en $\mathcal{M}_{mm}$.

(b) Para el caso $d = 1$ tenemos que $\Sigma_1 = \sigma_1^2$ y $\Sigma_2 = \sigma_2^2$ corresponden a las varianzas y $A = \sigma_1^{-1} \left(\sigma_1 \sigma_2^2 \sigma_1\right)^{1/2} \sigma_1^{-1} = \frac{\sigma_2}{\sigma_1}$, por lo cual el transporte óptimo viene dado por

$$T(x) = \mu_2 + \frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Notemos que el transporte óptimo se puede ver como tres transformaciones que llevan $\mathcal{N}(\mu_1, \Sigma_1)$ a $\mathcal{N}(\mu_2, \Sigma_2)$. Sea $x \in \mathcal{N}(\mu_1, \sigma_1^2)$, entonces $x - \mu_1 \in \mathcal{N}(0, \sigma_1)$, es decir, llevamos x al origen, luego al multiplicar por A tenemos $A(x - \mu_1) \in \mathcal{N}(0, \sigma_2^2)$, es decir, achicamos por σ_1 y amplificamos por σ_2, y por último, trasladamos a la segunda distribución de modo que $\mu_2 + A(x - \mu_1) \in \mathcal{N}(\mu_2, \sigma_2^2)$.

De manera general (para $d > 1$) la intución es la misma, primero trasladamos los datos al origen, luego en el origen deformamos la primera distribución hasta obtener la forma de la segunda distribución. Por último, una vez obtenemos la forma correcta, centramos en la media de la distribución objetivo.

(c) Sampleamos muestras de dos distribuciones normales. Nuestra distribución inicial será $\mathcal{N}(0, 3)$ y samplearemos 50 muestras, y la distribución objetivo será $\mathcal{N}(5, 2)$ y samplearemos 60 muestras.

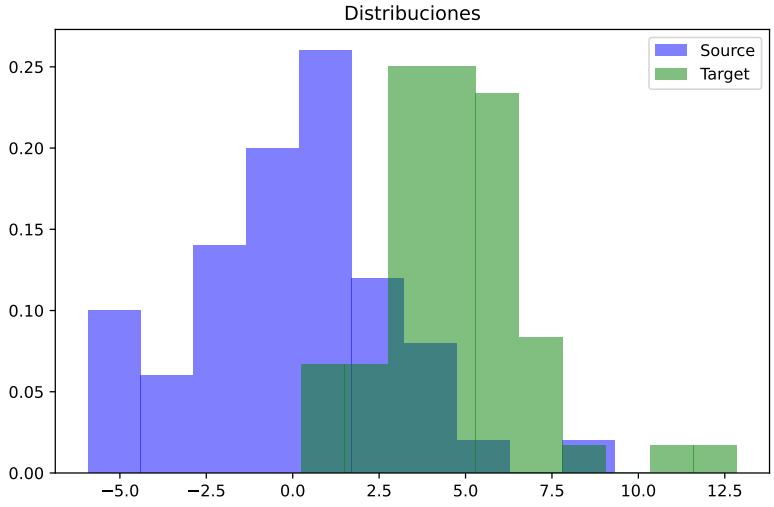


Figura 1: Muestras de distribuciones normales.

Aplicando el método anterior, obtenemos el siguiente transporte óptimo.

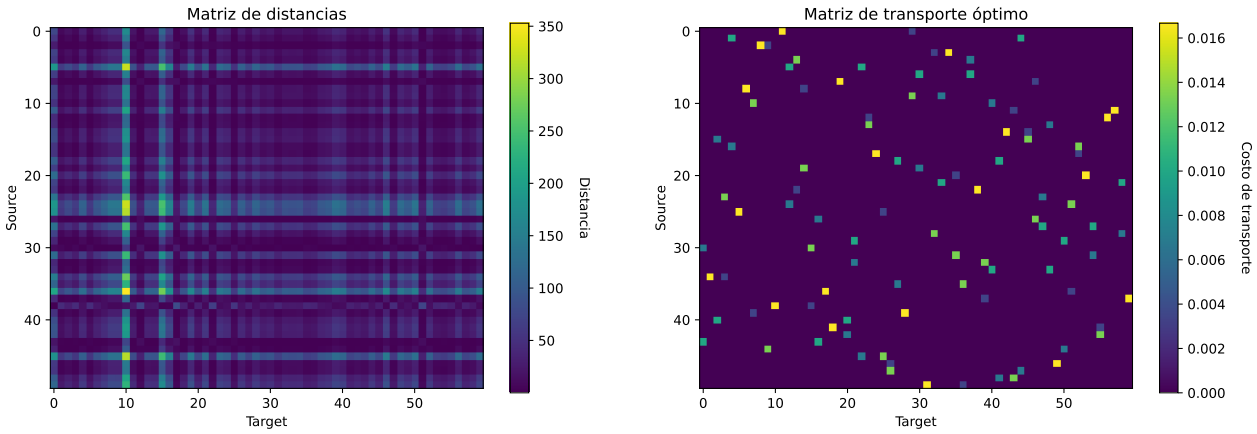


Figura 2: Matriz de distancias y transporte óptimo.

El valor óptimo obtenido por el método fue de 22.88, mientras que la distancia de Wasserstein entre las dos distribuciones es 25.10, por lo que obtuvimos un resultado parecido usando el método que estaba aproximando una normal mediante muestras.

(d) Sampleamos muestras de dos distribuciones normales en $\mathbb{R}^2$. La distribución inicial será $\mathcal{N}(\mu_1, \Sigma_1)$ de la cual samplearemos 10 muestras, y la distribución objetivo será $\mathcal{N}(\mu_2, \Sigma_2)$ de la cual samplearemos 15 muestras, donde

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 10 \\ 3 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

son las medias y matrices de covarianza respectivas.

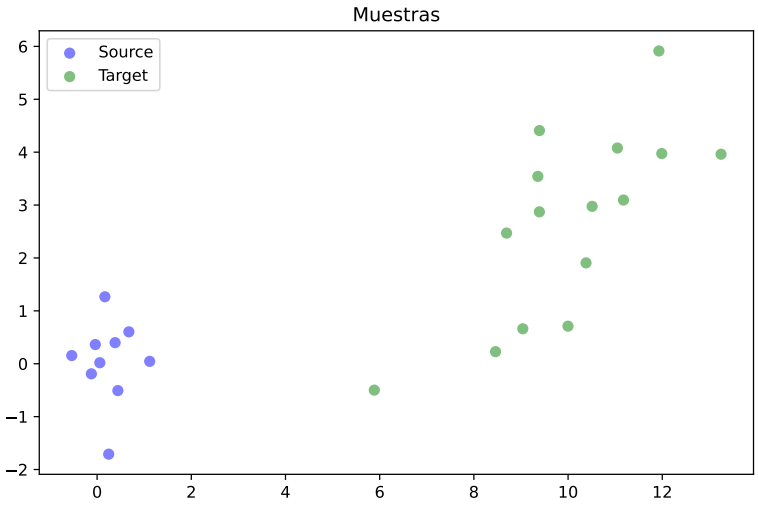


Figura 3: Muestras de distribuciones normales en $\mathbb{R}^2$

Obtenemos el siguiente transporte óptimo.

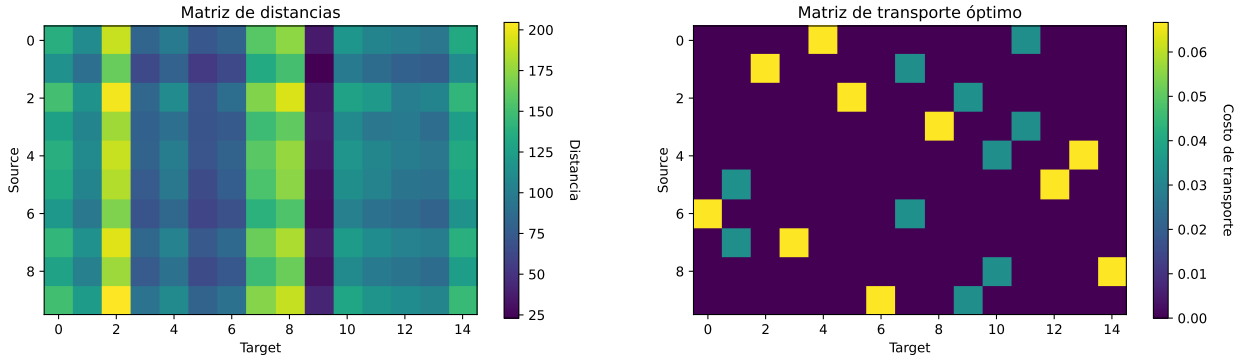


Figura 4: Matriz de distancias y transporte óptimo.

Esto se traduce en las siguientes asignaciones.

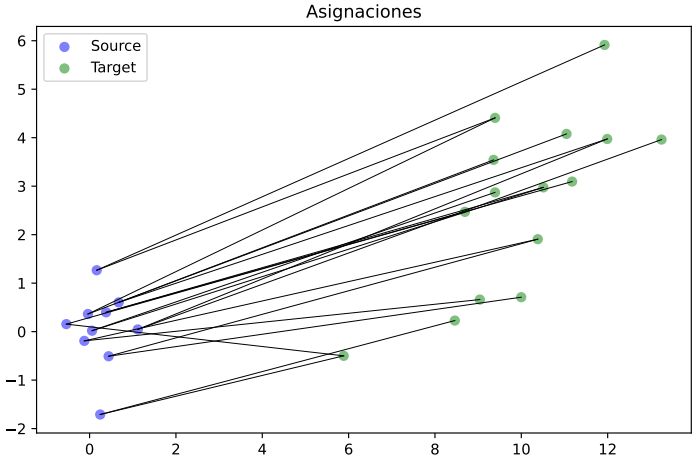
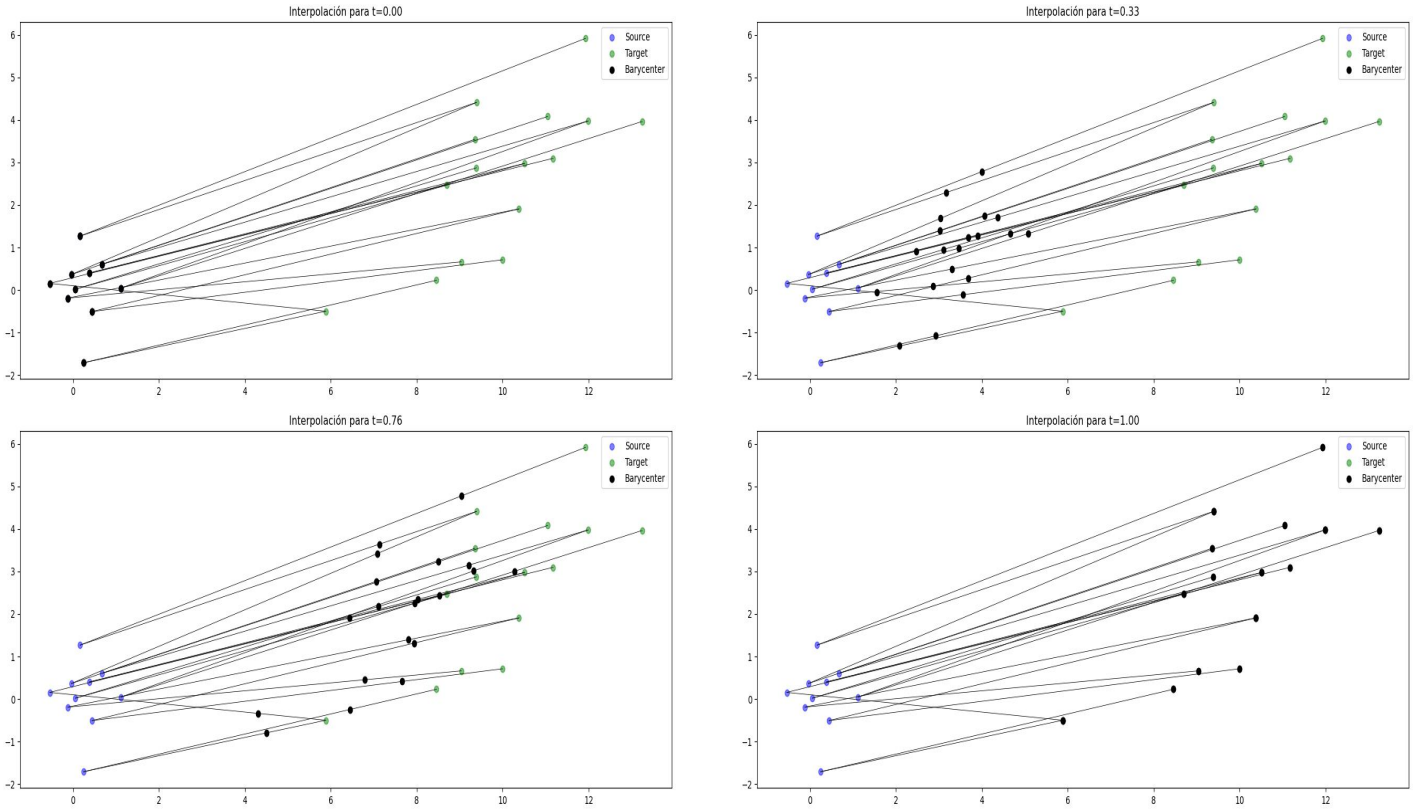


Figura 5: Asignaciones.

(e) El baricentro nos permite visualizar cómo ocurre la transformación de la distribución inicial hacia la distribución objetivo.



En el siguiente link se puede ver una animación para 100 valores de t entre 0 y 1.