Factor Analysis

(Chapter 21)[1]

David Barber

(adapted by Miguel Palomino)

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Distribución normal (o gaussiana) multidimensional

La función de densidad de una distribución gaussiana D dimensional para una variable continua $\mathbf{x}$ es

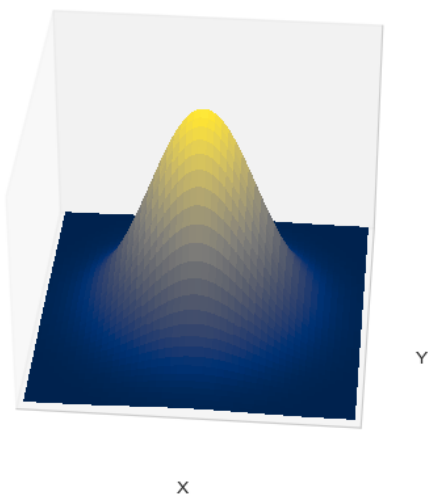
$$p(\mathbf{x}|\mathbf{m},\mathbf{S}) = \frac{1}{\sqrt{\det(2\pi\mathbf{S})}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right\}$$

donde $\mathbf{m}$ es la media y $\mathbf{S}$ la matriz de covarianzas.

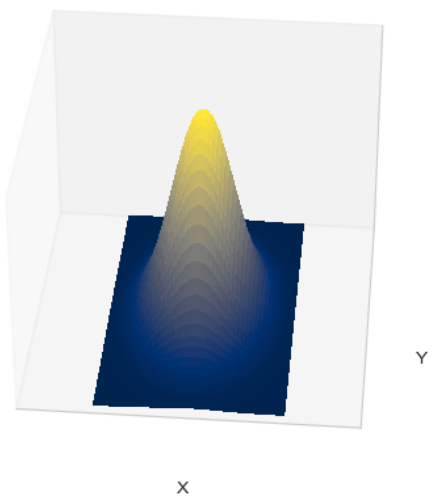
Se trata de la distribución conjunta más utilizada para variables aleatorias continuas. Su popularidad se debe a sus propiedades matemáticas y también a que en muchas situaciones la asunción de gaussianidad resulta bastante razonable.

Distribución gaussiana bidimensional

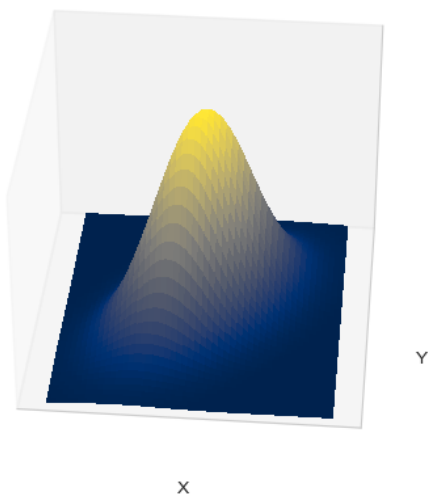
$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$



$$\mathbf{S} = \begin{pmatrix} 3 & 1.8 \\ 1.8 & 4 \end{pmatrix}$$



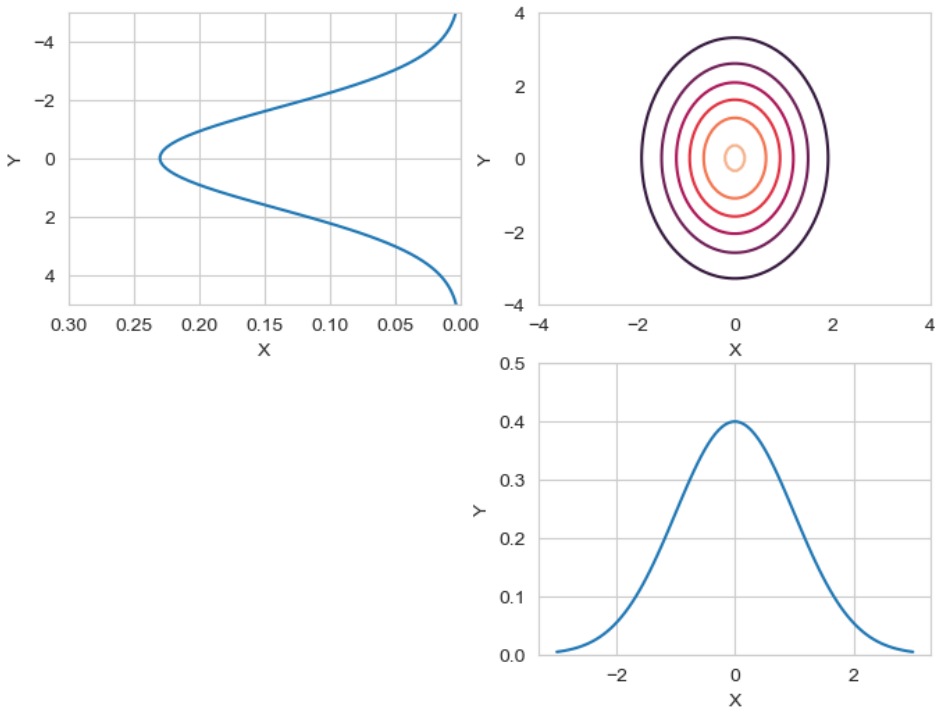
Distribución marginal de una gaussiana

Particionamos una variable $\mathbf{x}$ D-dimensional en dos partes $\mathbf{x}_1$ y $\mathbf{x}_2$ de modo que

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad y \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Entonces, las distribuciones marginales son normales:

$$p(\mathbf{x}_1) = \int p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$



Distribución condicionada y transformación lineal

Las distribuciones condicionadas también son normales:

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$$

donde

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Transformaciones lineales

Si $\mathbf{y} = \mathbf{M}\mathbf{x} + \boldsymbol{\eta}$, con $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, entonces

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}, \mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\mathsf{T} + \boldsymbol{\Sigma})$$

Factor Analysis

FA is essentially a probabilistic extension of Principal Components Analysis. It is very widely used in practice and one of the central tools in statistical analysis. $\mathbf{v}$ is 'visible' data vector. The dataset is then given by a set of vectors,

$$\mathcal{V} = \left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\}$$

where $\dim(\mathbf{v}) = D$. Our interest is to find a lower H-dimensional probabilistic description of this data.

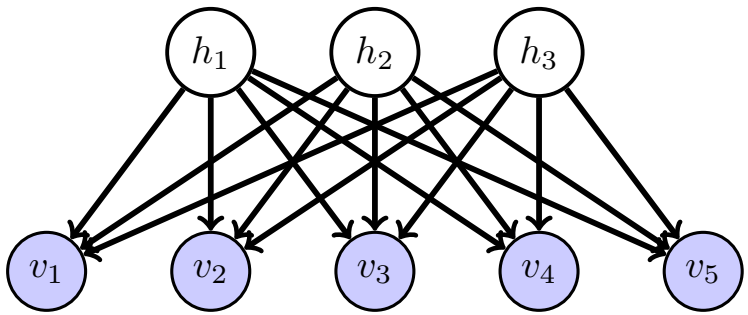
$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon}$$

where the noise $\boldsymbol{\epsilon}$ is Gaussian distributed,

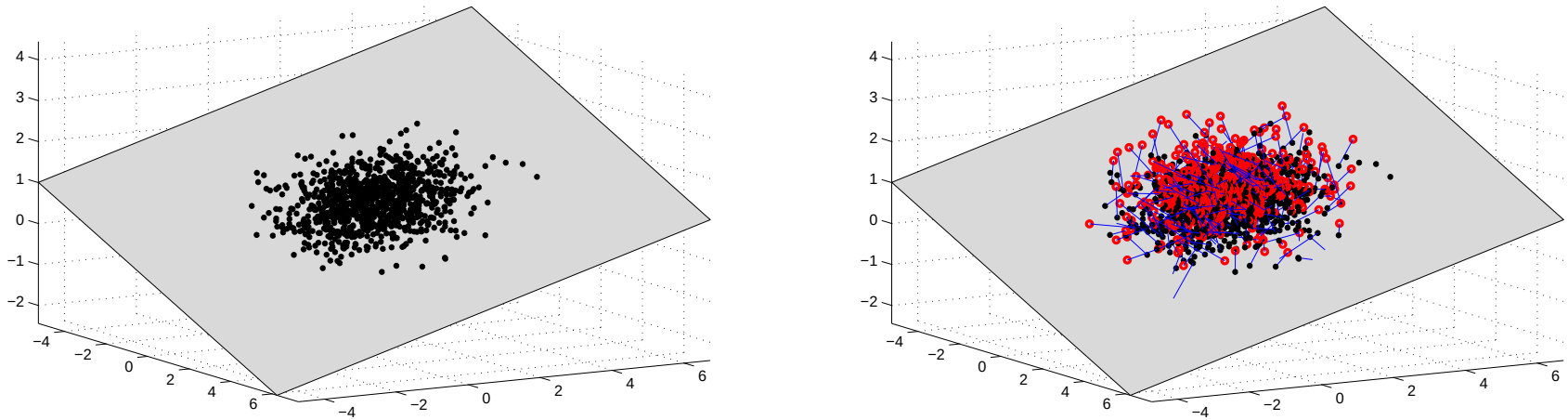
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Psi})$$

Graphical model

The coordinates $\mathbf{h}$ will be preferentially concentrated around values close to $\mathbf{0}$. If we sample a $\mathbf{h}$ from $p(\mathbf{h})$ and then draw a value for $\mathbf{v}$ using $p(\mathbf{v}|\mathbf{h})$, the sampled $\mathbf{v}$ vectors would produce a saucer or 'pancake' of points in the $\mathbf{v}$ space.



Pancakes



(a) (b)

Figure: Factor Analysis: 1000 points generated from the model. **(a)**: 1000 latent two-dimensional points $\mathbf{h}^n$ sampled from $\mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I})$. These are transformed to a point on the three-dimensional plane by $\mathbf{x}_0^n = \mathbf{c} + \mathbf{F}\mathbf{h}^n$. The covariance of $\mathbf{x}_0$ is degenerate, with covariance matrix $\mathbf{F}\mathbf{F}^\mathsf{T}$. **(b)**: For each point $\mathbf{x}_0^n$ on the plane a random noise vector is drawn from $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\boldsymbol{\Psi})$ and added to the in-plane vector to form a sample $\mathbf{x}^n$, plotted in red. The distribution of points forms a 'pancake' in space. Points 'underneath' the plane are not shown.

Una descripción probabilística

$$p\left(\mathbf{v}|\mathbf{h}\right) = \mathcal{N}\left(\mathbf{v}|\mathbf{F}\mathbf{h} + \mathbf{c}, \mathbf{\Psi}\right)$$

Para completar el modelo necesitamos especificar la distribución oculta $p(\mathbf{h})$. Una opción conveniente es una normal

$$p\left(\mathbf{h}\right) = \mathcal{N}\left(\mathbf{h}|\boldsymbol{\mu}_0, \mathbf{\Sigma}_0\right)$$

Se puede demostrar que la expresibilidad del modelo no requiere una media y covarianza generales, por lo que

$$p\left(\mathbf{h}\right) = \mathcal{N}\left(\mathbf{h}|\mathbf{0}, \mathbf{I}\right)$$

y puesto que la transformación lineal de una gaussiana lo sigue siendo, resulta:

$$p\left(\mathbf{v}\right) = \int p\left(\mathbf{v}|\mathbf{h}\right) p\left(\mathbf{h}\right) d\mathbf{h} = \mathcal{N}\left(\mathbf{v}|\mathbf{c}, \mathbf{F}\mathbf{F}^{\mathsf{T}} + \mathbf{\Psi}\right)$$

AF como distribución gaussiana con un rango reducido

Es habitual restringir $\mathbf{\Psi}$ para que sea diagonal; en caso contrario, podrían ignorarse las componentes latentes con $\mathbf{F} = \mathbf{0}$ y modelar cualquier covarianza.

$$p(\mathbf{v}) = \mathcal{N}\left(\mathbf{v}|\mathbf{c}, \mathbf{F}\mathbf{F}^{\mathsf{T}} + \mathbf{\Psi}\right)$$

Resulta así que se puede pensar en AF como en una distribución gaussiana con rango reducido. Si la covarianza no estuviera constreñida, el número de parámetros sería $D(D+1)$; en AF, el número de parámetros es $D(H+1)$, que es significativamente menor si $H \ll D$.

PCA probabilístico

$$\mathbf{\Psi} = \sigma^2\mathbf{I}$$

Análisis de factores

$$\mathbf{\Psi} = \mathrm{diag}\left(\psi_1, \ldots, \psi_D\right)$$

Algoritmo EM

Es fácil demostrar que el valor óptimo de $\mathbf{c}$ es la media de los datos $\bar{\mathbf{v}}$. Para los restantes, hay que considerar la energía:

$$-\sum_{n=1}^{N}\left\langle \frac{1}{2}(\mathbf{v}^n-\bar{\mathbf{v}}-\mathbf{F}\mathbf{h})^{\mathsf{T}}\boldsymbol{\Psi}^{-1}(\mathbf{v}^n-\bar{\mathbf{v}}-\mathbf{F}\mathbf{h})\right\rangle_{q(\mathbf{h}|\mathbf{v}^n)}-\frac{N}{2}\log\det(\boldsymbol{\Psi})$$

Tras algo de trabajo

$$\mathbf{F}=\mathbf{A}\mathbf{H}^{-1}$$

donde $\mathbf{A}=\sum_n(\mathbf{v}^n-\bar{\mathbf{v}})\langle\mathbf{h}\rangle^{\mathsf{T}}_{q(\mathbf{h}|\mathbf{v}^n)}$ y $\mathbf{H}=\sum_n\left\langle\mathbf{h}\mathbf{h}^{\mathsf{T}}\right\rangle_{q(\mathbf{h}|\mathbf{v}^n)}$, y

$$\boldsymbol{\Psi}=\frac{1}{N}\mathrm{diag}\Big(\sum_n(\mathbf{v}^n-\bar{\mathbf{v}})(\mathbf{v}^n-\bar{\mathbf{v}})^{\mathsf{T}}-\mathbf{F}\mathbf{A}^{\mathsf{T}}\Big)$$

(Barber también presenta otro algoritmo alternativo que no vamos a tratar.)

Probabilistic PCA

In the special case that

$$\mathbf{\Psi} = \sigma^2 \mathbf{I}$$

the optimal parameters have a closed form:

$$\sigma^2 = \frac{1}{D-H} \sum_{j=H+1}^{D} \lambda_j$$

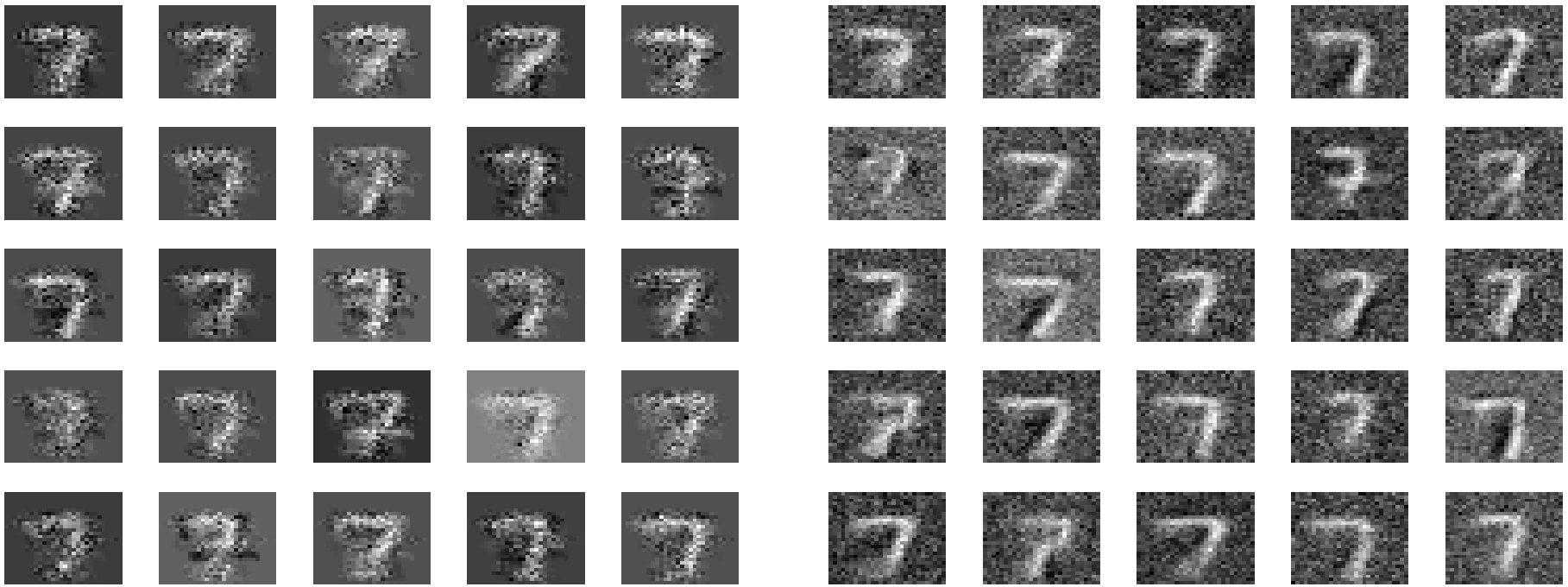
$$\mathbf{F} = \mathbf{U}_H \left(\mathbf{\Lambda}_H - \sigma^2 \mathbf{I}_H \right)^{\frac{1}{2}}$$

where $\mathbf{U}_H$, $\mathbf{\Lambda}_H$ are the eigenvectors and corresponding eigenvalues of the sample covariance $\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{v} - \bar{\mathbf{v}}) (\mathbf{v} - \bar{\mathbf{v}})^{\mathsf{T}}$.

The single-shot training nature of PPCA makes it an attractive algorithm and also gives a useful initialisation for Factor Analysis.

Classical PCA is recovered in the limit $\sigma^2 \to 0$.

PPCA versus FA



(a) Factor Analysis

(b) PPCA

Figure: **(a)**: 25 samples from the learned FA model of a dataset of handwritten '7s'. Note how the noise variance depends on the pixel, being zero for pixels on the boundary of the image. **(b)**: 25 samples from the learned PPCA model.

Mixtura de Analizadores de Factores

Una de las ventajas de los modelos probabilísticos es que se pueden combinar para construir modelos más complejos. En este caso tendríamos una mixtura de FA. Seguimos aquí la sección 28.3.3 del libro de Murphy y adoptamos su notación, de manera que $\mathbf{v}$, $\mathbf{h}$ y $\mathbf{F}$ se corresponden con $\mathbf{x}$, $\mathbf{z}$ y $\mathbf{W}$.

Modelo generativo

Primero muestrea un indicador oculto $m_n \in \{1, \ldots, K\}$ para elegir el cluster que genera el dato. Si $m_n = k$, muestreamos los factores ocultos $\mathbf{z}_n$ de una gaussiana y los pasamos a través de $\mathbf{W}_k$, que devuelve valores en el espacio visible. Por último, se añade ruido.

$$\mathbf{x}_n = \boldsymbol{\mu}_k + \mathbf{W}_k \mathbf{z}_n + \epsilon_n$$

Mixtura de FA: algoritmo EM

$$p\left(\mathbf{x}_n|\mathbf{z}_n, m_n = k\right) = \mathcal{N}\left(\mathbf{x}_n|\mathbf{W}_k\mathbf{z}_n + \boldsymbol{\mu}_k, \boldsymbol{\Psi}\right)$$
$$p\left(\mathbf{z}_n\right) = \mathcal{N}\left(\mathbf{z}_n|\mathbf{0}, \mathbf{I}\right)$$
$$p\left(m_n\right) = \mathbf{Cat}(m_n)$$

La distribución en el espacio visible resulta ser:

$$p\left(\mathbf{x}\right) = \sum_k \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{W}_k\mathbf{W}_k^\mathsf{T} + \boldsymbol{\Psi}\right)$$

Los parámetros de este modelo se pueden ajustar utilizando el algoritmo EM, a lo que dedicaremos uno de los ejercicios de la tarea de este tema.