UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

MIXTURES GAUSSIANES I ALGORISME EM

Autora: Andrea Baena Espejo

Director: Dr. Josep Fortiana Gregori
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 24 de gener de 2022

Abstract

In a statistical context, we can define a mixture model as a probabilistic model used to represent the presence of subpopulations within the same population. However, we can also say that a mixture model corresponds to a distribution (formed by a convex linear combination of other distributions) that represents the probability distribution of an observation in a population. Mixture models are used to create statistical inferences, approximations, and predictions about the properties of subpopulations based on observations made about the population studied, without the need to identify the corresponding subpopulation of each observation. In this project, we will study a particular case of mixture models: the Gaussian mixture model (mixture of multivariate Gaussian distributions).

The EM algorithm is a method that allows us to estimate the parameters of a statistical model when the data is incomplete or when the model contains unknown variables. In the case of mixture models, the unknown variables are those that tell us which component generated each observation in the sample. In this project, we will study the EM algorithm from different points of view, and we will use it to estimate the parameters of the Gaussian mixture model.

Resum

En un context estadístic, podem definir un model de mixtura com un model probabilístic que serveix per representar la presència de subpoblacions dins una mateixa població. No obstant, també podem dir que un model de mixtura correspon a una distribució (formada per una combinació lineal convexa d'altres distribucions) que representa la distribució de probabilitat d'una observació qualssevol d'una població. Els models de mixtura són usats per crear inferències estadístiques, aproximacions i prediccions sobre les propietats de les subpoblacions a partir de les observacions realitzades sobre la població estudiada, sense necessitat d'informació que n'identifiqui la subpoblació corresponent de cadascuna d'elles. En aquest treball estudiarem un cas particular dels models de mixtura: el model de mixtura de distribucions gaussianes multivariants.

L'algorisme EM és un mètode que ens permet estimar els paràmetres d'un model estadístic quan les dades són incompletes o quan el model conté variables desconegudes. En el cas dels models de mixtura, les variables desconegudes són aquelles que ens indiquen quina component ha generat cada observació de la mostra. En aquest treball estudiarem l'algorisme EM des de diferents punts de vista i l'emprarem per estimar els paràmetres del model de mixtura de gaussianes.

2020 Mathematics Subject Classification. 62H10, 62H12, 62H30

Agraïments

En primer lloc, vull agrair al meu tutor, el Dr. Josep Fortiana, per guiar-me durant tot el treball i haver estat sempre disposat a ajudar-me i resoldre els meus dubtes. Gràcies al seu esforç i la seva dedicació, el camí ha estat molt més fàcil.

També vull agrair a la meva família, amics i parella per haver estat al meu costat i haver-me donat el seu suport durant tots aquests anys. Sense ells no hauria estat possible arribar fins aquí.

Índex

1 Introducció

Tal com indica el títol, en aquest treball ens centrarem en l'estudi de dos grans temes: el model de mixtura de distribucions gaussianes i l'algorisme EM. L'objectiu d'aquest primer apartat és introduir tots dos conceptes i establir la base sobre la que treballarem en detall a la resta de seccions.

Quan analitzem un conjunt de dades, sovint fem hipòtesis que ens permeten simplificar els càlculs, com ara assumir que totes les observacions d'una mostra provenen d'una distribució específica (per exemple, una distribució gaussiana). Tanmateix, en molts casos, suposar que cada observació prové de la mateixa distribució unimodal és massa restrictiu. Sovint, les dades que estem intentant modelar són més complexes. Per exemple, poden ser multimodals, que contenen diverses regions amb massa probabilitat elevada. En aquests casos, podem fer ús dels **models de mixtura**.

Una mixtura de K distribucions contínues amb densitats $f_1, \dots f_K$, en **proporcions** $\pi_1, \dots, \pi_K$ ve donada per

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot f_k(x, \Theta_k). \tag{1.1}$$

on els **coeficients** de la mixtura han de satisfer

$$0 \leq \pi_k \leq 1 \quad \text{i} \quad \sum_{k=1}^{K} \pi_k = 1. \tag{1.2}$$

La mateixa expressió és vàlida per a una mixtura de K distribucions discretes. Nosaltres estudiarem el cas en què les funcions f_k són densitats de distribucions gaussianes multivariants. Aquest model és el que coneixem com a **model de mixtura de gaussianes**.

Un concepte que utilitzarem àmpliament durant tot el treball és el de **variable latent**. Diem que una variable és latent en un model de probabilitat, quan aquesta no es pot observar directament. Contemplar variables no observades per ajudar amb els càlculs és un truc molt habitual. Els models de mixtura es poden interpretar en termes de variables latents de la següent manera: per a cada variable observada x, podem coniderar una variable latent $z = (z_1, \dots, z_K)$, tal que tal que totes les components de z siguin 0 tret d'una, que prengui el valor 1, de manera que

$$p(z_k = 1) = \pi_k. \tag{1.3}$$

$$p(\, x \mid z_k = 1 \,) = f_k(x, \Theta_k). \tag{1.4}$$

Sovint ens referim a la variable z com a **component d'origen**, ja que ens dona informació de la component de la mixtura a la qual pertany l'observació x. La funció de densitat de probabilitat de x s'obté per marginalització

$$p(x) = \sum_{k=1}^{K} p(\, x \mid z_k = 1\,) \cdot p(z_k = 1) = \sum_{k=1}^{K} \pi_k \cdot f_k(x, \Theta_k). \tag{1.5}$$

Finalment, podem definir el valor

$$\gamma(z_k) \equiv p(\, z_k = 1 \mid x\,) \tag{1.6}$$

que, pel teorema de Bayes de les probabilitats condicionades, es pot calcular com

$$\gamma(z_k) = \frac{p(\, x \mid z_k = 1\,) \cdot p(z_k = 1)}{p(x)} = \frac{\pi_k \cdot f_k(x, \theta_k)}{\sum_{k=1}^{K} p(\, x \mid z_k = 1\,) \cdot p(z_k = 1)}. \tag{1.7}$$

Per la definició (1.7) de $\gamma(z_k)$, se satisà que

$$\sum_{k=1}^{K} \gamma(z_k) = 1. \tag{1.8}$$

Veiem que π_k representa la probabilitat a *priori* de $z_k = 1$ i $\gamma(z_k)$ la corresponent probabilitat posterior, un cop observada x.

Una de les aplicacions dels models de mixtura és l'anàlisi de clústers. **L'anàlisi de clústers** és una tècnica per classificar un conjunt d'observacions en grups (clústers) segons les característiques que són similars entre elles. Al pròxim apartart del treball parlarem d'un algorisme (no probabilístic) molt usat en la tasca d'agrupar dades: l' **algorisme *K-mitjanes***. Al llarg de la resta de seccions, anirem veient com encaixen els models de mixtura en aquest àmbit.

En el tercer apartat del treball veurem la formulació del model de mixtura de gaussianes en termes de variables latents. Després, plantejarem el **mètode d'estimació de màxima versemblança** per inferir els paràmetres del model (primer, per a una mostra extreta d'una distribució gaussiana multivariant i, posteriorment, per una mostra modelada a través d'una mixtura de gaussianes). Aquest estudi desembocarà en una primera aproximació a l'anomenat **algorisme EM**. Aquest algorisme constitueix una tècnica general per trobar estimadors de màxima versemblança en models de mixtura de distribucions (més generalment, en models amb variables latents).

A l'apartat 4 estudiarem en detall l'algorisme EM. Partirem del cas concret d'aplicació en el model de mixtura de gaussianes i després, deduïrem el cas general. Finalment, demostrarem que l'algorisme *K-mitjanes* introduït a l'apartat 2, correspon a un límit particular de l'algorisme EM.

A l'últim apartat teòric del treball, parlem de l'algorisme EM de manera més generalitzada. En veurem també diverses variants: primer explicarem com podem procedir quan algun dels passos de l'algorisme no es pot resoldre de manera convencional i, després, veurem una versió incremental més simple que es pot utilitzar si se satisfan certes hipòtesis.

Aquest treball tindrà com suport una part pràctica, en la qual il·lustrarem, usant el llenguatge de programació R, diversos exemples d'aplicació en dades reals i simulades dels algorismes estudiats.

2 Algorisme *K-mitjanes*

En aquesta primera secció del treball parlarem de l'algorisme *K-mitjanes*. A grans trets, podem dir que aquest algorisme és un mètode iteratiu que ens permet agrupar un conjunt d'observacions segons algunes de les seves característiques. Per exemple, suposem que hem mesurat l'amplada i la llargària dels pètals d'un conjunt de flors d'una certa població. Suposem també que dins d'aquesta població hi ha flors de tres tipus (subpoblacions) diferents, que no sabem distingir a simple vista. En aquest cas, podríem aplicar l'algorisme als valors d'amplada i llargària obtinguts, per tal d'intentar agrupar les dades segons els tres grups de flors existents a la població.

En els pròxims subapartats en parlarem en detall. Primer, donarem la definició de l'algorisme. Després, n'estudiarem certes propietats com la convergència i el valor òptim de clústers en els quals agrupar un conjunt de dades, quan no sabem de quants models diferents provenen.

2.1 Descripció de l'algorisme

Suposem que tenim un conjunt d'observacions $(x_1, \ldots, x_N)$ d'una població qualsevol. Suposem també que dins d'aquesta població hi ha K subpoblacions diferents però que, de la mostra que hem extret, no sabem a quina subpoblació correspon cada observació. Introduïm la següent nomenclatura:

- Cada subpoblació o grup s'anomena **clúster**. Denotem els K clústers per $\{S_1, \ldots, S_K\}$.
- Cada clúster el definim per un **prototip** o **centroide**. Escrivim el conjunt de tots els prototips en un vector $\mu = (\mu_1, \ldots, \mu_K)$.

En primer lloc, per a cada observació x_n del nostre conjunt de dades, hem de definir una nova variable discreta $r_n = (r_{n1}, \ldots, r_{nK})$ tal que

$$r_{nk} \in \{0, 1\} \quad \text{i} \quad r_{nk} = \begin{cases} 1 & \text{si} \quad x_n \in S_k \\ 0 & \text{si} \quad x_n \notin S_k. \end{cases} \tag{2.1}$$

Aquestes variables s'anomenen **assignacions**, ja que indiquen a quin clúster pertany cada observació. Les podem escriure conjuntament en una matriu d'incidència r de dimensió N × K. D'aquesta manera, a la fila n-èssima de la matriu tenim les assignacions corresponents al punt x_n. Aquestes variables satisfan les següents propietats

$$\sum_{k=1}^{K} r_{nk} = 1 \tag{2.2}$$

juntament amb

$$\sum_{n=1}^{N} r_{nk} = |S_k| \quad (\forall n). \tag{2.3}$$

L'algorisme *K-mitjanes* ens serveix per estimar els valors dels prototips μ_k i de les assignacions r_n . El procediment consisteix a resoldre un problema d'optimització, sent la **funció objectiu** la suma de les distàncies al quadrat de cada punt de la mostra al prototip del seu clúster. És a dir

$$J(\mu, r) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \cdot \|x_n - \mu_k\|^2. \tag{2.4}$$

Per optimitzar aquesta funció, hem de seguir els següents passos:

1. Inicialitzar els prototips μ (a l'apartat 2.2 veurem diferents mètodes per fer-ho) i les assignacions r (aleatòriament). Avaluar la funció J als valors que hàgim inicialitzat.

2. **Pas d'assignació:** minimitzar la funció J respecte r: per a cada observació x_n, establim

$$r_{nk}^* = \begin{cases} 1, & \text{si } \|x_n - \mu_k\| < \|x_n - \mu_j\| \ \ (\forall j \neq k) \\ 0, & \text{en cas contrari.} \end{cases} \tag{2.5}$$

3. **Pas d'actualització:** minimitzar la funció J respecte μ

$$J(\mu, r) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \cdot \|x_n - \mu_k\|^2 = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \cdot \|x_n - \mu_k\| = \sum_{k=1}^{K} J_k. \tag{2.6}$$

   Com que $J_k \geq 0$ $(\forall k)$, fixem k

$$0 = \frac{\partial J_k(\mu_k, r_{nk})}{\partial \mu_k} = -2 \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \cdot (x_n - \mu_k)$$

$$\mu_k^* = \frac{\sum_{n=1}^{N} r_{nk} \cdot x_n}{\sum_{n=1}^{N} r_{nk}} = \frac{1}{|S_k|} \sum_{x_n \in S_k} x_n. \tag{2.7}$$

   Per tant, el nou prototip de cada clúster és la mitjana dels punts que hi han estat assignats en el pas previ.

4. Avaluar la funció J als nous valors obtinguts μ^*, r^*. Si la funció ha variat en menys d'un ϵ petit (tolerància) respecte al valor previ, aturar l'algorisme. En cas contrari, tornar al pas 2.

2.2 Convergència

Proposició: l'algorisme *K-mitjanes* convergeix a un mínim local de la funció objectiu $J(\mu, r)$ definida a (2.4).

Demostració: per provar-ho, demostrarem que la funció J decreix monòtonament a cada iteració, tant en el pas d'assignació com en el pas d'actualització.

- Pas d'assignació

  Calculem la diferència entre la funció J abans i després de fer les noves assignacions:

$$\begin{aligned} J(\mu, r^*) - J(\mu, r) &= \sum_{n=1}^{N} \left(\sum_{k=1}^{K} r_{nk}^* \cdot \|x_n - \mu_k\|^2 - \sum_{k=1}^{K} r_{nk} \cdot \|x_n - \mu_k\|^2 \right) \\ &= \sum_{n=1}^{N} \Delta_n \leq 0. \end{aligned} \tag{2.8}$$

  La desigualtat anterior se satisfà ja que, per la tria que hem fet de r_{nk}^* (2.5), tenim que $\Delta_n \leq 0$ $(\forall n)$.

- Pas d'actualització

  Calculem la diferència entre la funció J abans i després d'actualitzar els prototips:

$$\begin{aligned} J(\mu^*, r^*) - J(\mu, r^*) &= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} r_{nk}^* \cdot \|x_n - \mu_k^*\|^2 - \sum_{n=1}^{N} r_{nk}^* \cdot \|x_n - \mu_k\|^2 \right) \\ &= \sum_{n=1}^{N} \Delta_k \leq 0. \end{aligned} \tag{2.9}$$

  La desigualtat anterior se satisfà ja que, per la tria que hem fet de μ_k^* (2.7), tenim que $\Delta_k \leq 0$ $(\forall k)$.

□

El mínim local al qual convergeixi l'algorisme pot no ser necessàriament un mínim global. Llavors, l'algorisme pot convergir a mínims locals diferents en funció dels prototips inicials que hàgim triat. En aquest sentit, podem destacar els següents mètodes per inicialitzar els prototips:

- **Aleatòriament**

  Triem K punts qualssevol de l'espai de dades o bé els seleccionem d'entre el nostre conjunt d'observacions. Aquesta segona opció és més habitual.

- **k-means++**

  Aquest mètode d'inicialització es basa en repartir els prototips inicials per cobrir la major part possible de l'espai de dades. L'objectiu és intentar minimitzar el nombre d'iteracions necessàries perquè l'algorisme convergeixi. El procediment és el següent:

  1. Triar el primer centroide μ_1 aleatòriament d'entre els punts del nostre conjunt d'observacions. Tots els punts poden ser escollits amb la mateixa probabilitat.

  2. Per a cada punt x_n, calcular la seva distància $D(x_n)$ al centroide més proper que ja hàgim inicialitzat.

  3. Triar el següent centroide de manera aleatòria entre els punts del nostre conjunt d'observacions, utilitzant una distribució de probabilitat ponderada tal que el punt x_n és escollit amb una probabilitat proporcional a $D(x_n)^2$.

  Repetir els passos 2 i 3 fins que hàgim inicialitzat els K centroides.

2.3 Estimació del K òptim

Un dels principals problemes d'aquest algorisme és decidir quin valor de grups K hem de fer servir com a paràmetre d'entrada. Vegem alguns dels mètodes més coneguts per trobar-ne el valor òptim.

- **Mètode del colze** (*elbow method*)

  Aquesta tècnica consisteix a executar l'algorisme *K-mitjanes* per a diversos valors de K dins d'un rang i comparar la suma dels errors al quadrat entre cada observació i el centroide que li ha estat assignat. Aquesta suma s'anomena *within sum of squares* i es calcula com

$$WSS = \sum_{k=1}^{K} \sum_{x_n \in S_k} \|x_n - \mu_k\| \tag{2.10}$$

  Observem que aquest valor correspon a la funció J (2.4), avaluada a les assignacions i els prototips obtinguts de l'algorisme. Podem pensar que el valor de l'expressió (2.10) ens dona una mesura global de l'error comès en l'agrupació. En general, en augmentar el nombre de grups K, l'error disminueix, ja que els grups són més petits i, en conseqüència, també ho són les distàncies. Si fem un gràfic per veure la relació entre K i el valor obtingut en la WSS, podem prendre com a valor òptim de clústers aquell punt en el qual veiem una reducció dràstica de l'error.

  Exemple: suposem que per una mostra donada, executem l'algorisme EM prenent $K = (1, \ldots, 10)$ i que, per a cada K, calculem el valor de la suma (2.10). Suposem que obtenim la següent representació gràfica:

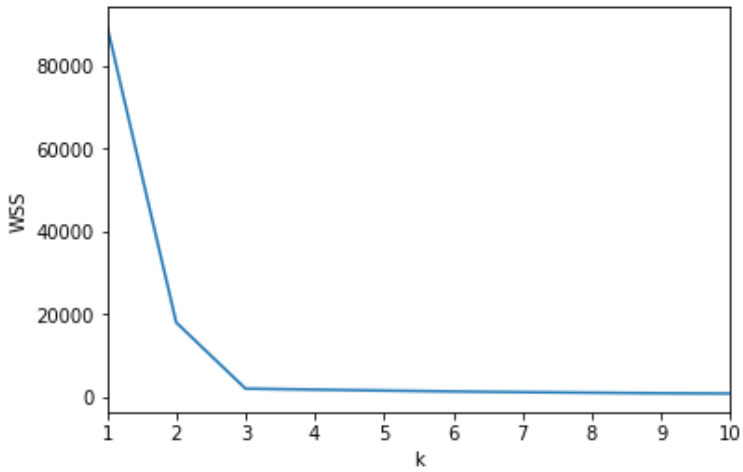

Figura 1: Gràfica K - WSS

A partir d'aquest resultat, podem veure clarament que el valor òptim que hem de triar és K=3. Tanmateix, en altres exemples la tria pot no ser tan evident. Per a aquests casos, podem recórrer al segon mètode de selecció, el *silhouette method*.

- **Mètode de la silueta** (*silhouette method*)

  En aquest mètode hem d'analitzar el valor de la silueta (*silhouette value*) de cada observació. El valor de la silueta compara com és de semblant un punt al seu clúster (cohesió) respecte a la resta de clústers (separació). Així, per a cada punt x_n de la nostra mostra, definim aquest valor com

$$s(n) = \frac{b(n) - a(n)}{\max\{a(n), b(n)\}}. \tag{2.11}$$

  – **Cohesió**: per una banda, $a(n)$ mesura la semblança de x_n amb el seu grup. Aquest valor es calcula com la mitjana de les distàncies de x_n a la resta de punts del clúster. Si suposem que $x_n \in S_k$, llavors tenim

$$a(n) = \frac{1}{|S_k| - 1} \sum_{x_j \in C_i} \|x_i - xj\|. \tag{2.12}$$

  Observem que $a(n) > 0$. Ens interessa obtenir un valor petit d'aquesta mesura, ja que això indica que el punt x_n és semblant al seu clúster i que, per tant, hem fet una bona assignació.

  – **Separació**: per l'altra banda, $b(n)$ mesura la diferència entre x_n i els punts de la resta de grups. Aquest valor es calcula com la mínima de les distàncies mitjanes de x_n a tots els punts de la resta de grups. És a dir

$$b(n) = \min_{k' \neq k} \frac{1}{|Sk'|} \sum_{x_j \in C_{k'}} \|x_i - x_j\|. \tag{2.13}$$

Observem també que $b(n) > 0$. Ens interessa que aquest valor sigui gran, ja que això indica que no hi ha cap altre grup que sigui molt semblant a x_n i que, per tant, hem fet una bona assignació. El clúster $S_{k'}$ per al qual s'assoleix aquest mínim l'anomenem clúster veí de x_n, ja que és el següent grup que millor s'ajusta al punt.

Definim $s(n) = 0$ si x_n és l'únic punt en el seu clúster. Aquesta restricció evita que el nombre de clústers augmenti significativament creant grups d'un sol punt. Veiem que $s(n) \in [-1, +1]$ i que, per la definició que hem donat de $a(n)$ i $b(n)$, volem que aquest valor sigui el més gran possible (proper a 1). Si per contra, hi ha molts punts amb un valor negatiu de $s(n)$, significa que hem creat massa (o massa pocs) grups.

Com que hem definit $s(n)$ per a cada punt de la mostra, per tal de poder fer una anàlisi global de l'error en l'agrupació, prenem el valor mitjà de tots els valors de silueta

$$SC = \frac{1}{N} \sum_{n=1}^{N} s(n). \tag{2.14}$$

Si fem un gràfic per veure la relació entre el nombre de grups K i el valor del SC, podem prendre com a K òptim aquell punt en el qual el *Silhouette Score* assoleix el seu màxim.

Exemple: suposem que per una mostra donada, executem l'algorisme EM prenent $K = (1, \ldots, 10)$ i que, per a cada K, calculem el valor de la mitjana (2.14). Suposem que obtenim la següent representació gràfica:

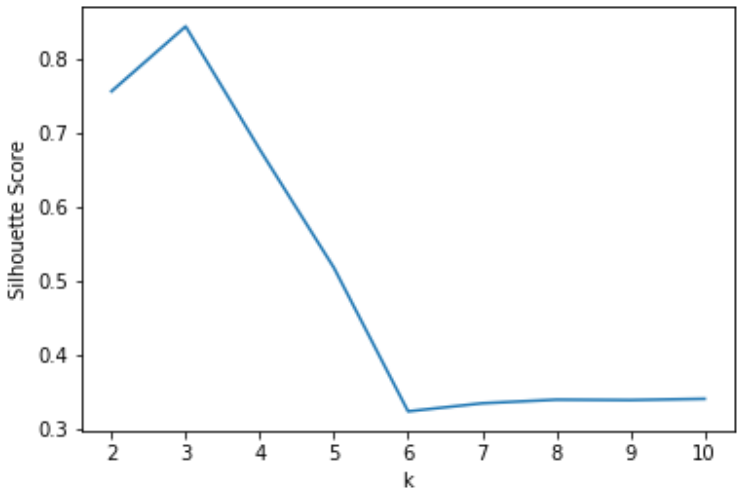

Figura 2: Gràfica K-SC

A partir d'aquest resultat, podem veure clarament que el valor òptim que hem de triar és K=3.

No hem de considerar aquests dos mètodes com a tècniques alternatives l'una de l'altra. De fet, la millor estratègia consisteix a utilitzar-los complementàriament amb l'objectiu de tenir la màxima informació possible per fer la tria del valor de K òptim.

2.4 Validació de la bondat de l'agrupació

Suposem que tenim unes dades que volem agrupar i que, per fer-ho, hem aplicat l'algorisme *K-mitjanes* per un cert valor K de clústers. Suposem que volem interpretar els resultats obtinguts i validar la coherència i la bondat de l'agrupació. Per fer-ho, podem emprar el mètode de la silueta, vist a l'apartat anterior. A més de servir-nos per estimar el nombre òptim del paràmetre K, aquest mètode també ens proporciona un marc per interpretar i validar la coherència dels resultats dins de l'anàlisi de clústers.

Procediment

El mètode de la silueta consisteix a representar, mitjançant un diagrama de barres, el valor de la silueta s_n de cada observació x_n. La gràfica obtinguda ens indica en quina mesura podem considerar correcta la classificació que ha fet l'algorisme de cada objecte, segons el següent criteri:

- Si $s_n > 0$ significa que l'observació x_n està ben agrupada. Com més proper a 1 sigui el valor de la silueta, més probable és que l'observació s'hagi assignat al clúster correcte.
- Si $s_n < 0$ significa que l'observació x_n s'ha assignat a un clúster incorrecte. Tanmateix, un valor negatiu de la silueta ens pot indicar la presència de valors atípics, com veurem després a un exemple.
- Si $s_n = 0$ significa que l'observació està entre dos clústers.

Vegem ara dos exemples d'utilització d'aquest mètode per analitzar els resultats obtinguts en l'agrupació amb *K-mitjanes*.

Exemple 1

Hem extret una mostra de llet de 25 espècies de mamífers diferents i hem analitzat la seva composició d'aigua, proteïna, lactosa, greix i minerals. Després, hem aplicat l'algorisme *K-mitjanes* per classificar aquestes espècies en grups, segons com és de semblant la composició de la seva llet respecte a les cinc variables mesurades. Abans, hem usat el mètode del colze amb $K = \{1, \ldots, 20\}$ per determinar el nombre de grups òptim a emprar en l'algorisme. Hem obtingut el valor de K=4 i l'hem usat com a paràmetre d'entrada de l'algorisme. Un cop executat i obtinguts els resultats de l'agrupació, volem validar-ne la coherència. Per fer-ho, calculem el valor de la silueta s_n de cada observació x_n i els representem gràficament, agrupant-los per clústers, segons les assignacions que ha realitzat l'algorisme. La gràfica resultant es mostra a continuació:

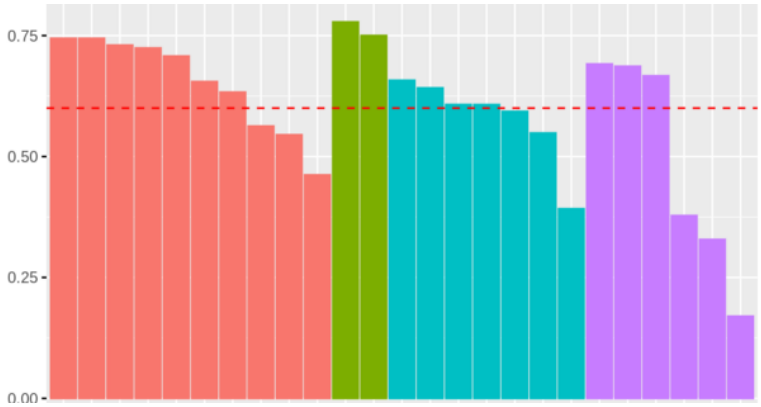

Figura 3: Gràfic de silueta 1

Veiem que no hi ha amplades de silueta negatives i que la majoria dels valors estan per sobre de 0,5. La mitjana de tots els valors de silueta és de 0,6. Per tant, a *priori* no hi ha cap indici que ens indiqui errors en l'agrupació.

Exemple 2

A partir de la base de dades del zoo de Barcelona, hem recollit els valors de certes característiques per a cada espècie d'animal que hi ha en captiveri. Sabem que aquestes espècies són de tres tipus: peixos, insectes i mamífers. Hem aplicat l'algorisme *K-mitjanes* usant $K = 3$ i, posteriorment, hem representat la gràfica de les siluetes obtingudes:

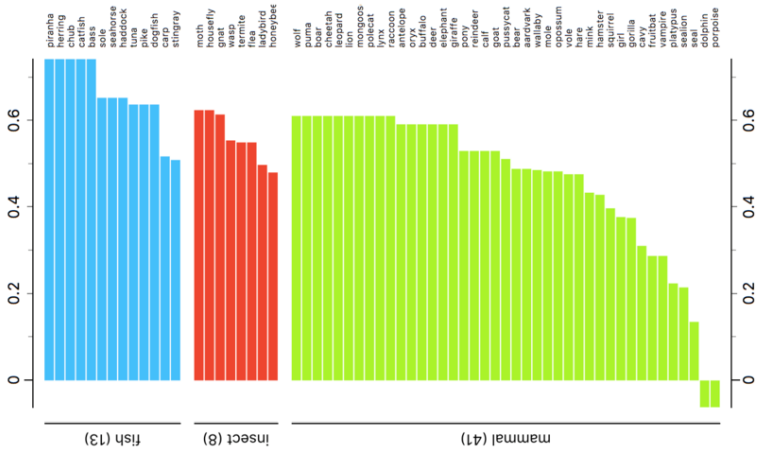

Figura 4: Gràfic de silueta 2

El dofí i la marsopa han estat classificats dins del grups dels mamífers. Tanmateix, la silueta els identifica com a valors atípics dins d'aquest grup. Això pot ser degut a que, per a les característiques que hem mesurat, els dofins i les marsopes són més semblants a les espècies d'un altre grup (deduïm, el de peixos) que no pas a les de mamífers.

3 Mixtures gaussianes

En aquest apartat parlem del model de mixtura de distribucions gaussianes (*Gaussian Mixture Model*). Veurem la seva formulació en termes de variables latents i parlarem també del mètode d'estimació de màxima versemblança per trobar-ne els paràmetres. Com a conclusió de l'apartat, deduirem una primera aproximació a l'algorisme EM.

3.1 Model de mixtura de gaussianes i variables latents

Tal com hem introduït a la primera secció del treball, sovint volem modelar dades que contenen diverses regions amb massa probabilitat elevada. Per a aquests casos, hem vist que podem fer ús dels models de mixtura de distribucions. Vegem-ne ara el cas concret en el qual les densitats que formen les components de la mixtura són gaussianes multivariants. L'expressió (1.1) es tradueix en

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\, x \mid \mu_k, \Sigma_k \,) \tag{3.1}$$

on

$$\mathcal{N}(\, x \mid \mu, \Sigma \,) = \frac{1}{(2\pi)^{d/2}\ |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T\ \Sigma^{-1}\ (x-\mu) \right\}. \tag{3.2}$$

Recordem que els coeficients de la mixtura han de satisfer

$$0 \leq \pi_k \leq 1 \quad \text{i} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{3.3}$$

A continuació, per a cada variable observada x, podem considerar una variable latent $z = (z_1, \ldots, z_K)$, tal que totes les components de z siguin 0 tret d'una, que prengui el valor 1, de manera que

$$p(z_k = 1) = \pi_k \tag{3.4}$$

$$p(\, x \mid z_k = 1 \,) = \mathcal{N}(\, x \mid \mu_k, \Sigma_k \,). \tag{3.5}$$

Llavors, per la definició de densitat de probabilitat condicionada, tenim que

$$p(\, x \mid z \,) = \prod_{k=1}^{K} \mathcal{N}(\, x \mid \mu_k, \Sigma_k \,)^{z_k} \tag{3.6}$$

i la distribució de z com

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{3.7}$$

La variable z de vegades es coneix com a **component d'origen**, ja que ens dona informació de la component de la mixtura a la qual pertany l'observació x. Finalment, la densitat de probabilitat de x s'obté per marginalització

$$p(x) = \sum_{k=1}^{K} p(\, x \mid z_k = 1\,) \cdot p(z_k = 1) = \pi_k \cdot \mathcal{N}(\, x \mid \mu_k, \Sigma_k\,). \tag{3.8}$$

Aquest últim resultat coincideix amb l'expressió (3.1). D'aquesta manera, hem trobat una formulació del model de mixtura de gaussianes en termes d'una variable latent explícita. Tot i que a *priori* pot semblar que no hem guanyat gaire amb aquesta discussió, ara podem treballar amb la densitat de probabilitat conjunta de $\{x, z\}$

$$p(x, z) = p(\, x \mid z\,) \cdot p(z) = \prod_{k=1}^{K} [\, \pi_k \cdot \mathcal{N}(\, x \mid \mu_k, \Sigma_k\,)\,]^{z_k} \tag{3.9}$$

en comptes d'amb la densitat de probabilitat marginal de x. Això ens serà molt útil en les properes seccions del treball, quan vulguem resoldre el problema de màxima versemblança aplicant l'algorisme EM, ja que ens simplificarà de manera notable els càlculs.

També adaptem l'expressió (1.7) pel cas particular de la mixtura de gaussianes i obtenim que

$$\gamma(z_k) \equiv p(\, z_k = 1 \mid x\,) = \frac{\pi_k \cdot \mathcal{N}(\, x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\, x \mid \mu_j, \Sigma_j)}. \tag{3.10}$$

Per la definició de probabilitat, aquests valors satifan que

$$\sum_{k=1}^{K} \gamma(z_{nk}) = 1. \tag{3.11}$$

Per tant, π_k representa la probabilitat a *priori* de $z_k = 1$ i $\gamma(z_k)$ la corresponent probabilitat posterior, un cop observada x. Podem interpretar les probabilitats posteriors $\gamma(z_k)$ com a assignacions probabilístiques: un cop hem observat x, la probabilitat que aquesta observació hagi estat generada per la component k de la mixtura, és $\gamma(z_k)$. Aquest valor jugarà també un paper molt important quan parlem de l'algorisme EM.

3.2 Estimació de màxima versemblança

Comencem aquesta secció recordant la definició de la funció de versemblança. Sigui $X_1 \dots X_N$ una mostra aleatòria (no necessàriament simple) d'una població X amb funció de densitat $p(\, X \mid \Theta)$, on Θ és el conjunt de paràmetres del model. Per a cada mostra particular $\mathcal{X} = (x_1 \dots x_N)$, es defineix la funció de versemblança com la funció de densitat

conjunta de $X_1 \dots X_N$ avaluada a $(x_1 \dots x_N)$

$$\mathcal{L}(\mathcal{X}, \Theta) = p(\, \mathcal{X} \mid \Theta\,). \tag{3.12}$$

En el nostre cas, treballarem sempre amb mostres aleatòries simples. És a dir, assumirem que les variables aleatòries $X_1 \dots X_N$ són i.i.d. D'aquesta manera, la funció de versemblança d'una mostra $\mathcal{X} = (x_1, \dots, x_N)$ es pot escriure com

$$\mathcal{L}(\mathcal{X}, \Theta) = \prod_{n=1}^{N} p(\, x_n \mid \Theta\,). \tag{3.13}$$

En endavant, quan parlem de la funció de versemblança, ens referirem sempre a aquesta última expressió.

Donat un conjunt d'observacions $\mathcal{X}$, l'estimació de màxima versemblança consisteix a trobar els paràmetres Θ que maximitzen la funció (3.13). Sovint, però, s'acostuma a maximitzar el logaritme de la versemblança, ja que els càlculs, en general, són més senzills

$$0 = \frac{\partial \log \mathcal{L}(\mathcal{X}, \Theta)}{\partial \Theta} = \frac{\partial \log p(\, \mathcal{X} \mid \Theta\,)}{\partial \Theta}. \tag{3.14}$$

En molts casos, aquest problema no es pot resoldre directament, perquè l'expressió resultant (3.14) és intractable analíticament. Per a aquests casos, cal emprar altres mètodes més sofisticats. N'estudiarem un d'ells, l'algorisme EM, en les properes seccions.

Ara que ja hem introduït el concepte de versemblança de forma general i hem explicat en què consisteix l'estimació màxim versemblant, passem a estudiar dos casos particulars: el de la distribució normal multivariant i el del model de mixtura de gaussianes, en aquest ordre.

Distribució gaussiana

Suposem que el nostre conjunt de dades $\mathcal{X} = (x_1, \dots, x_N)$ es regeix per una distribució gaussiana multivariant. A partir de l'expressió (3.2), calculem el logaritme de la funció de versemblança

$$\begin{aligned} \log \mathcal{N}(\, \mathcal{X} \mid \mu, \Sigma\,) &= \sum_{n=1}^{N} \log \frac{1}{(2\pi)^{d/2}\, |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x_n - \mu)^T\ \Sigma^{-1}\ (x_n - \mu) \right\} \\ &= -\frac{N\,d}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T\ \Sigma^{-1}\ (x_n - \mu). \end{aligned} \tag{3.15}$$

Maximitzem l'expressió anterior respecte cadascun dels paràmetres, prenent derivades parcials i igualant-les a 0:

- **Respecte μ**

$$0 = \frac{\partial \log \mathcal{N}(\mathcal{X} \mid \mu, \Sigma)}{\partial \mu} = \sum_{n=1}^{N} (x_n - \mu). \tag{3.16}$$

Reordenant els termes i aïllant μ, obtenim el següent resultat

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n. \tag{3.17}$$

És a dir, el paràmetre μ màxim versemblant és la mitjana mostral.

- **Respecte Σ**

$$0 = \frac{\partial \log \mathcal{N}(\mathcal{X} \mid \mu, \Sigma))}{\partial \Sigma} = \sum_{n=1}^{N} \left\{ (x_n - \mu)(x_n - \mu)^T - \Sigma \right\}. \tag{3.18}$$

Reordenem els termes i aïllem Σ

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^T. \tag{3.19}$$

És a dir, el paràmetre Σ màxim versemblant és la covariància mostral.

Model de mixtura de gaussianes

Suposem ara que volem modelar les nostres observacions $\mathcal{X} = (x_1, \ldots, x_N)$ utilitzant el model de mixtura de gaussianes. Siguin $\mathcal{Z} = (z_1, \ldots, z_N)$ les corresponents variables latents definides a (3.4). A partir de l'expressió (3.1), calculem el logaritme de la funció de versemblança

$$\log p(\mathcal{X} \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_n \mid \pi_k, \mu_k, \Sigma_k) \right\}. \tag{3.20}$$

El problema de maximitzar el logaritme de la versembança per a un model de mixtura de gaussianes resulta més complicat que en el cas d'una única distribució gaussiana multivariant. Aquesta dificultat addicional és deguda al fet que, en l'expressió (3.20) apareix el sumatori sobre k dintre del logaritme, provocant que aquest no actuï directament sobre les densitats gaussianes.

Tot i això, podem prosseguir calculant les derivades parcials respecte cadascun dels paràmetres i igualant-les a 0:

- **Respecte** μ_k

$$0 = \frac{\partial \log p(\,\mathcal{X} \mid \pi, \mu, \Sigma)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \cdot \mathcal{N}(\,x_n \mid \mu_k, \Sigma_k\,)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\,x_n \mid \mu_j, \Sigma_j\,)} \Sigma_k^{-1}(x_n - \mu_k). \tag{3.21}$$

Veiem que a la banda dreta de la igualtat apareixen les probabilitats posteriors $\gamma(z_{nk})$. Podem reescriure l'expressió anterior com

$$0 = \sum_{n=1}^{N} \gamma(z_{nk})\, \Sigma_k^{-1}(x_n - \mu_k). \tag{3.22}$$

Finalment, multiplicant per Σ_k i reordenant els termes, obtenim que

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, x_n \tag{3.23}$$

on hem definit

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{3.24}$$

Interpretem N_k com el nombre efectiu de punts modelats per la component k de la mixtura. La forma que pren μ_k és anàloga a la que obteníem en el cas de la distribució gaussiana multivariant a l'expressió (3.17). Ara, però, es tracta d'una mitjana ponderada, on en factor de ponderació del punt x_n ve donat per la probabilitat posterior $\gamma(z_{nk})$.

- **Respecte** Σ_k

$$0 = \frac{\partial \log p(\,\mathcal{X} \mid \pi, \mu, \Sigma)}{\partial \Sigma_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left[(x_n - \mu_k)(x_n - \mu_k)^T - \Sigma_k\right]. \tag{3.25}$$

Operant i reordenant termes, obtenim

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T. \tag{3.26}$$

També l'estimació del paràmetre Σ_k és anàloga al cas de la distribució gaussiana multivariant. La diferència es que, ara, cada punt està ponderat per la corresponent probabilitat posterior $\gamma(z_{nk})$ i al denominador trobem el nombre efectiu de punts assignats a la component k de la mixtura, N_k.

- **Respecte π_k.**

  Per a calcular la derivada parcial respecte els coeficients de la mixtura, cal tenir especial cura. Hem de tenir en compte que aquests paràmetres han de satisfer la condició (3.3). Per tal d'imposar aquesta restricció, utilitzem un multiplicador de Lagrange λ, que introduïm a l'expressió de la versemblancça de la següent manera

$$\log p(\, \mathcal{X} \mid \pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right). \tag{3.27}$$

  Ara, podem procedir com hem fet pels altres paràmetres: prenem la derivada parcial de l'expressió anterior respecte π_k i la igualem a 0

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\, x_n \mid \mu_k, \Sigma_k \,)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\, x_n \mid \mu_j, \Sigma_j \,)} + \lambda. \tag{3.28}$$

  Multipliquem l'expressió anterior per π_k

$$\begin{aligned} 0 &= \sum_{n=1}^{N} \frac{\pi_k \cdot \mathcal{N}(\, x_n \mid \mu_k, \Sigma_k \,)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\, x_n \mid \mu_j, \Sigma_j \,)} + \lambda \cdot \pi_k \\ &= \sum_{n=1}^{N} \gamma(z_{nk}) + \lambda \cdot \pi_k. \end{aligned} \tag{3.29}$$

  Ara, sumem respecte k

$$0 = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) + \lambda \cdot \sum_{k=1}^{K} \pi_k = N + \lambda \tag{3.30}$$

  on hem usat la propietat (3.11). Per tant, $\lambda = -N$. Substinuïm aquest resultat a l'expressió (3.29) i obtenim que l'estimació de màxima versemblança dels coeficients de la mixtura és

$$\pi_k = \frac{N_k}{N}. \tag{3.31}$$

  Per tant, π_k representa la mesura en que la corresponent component k de la mixtura explica els punts del conjunt de dades.

Les expressions (3.23), (3.26), (3.31) que hem trobat, no ens donen una solució tancada pels paràmetres de la mixtura, ja que les probabilitats $\gamma(z_{nk})$ depenen dels mateixos paràmetres (recordem l'expressió (3.10)). Ens trobem en la següent situació: si

coneguéssim els paràmetres, podríem calcular les probabilitats posteriors $\gamma(z_{nk})$; si coneguéssim els posteriors $\gamma(z_{nk})$, podríem calcular fàcilment els paràmetres. Això dona pas a un esquema iteratiu relativament simple per trobar estimadors de màxima versemblança, que resulta ser un cas particular de l'algorisme EM. Ho veiem en el següent apartat.

3.3 Deducció de l'algorisme EM

El nostre objectiu és, donat un model de mixtura de gaussianes, maximitzar el logaritme de la funció de versemblança respecte als paràmetres. Dels resultats que hem obtingut a l'apartat anterior, es dedueix un mètode iteratiu que ens permet aproximar-ne una solució.

En primer lloc, hem de triar uns valors inicials de les mitjanes, covariàncies i coeficients de la mixtura. Després, hem d'alternar entre dos passos d'actualització, els quals anomenem pas E (*estimació*) i pas M (*maximització*). En el pas E utilitzem els valors actuals dels paràmetres per avaluar les probabilitats posteriors. Després, en el pas M, fem servir aquestes probabilitats per reestimar les mitjanes, les covariàncies i els coeficients de la mixtura, segons les expressions (3.23), (3.26) i (3.31).

Aquest mètode iteratiu s'anomena algorisme EM. En resumim els passos a continuació:

1. Inicialitzar les mitjanes μ_k, les covariaàncies Σ_k i els coeficients π_k i avaluar el valor inicial del logaritme de la versemblança.

2. **Pas E** (*estimació*): avaluar les probabilitats posteriors $\gamma(z_{nk})$ utilitzant els valors actuals dels paràmetres μ_k, Σ_k i π_k.

$$\gamma(z_{nk}) = \frac{\pi_k \cdot \mathcal{N}(\, x_n \mid \mu_k, \Sigma_k \,)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\, x_n \mid \mu_j, \Sigma_j \,)}. \tag{3.32}$$

3. **Pas M** (*maximització*): reestimar els paràmtres μ_k, Σ_k i π_k emprant les probabilitats $\gamma(z_{nk})$ calculades al pas E.

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, x_n \tag{3.33}$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \tag{3.34}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{3.35}$$

on

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{3.36}$$

4. Avaluar el logaritme de la versemblança

$$\log p(\, \mathcal{X} \mid \pi, \mu, \Sigma\,) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\, x_n \mid \pi_k, \mu_k, \Sigma_k\,) \right\} \tag{3.37}$$

i comprovar si hi ha hagut convergència dels paràmetres o de la mateixa funció. En cas afirmatiu, aturar l'algorisme. El resultat de l'estimació de màxima versemblança correspon al valors obtinguts pels paràmetres en l'última iteració. En cas contrari, tornar al pas 2.

4 L'algorisme EM (variables latents)

En aquesta secció del treball parlarem en detall de l'algorisme EM i destacarem el paper determinant que juguen les variables latents. Primer, discutirem el cas general i després ens centrarem en el cas particular del model de mixtura de distribucions gaussianes.

4.1 Definició

L'objectiu de l'algorisme EM és trobar solucions del problema de màxima versemblança en models de mixtura de distribucions o, més generalment, en models amb variables latents. Suposem que tenim un conjunt d'observacions $\mathcal{X} = \{x_1, \ldots, x_N\}$ i considerem un model amb densitat de probabilitat p_Θ i variables latents $\mathcal{Z} = \{z_1, \ldots, z_N\}$. Introduïm en aquest apartat una nova nomenclatura, que usarem al llarg de les properes seccions:

- Anomenem $\{\mathcal{X}, \mathcal{Z}\}$ el **conjunt de dades complet**.
- Ens referim a les dades realment observades $\mathcal{X}$ com a **dades incompletes**.

Podem escriure el logaritme de la versemblança de la funció de desnitat de probabilitat marginal de $\mathcal{X}$ en termes de la funció de densitat de probabilitat conjunta de $\{\mathcal{X}, \mathcal{Z}\}$ com

$$\log p(\, \mathcal{X} \mid \Theta \,) = \log \left\{ \sum_z p(\, \mathcal{X}, \mathcal{Z} \mid \Theta \,) \right\}. \tag{4.1}$$

Observem que a l'expressió anterior el logaritme no actua directament sobre la funció de densitat de probabilitat conjunta, a causa del sumatori que apareix sobre la variable latent. Això provoca que en molts casos l'expressió (4.1) no es pugui maximitzar directament. Per fer front a aquest inconvenient, a partir d'ara treballarem amb la funció de densitat de probabilitat conjunta de $\{\mathcal{X}, \mathcal{Z}\}$, en comptes de fer-ho amb la marginal de $\mathcal{X}$. Ens podem trobar en dues situacions, que detallem a continuació.

Variables latents conegudes

Suposem que, per a cada observació x_n, coneixem el valor de la seva corresponent variable latent z_n. En aquest cas, considerem el problema de maximitzar el logaritme de la versemblança del conjunt de dades complet

$$0 = \frac{\partial \log p(\, \mathcal{X}, \mathcal{Z} \mid \Theta \,)}{\Theta}. \tag{4.2}$$

És molt probable que aquest problema sigui més simple de resoldre que no pas la que trobàvem a (4.1).

Valor esperat de les variables latents

Aquesta és la situació en la que ens trobarem a la pràctica: no sabem quins són els valors de les variables latents i només coneixem la seva distribució posterior $p(\, \mathcal{Z} \mid \mathcal{X}, \Theta \,)$. Llavors, considerem el valor esperat del logaritme de la versemblança del conjunt de dades complet sota la distribució posterior de $\mathcal{Z}$

$$\mathbb{E}_{\mathcal{Z}}\left[\log\, p(\,\mathcal{X},\mathcal{Z}\mid\Theta\,)\right]=\sum_{z} p(\,\mathcal{Z}\mid\mathcal{X},\Theta\,)\cdot\log\, p(\,\mathcal{X},\mathcal{Z}\mid\Theta\,) \tag{4.3}$$

i maximitzem aquest valor aplicant l'algorisme EM. Primer, hem de triar un valor inicial pels paràmetres Θ_0 i establim $\Theta^{old} \leftarrow \Theta_0$. Després hem d'iterar en dos passos:

- **Pas E** (esperança): utilitzem el valor actual dels paràmetres Θ^{old} per trobar la distribució posterior de les variables latents $p(\,\mathcal{Z}\mid\mathcal{X},\Theta^{old}\,)$. Després, calculem l'esperança del logaritme de la versemblança de $\{\mathcal{X},\mathcal{Z}\}$, avaluat a un valor general de Θ. Denotem per $Q(\Theta,\Theta^{old})$ aquest valor esperat i el calculem com

$$Q(\Theta,\Theta^{old})=\sum_{z} p(\,\mathcal{Z}\mid\mathcal{X},\Theta^{old}\,)\cdot\log\, p(\,\mathcal{X},\mathcal{Z}\mid\Theta\,). \tag{4.4}$$

- **Pas M**: (maximització): maximitzem la funció Q per trobar els nous valors dels paràmetres, que denotem per Θ^{new}

$$\Theta^{new}=\arg\max_{\Theta} Q(\Theta,\Theta^{old}). \tag{4.5}$$

Observem com en la definició de $Q(\Theta,\Theta^{old})$ que hem donat a (4.4), el logaritme sí que actua directament sobre la distribució conjunta. Aquest fet és el que ens simplifica els càlculs respecte al problema de maximitzar l'expressió (4.1).

Com a conclusió d'aquesta secció, donem un resum esquemàtic dels passos que cal seguir per aplicar l'algorisme EM.

1. Inicialitzar els valors dels paràmertes Θ_0 i establir $\Theta^{old} \leftarrow \Theta_0$.

2. **Pas E** (esperança): avaluar $p(\,\mathcal{Z}\mid\mathcal{X},\Theta^{old}\,)$ i calcular $Q(\Theta,\Theta^{old})$.

3. **Pas M** (maximització): maximitzar la funció $Q(\Theta,\Theta^{old})$ per trobar els nous paràmetres, que denotem per Θ^{new}.

4. Comprovar si hi ha hagut convergència dels paràmetres o de la mateixa funció. En cas afirmatiu, aturar l'algorisme. El resultat de l'estimació de màxima versemblança és el valor dels paràmetres Θ^{new}. En cas contrari, establir

$$\Theta^{old} \leftarrow \Theta^{new} \tag{4.6}$$

   i tornar al pas 2.

4.2 Aplicació en el model de mixtura de gaussianes

Vegem ara com podem aplicar aquesta definició de l'algorisme EM en el cas particular del model de mixtura de gaussianes. El nostre objectiu és maximitzar el logaritme de la funció de versemblança que hem calculat a (3.20).

Variables latents conegudes

Suposem que coneixem els valors de les variables latents $\mathcal{Z}$. Llavors, considerem el problema de maximitzar la versemblança del conjunt complet de dades. De les expressions (3.6) i (3.7) tenim que la versemblança de $\{\mathcal{X}, \mathcal{Z}\}$ és de la forma

$$p(\,\mathcal{X}, \mathcal{Z} \mid \mu, \Sigma, \pi\,) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \mathcal{N}(\,x_n \mid \mu_k, \Sigma_k\,)^{z_{nk}}. \tag{4.7}$$

Prenent el logaritme obtenim

$$\log p(\,\mathcal{X}, \mathcal{Z} \mid \mu, \Sigma, \pi\,) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{\log \pi_k + \log\, \mathcal{N}(\,x_n \mid \mu_k, \Sigma_k\,)\right\}. \tag{4.8}$$

Si comparem aquesta expressió amb la del logaritme de la versemblança de les dades incompletes (3.20), veiem com el sumatori sobre k i el logaritme han intercanviat posicions. Ara, el logaritme actua directament sobre la distribució gaussiana que, alhora, pertany a la família de les funcions exponencials. En conseqüència, estimar els paràmetres màxim versemblants és molt més simple. Vegem-ho calculant de nou les derivades parcials respecte cadascun dels paràmetres i igualant-les a 0:

- **Respecte** μ_k

$$0 = \frac{\partial \log p(\,\mathcal{X}, \mathcal{Z} \mid \mu, \Sigma, \pi\,)}{\partial \mu_k} = \sum_{n=1}^{N} z_{nk}\, \frac{N(\,x_n \mid \mu_k, \Sigma_k\,)}{N(\,x_n \mid \mu_k, \Sigma_k\,)}\, \Sigma_k^{-1}(x_n - \mu_k). \tag{4.9}$$

  Multiplicant per Σ_k i reordenant els termes, tenim

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} \cdot x_n \tag{4.10}$$

  on definim

$$N_k = \sum_{n=1}^{N} z_{nk}. \tag{4.11}$$

  Observem que, a diferència de la definició (3.24), ara N_k representa el nombre real de punts assignats a la corresponent component k de la mixtura. A més, el paràmetre

μ_k màxim versemblant que hem obtingut és anàleg a quan tenim una única distribució normal, però ara només implica als punts assignats a la component k.

- **Respecte** Σ_k

$$0 = \frac{\partial \log\, p(\,\mathcal{X} \mid \mu, \Sigma\,))}{\partial \Sigma_k} = \sum_{n=1}^{N} z_{nk} \left\{ (x_n - \mu)(x_n - \mu)^T - \Sigma_k \right\} \tag{4.12}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} (x_n - \mu)(x_n - \mu)^T \tag{4.13}$$

També el paràmetre Σ_k màxim versemblant és anàleg a quan tenim una única distribució normal, però en aquest cas només recull els punts assignats a la component k de la mixtura.

- **Respecte** π_k.

Els paràmetres π_k han de satisfer la condició (3.3). Per introduir aquesta restricció, usem un multiplicador de Lagrange λ

$$\log\, p(\,\mathcal{X} \mid \mu, \Sigma\,) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{4.14}$$

i maximitzem aquesta quantitat

$$0 = \frac{1}{\pi_k} \sum_{n=1}^{N} z_{nk} + \lambda = \frac{N_k}{\pi_k} + \lambda \tag{4.15}$$

Multipliquem per π_k

$$0 = N_k + \lambda \cdot \pi_k. \tag{4.16}$$

i sumem respecte k

$$0 = \sum_{k=1}^{K} N_k - \lambda \sum_{k=1}^{K} \pi_k = N + \lambda. \tag{4.17}$$

És a dir que $\lambda = -N$. Substituïm aquest resultat a (4.16) i obtenim que

$$\pi_k = \frac{N_k}{N}. \tag{4.18}$$

Per tant, el coeficient π_k representa la fracció de punts del conjunt de dades assignats a la corresponent component de la mixtura.

D'aquesta manera, hem obtingut una solució tancada pels paràmetres que maximitza el logaritme de la versemblança del conjunt complet de dades.

Valor esperat de les variables latents

A la pràctica, no coneixem els valors de les variables latents. Per tant, considerem el valor esperat del logaritme de la versemblança de $\{\mathcal{X}, \mathcal{Z}\}$ sota a la distribució posterior de $\mathcal{Z}$. Usant les expressions (3.6), (3.7) i (3.8) i aplicant el teorema de Bayes de les probabilitats condicionades, tenim que

$$p(\, \mathcal{Z} \mid \mathcal{X}, \mu, \Sigma, \pi\,) = \prod_{n=1}^{N} \frac{\prod_{k=1}^{K} [\, \pi_k \cdot \mathcal{N}(\, x_n \mid \mu_k, \Sigma_k\,)\,]^{z_{nk}}}{\sum_{k=1}^{K} [\, \pi_k \cdot \mathcal{N}(\, x_n \mid \mu_k, \Sigma_k\,)\,]^{z_{nk}}}. \tag{4.19}$$

Observem que aquesta expressió factoritza sobre n. Això vol dir que les variables latents $\{z_n\}$ són independents sota la seva distribució posterior. Llavors, partint de l'expressió (3.10), tenim que el valor esperat de z_{nk} ve donat per

$$\mathbb{E}\,[z_{nk}] = \frac{\pi_k \cdot \mathcal{N}(\, x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\, x_n \mid \mu_j, \Sigma_j)} = \gamma(z_{nk}) \tag{4.20}$$

que és, precisament, la probabilitat posterior $\gamma(z_{nk})$. Finalment, l'esperança del logaritme de la versemblança del conjunt complet de dades sota la distribució posterior de $\mathcal{Z}$ és

$$\mathbb{E}_{\mathcal{Z}}\,[\log\, p(\, \mathcal{X}, \mathcal{Z} \mid \mu, \Sigma, \pi\,)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})\, \{\log\, \pi_k + \log\, \mathcal{N}(\, x_n \mid \pi_k, \mu_k, \Sigma_k\,)\} \tag{4.21}$$

Llavors, seguim el següent esquema iteratiu:

1. Inicialitzar els valors dels paràmetres μ_k^{old}, Σ_k^{old} i π_k^{old}

2. **Pas E**: usar aquests valors per avaluar les probabilitats posteriors donades per (4.20)

3. **Pas M**: fixades les probabilitats $\gamma(z_{nk})$, maximitzar l'expressió (4.21) respecte a cadascun dels paràmetres. Com a resultat obtenim solucions μ^{new}, Σ^{new} i π^{new}.

4. Comprovar si se satisfan les condicions que hàgim definit de convergència. En cas afirmatiu, aturar l'algorisme. En cas contrari, establir

$$\mu_k^{old} \leftarrow \mu_k^{new} \tag{4.22}$$

$$\Sigma_k^{old} \leftarrow \Sigma_k^{new} \tag{4.23}$$

$$\pi_k^{old} \leftarrow \pi_k^{new} \tag{4.24}$$

i tornar al pas 2.

4.3 Relació amb l'algorisme *K-mitjanes*

En aquest apartat mostrarem la relació que hi ha entre els dos algorismes que hem estudiat: l'algorisme *K-mitjanes* i l'algorisme EM. Veurem que l'algorisme *K-mitjanes* es dedueix com un límit particular de EM aplicat al model de mixtura de gaussianes.

Considerem un model de mixtura de K gaussianes en el qual les matrius de covariàncies són de la forma

$$\Sigma_k = \epsilon \mathbf{I} \quad (\forall k) \tag{4.25}$$

on ϵ és el paràmetre de la variància (constant fixada) i $\mathbf{I}$ és la matriu identitat.

Proposició: L'algorisme *K-mitjanes* és equivalent a aplicar l'algorisme EM en el model de mixtura de gaussianes definit a (4.25) , quan $\epsilon \to 0$.

Demostració:

Considerem el model de mixtura de gaussianes de la proposició. Per hipòtesi, la funció de densitat de cada component de la mixtura és de la forma

$$\mathcal{N}(\, x \mid \mu_k, \epsilon \mathbf{I}\,) = \frac{1}{(2\pi)^{1/2}} \exp\left\{ -\frac{1}{2\epsilon} \|x - \mu_k\|^2 \right\}. \tag{4.26}$$

Apliquem l'algorisme EM al model. De l'expressió (3.10) tenim que les probabilitats posteriors $\gamma(z_{nk})$ per a un punt x_n venen donades per

$$\gamma(z_{nk}) = \frac{\pi_k \cdot \exp\{-\|x_n - \mu_k\|^2/2\epsilon\}}{\sum_{j=1}^{K} \pi_j \cdot \exp\{-\|x_n - \mu_j\|^2/2\epsilon\}}. \tag{4.27}$$

Prenem ara el límit quan $\epsilon \to 0$ i analitzem com les equacions pròpies de l'algorisme EM es trasnformen en les equacions donades per l'algorisme *K-mitjanes*.

- **Assignacions**

  Suposem que k' és tal que $\|x_n - \mu_{k'}\| < \|x_n - \mu_k\|$ $(\forall k)$. Llavors se satisfà que

$$\gamma(z_{nk}) \to \begin{cases} 1, & \text{si } k = k' \\ 0, & \text{si } k \neq k'. \end{cases} \tag{4.28}$$

  Ho demostrem a continuació:

$$\lim_{\epsilon\to 0}\gamma(z_{nk}) = \lim_{\epsilon\to 0}\frac{\pi_k \exp\left\{-\|x_n-\mu_k\|^2/2\epsilon\right\}}{\sum_{j=1}^K \pi_j \exp\left\{-\|x_n-\mu_j\|^2/2\epsilon\right\}}$$

$$= \lim_{\epsilon\to 0}\frac{1}{1+\sum_{j\neq k}\pi_j/\pi_k \exp\left\{-\left[\,\|x_n-\mu_j\|^2-\|x_n-\mu_k\|^2\,\right]/2\epsilon\right\}}.$$

Tenint en compte que

$$\pi_j/\pi_k \exp\left\{-\left[\|x_n-\mu_j\|^2-\|x_n-\mu_k\|^2\right]/2\epsilon\right\} \;\to\; \begin{cases} 0, & \text{si } \|x_n-\mu_k\| < \|x_n-\mu_j\| \\ \infty, & \text{en cas contrari.} \end{cases}$$

obtenim el resultat (4.28). Aquest raonament només és vàlid si els coeficients de la mixtura satisfan que $\pi_k \neq 0$ $(\forall k)$. □

D'aquesta manera, obtenim una assignació determinista dels punts a les components de la mixtura, com succeeix en l'algorisme *K-mitjanes*. En concret, hem demostrat que

$$\gamma(z_{nk}) \;\to\; r_{nk} \tag{4.29}$$

on r_{nk} està definida per (2.5). Cada punt s'assigna a la component de la mixtura amb la mitjana més propera.

- **Mitjanes**

  L'estimació dels paràmetres μ_k de EM donada per (3.23) es redueix al resultat (2.7) de *K-mitjanes*

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^N \gamma(z_{nk})\, x_n = \frac{1}{|S_k|}\sum_{x_n\in S_k} x_n. \tag{4.30}$$

  on hem usat que

$$N_k = \sum_{k=1}^K \gamma(z_{nk}) = \sum_{k=1}^K r_{nk} = |S_k|. \tag{4.31}$$

- **Coeficients de la mixtura**

  A través de l'expressió (3.31) podem reestimar els paràmetres π_k perquè siguin la fracció de punts assignats a la component k

$$\pi_k = \frac{N_k}{N} = \frac{|S_k|}{N}. \tag{4.32}$$

Aquests paràmetres, però, ja no juguen un paper determinant en l'algorisme.

- **Valor esperat**

  Per últim, el límit del valor esperat del logaritme de la versemblança del conjunt complet de dades donat per (4.21) es transforma en

$$\mathbb{E}_{\mathcal{Z}}\left[\log\, p(\,\mathcal{X}, \mathcal{Z} \mid \mu, \Sigma, \pi\,)\right] \;\to\; -\,\frac{1}{2}\underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\,\|x-\mu_k\|^2}_{J} + \text{const.} \tag{4.33}$$

  Amb aquesta última expressió veiem que, sota les hipòtesis de la proposiciío, maximitzar el valor esperat del logaritme de la versemblança del conjunt complet de dades quan apliquem EM és equivalent a minimitzar la mesura de distorsió J (2.4) de l'algorisme *K-mitjanes*.

□

5 L'algorisme EM (en general)

En aquesta última secció del treball parlarem de l'algorisme EM de forma més general i provarem que, efectivament, aquest algorisme maximitza la funció de versemblança a cada iteració. Per últim, a partir de la definició, en veurem diverses variants que podem emprar sota certes condicions.

5.1 Definició i convergència

L'algorisme EM és un mètode general per trobar una estimació dels paràmetres de màxima versemblança en models probabilístics amb variables latents. Considerem, doncs, un model en el qual denotem per $\mathcal{X}$ el conjunt de les variables observades, i per $\mathcal{Z}$ el conjunt de les variables latents. Suposem que la distribució conjunta depèn d'conjunt de paràmetres Θ. El nostre objectiu és maximitzar la funció

$$p(\,\mathcal{X} \mid \Theta\,) = \sum_{z} p(\,\mathcal{X}, \mathcal{Z} \mid \Theta\,). \tag{5.1}$$

En aquesta expressió estem suposant que les variables $\mathcal{Z}$ són discretes. La discussió és idèntica en el cas de tenir variables latents contínues o una combinació dels dos tipus, canviat el sumatori per una integral on sigui necessari.

Suposem que optimitzar l'expressió (5.1) és difícil però que, en canvi, optimitzar la funció de versemblança del conjunt complet de dades $p(\,\mathcal{X}, \mathcal{Z} \mid \Theta\,)$ és un problema significativament més senzill de resoldre. A continuació, introduïm una distribució $q(\mathcal{Z})$ definida sobre les variables latents.

Proposició: per a qualsevol elecció que fem de $q(\mathcal{Z})$, se satisfà la següent descomposiciío

$$\log p(\,\mathcal{X} \mid \Theta\,) = \mathcal{L}(q, \Theta) + \mathrm{KL}(q \,\|p\,) \tag{5.2}$$

on definim

$$\mathcal{L}(q, \Theta) = \sum_{z} q(\mathcal{Z}) \log \left\{ \frac{p(\,\mathcal{X}, \mathcal{Z} \mid \Theta\,)}{q(\mathcal{Z})} \right\} \tag{5.3}$$

i

$$\mathrm{KL}(q \,\|\, p) = -\sum_{z} q(\mathcal{Z}) \log \left\{ \frac{p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,)}{q(\mathcal{Z})} \right\}. \tag{5.4}$$

La mesura KL que definida t a (5.4) s'anomena **divergència de Kullback-Leibler**. En aquest cas, ens dona una mesura de la distància entre $q(\mathcal{Z})$ i la corresponent distribució posterior $p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,)$.

Demostració de la proposició:

$$\mathcal{L}(q,\Theta) + \mathrm{KL}(q \,\|\, p) = \sum_z q(\mathcal{Z}) \log \left\{ \frac{p(\,\mathcal{X}, \mathcal{Z} \mid \Theta\,)}{q(\mathcal{Z})} \right\} - \sum_z q(\mathcal{Z}) \log \left\{ \frac{p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,)}{q(\mathcal{Z})} \right\}$$

$$= \sum_z q(\mathcal{Z}) \left[\log \left\{ \frac{p(\,\mathcal{X}, \mathcal{Z} \mid \Theta\,)}{q(\mathcal{Z})} \right\} - \log \left\{ \frac{p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,)}{q(\mathcal{Z})} \right\} \right]$$

$$= \sum_z q(\mathcal{Z}) \log p(\,\mathcal{X} \mid \Theta\,) = \log p(\,\mathcal{X} \mid \Theta\,) \sum_z q(\mathcal{Z}).$$

Usem que, per la definició de probabilitat, se satisfà que

$$\sum_z q(\mathcal{Z}) = 1 \tag{5.5}$$

i obtenim el resultat que volíem provar. □

Propietats de la divergència de Kullback-Leibler

1. $\mathrm{KL}(q\|p) \geq 0$

2. $\mathrm{KL}(q\|p) = 0 \;\Leftrightarrow\; q(\mathcal{Z}) = p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,)$

3. $\mathcal{L}(q,\Theta) \leq \log\, p(\,\mathcal{X} \mid \Theta\,)$. És a dir, que $\mathcal{L}(q,\Theta)$ és una cota inferior de $\log p(\,\mathcal{X} \mid \Theta\,)$.

A partir de la descomposició (5.2) podem donar una definició general de l'algorisme EM. Suposem que el valor actual dels paràmetres és Θ^{old} i, a continuació, iterem en dues fases:

- **Pas E** (esperança): fixat el valor de Θ^{old}, maximitzem $\mathcal{L}(q, \Theta^{old})$ respecte $q(\mathcal{Z})$. Tenint en compte que s'ha de satisfer la igualtat (5.2) i que el valor $\log p(\,\mathcal{X} \mid \Theta^{old}\,)$ no depèn de $q(\mathcal{Z})$, deduïm que el valor màxim de $\mathcal{L}(q,\Theta)$ s'assoleix quan $\mathrm{KL}(q \,\|\, p) = 0$. Per tant, el resultat d'aquest pas és

$$q(\mathcal{Z}) = p(\,\mathcal{Z} \mid \mathcal{X}, \Theta^{old}\,) \tag{5.6}$$

- **Pas M** (maximització): fixada $q(\mathcal{Z})$, maximitzem $\mathcal{L}(q, \Theta^{old})$ respecte a Θ per trobar un nou valor Θ^{new}. En aquest pas, augmenta el valor de la cota inferior $\mathcal{L}(q, \Theta^{old})$ (si no era ja un màxim) i, en conseqüència, augmenta el valor del logaritme de la versemblança.

  Donat que $q(\mathcal{Z})$ és fixada, el seu valor s'avalua a Θ^{old} i, per tant, no serà igual al valor de la nova probabilitat posterior $p(\,\mathcal{Z} \mid \mathcal{X}, \Theta^{new}\,)$. Llavors tindrem un valor estrictament positiu de la divergència KL. Això provoca que l'augment del logaritme de la versemblança sigui major que l'augment que experimenta la cota inferior. Ho podem veure esquemàticament en la següent figura:

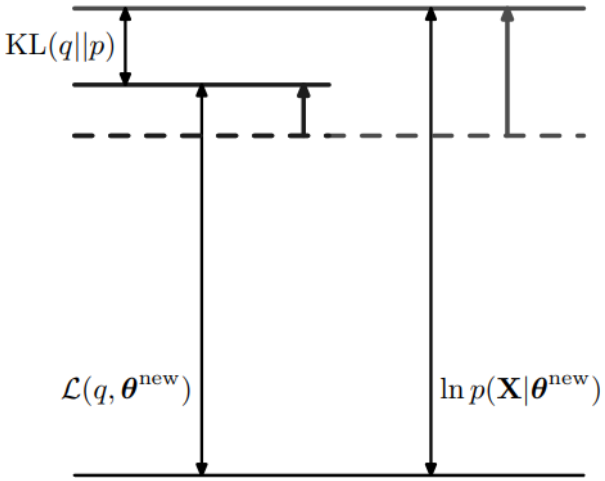

Figura 5: Esquema pas M

Per tant, veiem que tant el pas E com el pas M de l'algorisme fan augmentar el valor d'una cota inferior (ben definida) del logaritme de la versemblança. A cada cicle complet de EM es recalculen els paràmetres del model de manera que la versemblança augmenta.

5.2 Variants de l'algorisme

En aquest apartat veurem diverses variants de l'algorisme EM. En primer lloc, l'utilitzarem per maximitzar la distribució posterior $p(\Theta \mid \mathcal{X})$. Després, parlerem de dues generalitzacions que existeixen de l'algorisme i que es poden usar quan el pas M és difícil de resoldre: l'algorisme EM generalitzat (GEM) i l'algorisme EM condicional (ECM). Tanmateix, no entrarem en detall de com aplicar-les. Finalment, veurem una variant incremental que podem fer servir en el cas de tenir variables i.i.d.

Algorisme EM per maximitzar la distribució posterior dels paràmetres

Suposem que volem maximitzar la distribució posterior $p(\Theta \mid \mathcal{X})$ en un model en el qual hem introduït una probabilitat a *priori* dels paràmetres $p(\Theta)$. En aquest escenari podem escriure

$$p(\Theta \mid \mathcal{X}) = \frac{p(\Theta, \mathcal{X})}{p(\mathcal{X})} = \frac{p(\mathcal{X} \mid \Theta) \cdot p(\Theta)}{p(\mathcal{X})} \tag{5.7}$$

Per tant, prenent el logaritme, tenim que

$$\log p(\Theta \mid \mathcal{X}) = \log p(\mathcal{X} \mid \Theta) + \log p(\Theta) - \log p(\mathcal{X}). \tag{5.8}$$

Utilitzant la descomposició (5.2) veiem que

$$\begin{aligned} \log p(\Theta \mid \mathcal{X}) &= \mathcal{L}(q, \Theta) + \mathrm{KL}(q \| p) + \log p(\Theta) - \log p(\mathcal{X}) \\ &\geq \mathcal{L}(q, \Theta) + \log p(\Theta) - \underbrace{\log p(\mathcal{X})}_{\text{const.}}. \end{aligned} \tag{5.9}$$

Per tant, maximitzem la part dreta d'aquesta desigualtat, que és una cota inferior de $\log p(\,\Theta \mid \mathcal{X})$. Ho fem en els dos passos de l'algorisme:

- **Pas E** (esperança): maximitzem la cota inferior respecte a q. Donat que q només apareix al terme $\mathcal{L}(q, \Theta)$, les equacions d'aquest pas són idèntiques a les que hem trobat a l'apartat anterior.
- **Pas M** (maximització): maximitzem la cota inferior respecte a Θ. Observem que ara apareix un nou terme $\log p(\Theta)$ que cal tenir el compte en a l'hora de calcular la derivada parcial.

Generalitzacions de EM

En models complexes pot donar-se el cas que, o bé el pas E o bé el pas M (o tots dos) siguin intractables. Parlarem ara de dos mètodes que podem emprar en cas que el problema estigui en la resolució del pas M.

- **Algorisme EM generalitzat** (GEM): en aquesta variant, en comptes d'intentar maximitzar $\mathcal{L}(q, \Theta)$ respecte a Θ, busquem un nou paràmetre de Θ que n'incrementi el valor. Una manera de fer-ho és emprar estratègies d'optimització no lineals, com pot ser l'algorisme del gradient conjugat. (Sanjay-Gopal and Hebert, 1988)

- **Algorisme EM condicional** (ECM): en aquest cas, realitzem una partició dels paràmetres en grups. D'aquesta manera, subdividim el pas M en diversos passos, cadascun dels quals implica maximitzar $\mathcal{L}(q, \Theta)$ respecte a un dels grups de paràmetres, mantenint la resta fixats. (Meng and Rubin, 1993).

Variant incremental

Suposem que el nostre conjunt complet de dades $\{\mathcal{X}, \mathcal{Z}\}$ està format per variables i.i.d. Llavors, per hipòtesi tenim la següent igualtat

$$p(\mathcal{X}, \mathcal{Z}) = \prod_{n=1}^{N} p(x_n, z_n) \tag{5.10}$$

Per marginalització, obtenim que

$$p(\mathcal{X}) = \sum_{z} p(\mathcal{X}, \mathcal{Z}) = \sum_{z} \prod_{n=1}^{N} p(x_n, z_n) = \prod_{n=1}^{N} \sum_{z} p(x_n, z_n) = \prod_{n=1}^{N} p(x_n). \tag{5.11}$$

Com que hem assumit que les variables $\{\mathcal{X}, \mathcal{Z}\}$ són i.i.d, les $\mathcal{X}$ també ho són entre elles. Finalment, calculem la distribució posterior de $\mathcal{Z}$ usant la fórmula de Bayes

$$p(\,\mathcal{Z} \mid \mathcal{X}, \Theta\,) = \prod_{n=1}^{N} \frac{p(x_n, z_n)}{p(x_n)} = \prod_{n=1}^{N} p(\,z_n \mid x_n, \Theta\,). \tag{5.12}$$

Observem que aquesta expressió també factoritza sobre n.

Sota aquestes condicions podem aplicar una versió incremental de l'algorisme EM, en la qual en cada cicle només tractem un dels punts de dades. En el pas E, en comptes de recalcular totes les probabilitats posterios $\gamma(z_{nk})$, només reavaluem les del punt triat. Després, en el corresponent pas M, podem necessitar els valros de $\gamma(z_{nk})$ de la resta de punts per fer alguns càlculs. Tanmateix, si les components de la mixtura amb la qual estem treballant pertanyen a la família de les funcions exponencials, això no succeeix. Ho veiem a continuació pel cas de la mixtura de gaussianes.

Prenem un punt qualsevol x_m i suposem que en el pas E hem recalculat només les probabilitats posteriors $\gamma(z_{nk})$ d'aquest punt, les quals denotem per $\gamma^{old}(z_{mk})$ i $\gamma^{new}(z_{mk})$. A partir de la definició (3.24), obtenim

$$N_k^{new} = N_k^{old} + \gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk}). \tag{5.13}$$

Amb això i usant (3.23) podem calcular les mitjanes com

$$\mu_k^{new} = \mu_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \right) (x_m - \mu_k^{old}). \tag{5.14}$$

Demostrem com s'obté l'expressió anterior:

$$\begin{aligned}
\mu_k^{new} &= \frac{1}{N_k^{new}} \sum_{n=1}^{N} \gamma(z_{nk})\, x_n \\
&= \frac{1}{N_k^{new}} \left[\sum_{n \neq m} \gamma(z_{nk})\, x_n + \gamma^{new}(z_{mk})\, x_m + \gamma^{old}(z_{mk})\, x_m - \gamma^{old}(z_{mk})\, x_m \right] \\
&= \frac{1}{N_k^{new}} \left[\mu_k^{old} N_k^{old} + (\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk}))\, x_m \right] \\
&= \mu_k^{old}\, \frac{N_k^{new} - (\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk}))}{N_k^{new}} + \frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}}\, x_m \\
&= \mu_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \right) (x_m - \mu_k^{old}).
\end{aligned}$$

Per a calcular les covariàncies, a partir de l'expressió (3.26) obtenim la següent fórmula incremental

$$\Sigma_k^{new} = \Sigma_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \right) \left[(x_m - \mu_k)(x_m - \mu_k)^T - \Sigma_k^{old} \right]. \tag{5.15}$$

El procediment per obtenir aquest resultat és anàleg al que hem emprat per a les mitjanes

$$\Sigma_k^{new} = \frac{1}{N_k^{new}} \sum_{n=1}^{N} \gamma(z_{nk}) \left(x_n - \mu_k\right) \left(x_n - \mu_k\right)^T$$

$$= \frac{1}{N_k^{new}} \left[\sum_{n \neq m} \gamma(z_{nk}) \left(x_n - \mu_k\right) \left(x_n - \mu_k\right)^T + \gamma^{new}(z_{mk}) \left(x_m - \mu_k\right) \left(x_m - \mu_k\right)^T \right.$$

$$\left. + \gamma^{old}(z_{mk}) \left(x_m - \mu_k\right) \left(x_m - \mu_k\right)^T - \gamma^{old}(z_{mk}) \left(x_m - \mu_k\right) \left(x_m - \mu_k\right)^T \right]$$

$$= \frac{1}{N_k^{new}} \left[\Sigma_k^{old} N_k^{old} + \left(\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})\right) \left(x_m - \mu_k\right) \left(x_m - \mu_k\right)^T \right]$$

$$= \Sigma_k^{old} \frac{N_k^{new} - \left(\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})\right)}{N_k^{new}} + \frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \left(x_m - \mu_k\right) \left(x_m - \mu_k\right)^T$$

$$= \Sigma_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}} \right) \left[(x_m - \mu_k)(x_m - \mu_k)^T - \Sigma_k^{old} \right].$$

Per últim, l'equació resultant per a calcular els coeficients de la mixtura, partint de (3.31), és

$$\pi_k^{new} = \frac{N_k^{new}}{N} = \frac{N_k^{old} + \gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N} = \pi_k^{old} + \frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N}. \quad (5.16)$$

6 Implementacions en R

Com a adjunt, s'entrega un document anomenat "codeTFG" en format .html que conté exemples d'aplicació dels dos algorismes treballats en aquesta memòria:

- **Algorisme K-mitjanes**

  S'ha aplicat la funció *kmeans* continguda al paquet 'stats' de R sobre una base de dades reals (continguda també al paquet). Posteriorment, s'ha implementat una funció pròpia que executa l'algorismse i s'ha usat sobre les mateixes dades. S'han comparat tots dos resultats.

  Per últim s'ha usat l'algorisme per comprimir una imatge en format .jpg, usant K=10 clústers (colors).

- **Algorisme EM**

  S'ha fet ús de les funcions incloses al paquet 'rebmix' de R. S'ha simulat un conjunt d'observacions d'un model de mixtura de gaussianes per després usar-les per aplicar les funcions d'estimació del paquet. S'estudien diferents tipus d'estratègies d'estimació i s'analitzen els resultats obtinguts en cada cas.

Totes les explicacions sobre les funcions emprades i els procediments que s'han seguit es detallen al document adjunt.

7 Conclusions

En aquest últim apartat de la memòria farem un breu resum dels aspectes més importants del treball i analitzarem els resultats obtinguts en cada apartat.

El primer objectiu d'aquest treball era definir i entendre què són i per què serveixen els models de mixtura de distribucions per a després, aprofundir en l'estudi d'un cas particular: el model de mixtura de distribucions gaussianes. Des del primer apartat d'introducció hem començat a assolir aquest propòsit: hem entès què és una mixtura, quina utilitat té i n'hem donat la definició. Hem descobert també un nou concepte, el de variable latent, com a variable pertanyent a un model, però que no pot ser observada directament. A partir d'aquí, hem vist la interpretació dels models de mixtura en termes de variables latents discretes. La resta de seccions han girat entorn de les mixtures.

L'altre tema rellevant que es pretenia tractar en aquest projecte és l'algorisme EM. Hem vist que aquest algorisme constitueix un dels mètodes més importants per inferir els paràmetres màxim versemblants d'un model amb variables latents. Donat que hem interpretat els models de mixtura en termes de variables latents, aquest algorisme ens ha servit per inferir-ne els paràmetres. Finalment, a l'última secció del treball, hem donat una definició més general d'aquest algorisme i hem provat que, en efecte, maximitza la versemblança a cada iteració.

A més, hem descobert que els models de mixtura també es poden usar per classificar dades. Per això, hem estudiat en detall un algorisme de classificació molt emprat en l'àmbit de l'anàlisi de clústers: l'algorisme *K-mitjanes*. No sols n'hem donat la definició, hem vist com s'aplica i hem estudiat les seves propietats, sinó que l'hem relacionat directament amb l'algorisme EM, com a un límit particular.

Un últim aspecte important és que hem reproduït tots dos algorismes estudiats i els hem aplicat en dades reals i simulades per entendre com funcionen més en profunditat i poder visualitzar els resultats que s'obtenen.

En general, podem dir que els objectius que teníem marcats a l'inici d'aquest projecte s'han assolit. Tanmateix, tant els models de mixtura, com l'algorisme EM (i també l'algorisme K-mitjanes en l'anàlisi de clústers) són àmbits d'estudi molt amplis i hi ha gran varietat d'aplicacions que es poden investigar.

Com a línies d'investigació es proposa aprofundir en l'estudi de les variants de l'algorisme EM: l'algorisme EM generalitzat i l'algorisme ECM. En aquest treball les hem mencionat i referenciat, però no les hem tractat en detall. Pot resultar interessant i es podria considerar estudiar-les en un altre TFG.

Referències

[1] Kotz S. et al. (2006). *EM Algorithm, Encyclopedia of Statistial Sciences*, 2nd edition, pp.1918-1926.

[2] Bishop C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[3] Bilmes J.A. (1998). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, International Computer Science Institute (Berkeley), Computer Science Division, Department of Electrical Engineering and Computer Science (TR-97-021).

[4] Hastie T.H., Tibshirani R. and Friedman M. (2009). *The Elements of Statistical Learning*, 2nd edition

[5] Meng X. and Rubin D.B. (1993). *Maximum likelihood estimation via the ECM algorithm: A general framework*, Oxford University Press on behalf of Biometrika Trust.

[6] Sanjay-Gopal S. and Hebert T.J. (1988). *Bayesian Pixel Classification Using Spatially VariantFinite Mixtures and the Generalized EM Algorithm*, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 7, NO. 7.

[7] Neal, R. M. and Hinton, G.E (1999). *A new view of the EM algorithm that justifies incremental and other variants*. In M. I. Jordan (Ed.), Learning in Graphical Models, pp. 355-368. MIT Press.

[8] `https://es.wikipedia.org/wiki/Silhouette_(clustering)`

[9] `https://cran.r-project.org/web/packages/rebmix/rebmix.pdf`

[10] `https://stephens999.github.io/fiveMinuteStats/index.html`

[11] `https://github.com/stephens999/fiveMinuteStats`

[12] `https://github.com/fonluiz`