# Modelos de Soporte No Supervisado

## Model-Based Clustering

# Model-Based Clustering

- Los métodos de agrupamiento tradicionales, como la agrupación jerárquica y k-means clustering, son heurísticos y no se basan en modelos formales.

- Además, el algoritmo k-means suele inicializarse aleatoriamente, por lo que diferentes ejecuciones de k-means a menudo arrojará resultados diferentes.

- Además, k-means requiere que el usuario especifique la cantidad óptima de cluster.

- Una alternativa es la agrupación basada en modelos, que considera que los datos provienen de una distribución que es una mezcla de dos o más grupos (Chris Fraley y Adrian E. Raftery, 2002 y 2012).

- A diferencia de k-means, el clustering basado en modelos usa un criterio de asignación suave, donde cada punto de datos tiene una probabilidad de pertenecer a cada grupo.

# Model-Based Clustering

- En la agrupación basada en modelos, los datos se consideran como procedentes de una mezcla de densidad.

- Cada componente (cluster) k está modelado por la distribución normal o Gaussiana que se caracteriza por los parámetros:

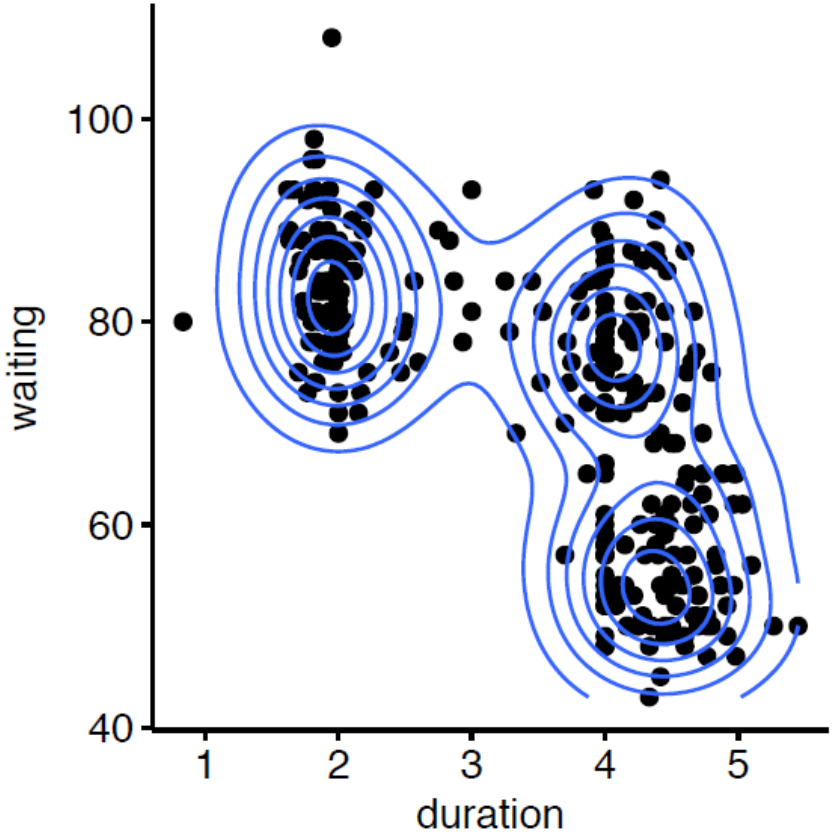
  μk: vector medias,
  Σk: matriz de varianzas-covarianzas,
  P(i): Una probabilidad asociada en la mezcla

Cada punto tiene una probabilidad de perteneciente a cada grupo.

# Model-Based Clustering

- Considere los datos “old faithful geyser data” [in MASS R package], y utilizando el paquete “ggpubr”:

```
library("MASS")
data("geyser")
library("ggpubr")
ggscatter(geyser, x = "duration", y = "waiting")+
geom_density2d() # Add 2D density
```



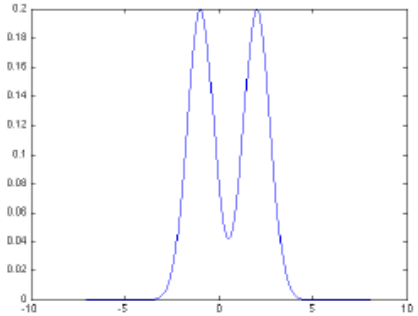
# Model-Based Clustering

- El gráfico anterior sugiere al menos 3 grupos en la mezcla.
- La forma de cada uno de los 3 grupos parece ser aproximadamente elíptica, lo que sugiere tres distribuciones normales bivariadas.
- Como las 3 elipses parecen ser similares en términos de volumen, forma y orientación, podríamos anticipar que los tres componentes de esta mezcla podrían tener matrices de covarianza homogéneas.

# Model-Based Clustering

## Mixture of Gaussians

- Generally: $X \sim \text{Multinomial}(\theta)$, $Z|X=k \sim \mathcal{N}(\mu_k, \Sigma_k)$
- Example: $P(X=1)=\frac{1}{2}, P(X=2)=\frac{1}{2}$, $Z|X=1 \sim \mathcal{N}(-1,1)$, $Z|X=2 \sim \mathcal{N}(2,1)$, $\rightarrow Z \sim \frac{1}{2}\mathcal{N}(-1,1)+\frac{1}{2}\mathcal{N}(2,1)$



- ML Objective: given data $z^{(1)}, \ldots, z^{(m)}$

$$\max_{\theta,\mu,\Sigma} \sum_{i=1}^{m} \log \sum_{k=1}^{n} \theta_k \frac{1}{(2\pi)^{d/2}|\Sigma_k|} e^{-\frac{1}{2}(z-\mu_k)^\top \Sigma_k^{-1}(z-\mu_k)}$$

- Setting derivatives w.r.t. $\theta$, $\mu$, $\Sigma$ equal to zero does not enable to solve for their ML estimates in closed form

We can evaluate function → we can in principle perform local optimization. In this lecture: "EM" algorithm, which is typically used to efficiently optimize the objective (locally)

## Expectation Maximization (EM)

- Example:
  - Model: $P(X=1)=\frac{1}{2}, P(X=2)=\frac{1}{2}$, $Z|X=1 \sim \mathcal{N}(\mu_1,1)$, $Z|X=2 \sim \mathcal{N}(\mu_2,1)$
  - Goal:
    - Given data $z^{(1)}, \ldots, z^{(m)}$ (but no $x^{(i)}$ observed)
    - Find maximum likelihood estimates of $\mu_1$, $\mu_2$
- EM basic idea: if $x^{(i)}$ were known → two easy-to-solve separate ML problems
- EM iterates over
  - **E-step**: For i=1,…,m fill in missing data $x^{(i)}$ according to what is most likely given the current model $\mu$
  - **M-step**: run ML for completed data, which gives new model $\mu$

# Model-Based Clustering

**Estimación de los parámetros del modelo**

- Los parámetros del modelo se pueden estimar utilizando la Expectativa-Maximización (EM)
- El algoritmo inicia por el agrupamiento jerárquico basado en modelos.
- Cada grupo k está centrado en la media μk, con densidad aumentada para puntos cercanos a la media.
- Las características geométricas (forma, volumen, orientación) de cada grupo están determinadas por la  matriz de varianza-covarianza Σk.
- Existen diferentes parametrizaciones posibles de Σk disponibles en el paquete R mclust (ver? mclustModelNames).

# Model-Based Clustering

- Las opciones de parametrización disponibles, en el paquete mclust, están representadas por identificadores que incluyen:

EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV y VVV

- El primer identificador se refiere al volumen, el segundo a la forma y el tercero a la orientación.
- E significa "igual", V para "variable" y I para "ejes de coordenadas".

Por ejemplo:

- EVI denota un modelo en el cual los volúmenes de todos los conglomerados son iguales (E), las formas de los conglomerados pueden variar (V), y la orientación es la identidad (I) o ejes de coordenadas.
- EEE significa que los clusters tienen el mismo volumen, forma y orientación en espacio p-dimensional.
- VEI significa que los clústeres tienen un volumen variable, la misma forma y orientación igual a ejes coordenados.

# Model-Based Clustering

**Elegir el mejor modelo**

- El paquete Mclust usa la máxima verosimilitud e adaptarse a todos estos modelos, con diferentes parametrizaciones de matriz de covarianza, para un rango de k componentes.

- El mejor modelo se selecciona usando el Criterio de información bayesiano o BIC.

- Alta puntuación BIC indica evidencia sólida para sustentar el modelo correspondiente.

# Model-Based Clustering

**Bayesian information criterion** (**BIC**) or **Schwarz criterion** (also **SBC**, **SBIC**) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model.

- $x$ = the observed data;
- $n$ = the number of data points in $x$, the number of observations, or equivalently, the sample size;
- $k$ = the number of free parameters to be estimated. If the estimated model is a linear regression, $k$ is the number of regressors, including the intercept;
- $p(x|k)$ = the probability of the observed data given the number of parameters; or, the likelihood of the parameters given the dataset;
- $L$ = the maximized value of the likelihood function for the estimated model.

The formula for the BIC is:[3][4]

$$-2 \cdot \ln p(x|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \ln(n).$$

# Model-Based Clustering

**Ejemplo**

- Comenzamos por instalar el paquete mclust: install.packages ("mclust")

- Aquí, ilustramos la agrupación basada en modelos en el conjunto de datos de diabetes consistente en tres mediciones y el diagnóstico para 145 sujetos descritos de la siguiente manera:

```
library("mclust")
data("diabetes")
head(diabetes, 3)
```

# Model-Based Clustering

| ## class | glucose | insulin | sspg |
|---|---|---|---|
| ## 1 Normal | 80 | 356 | 124 |
| ## 2 Normal | 97 | 289 | 117 |
| ## 3 Normal | 105 | 319 | 143 |

clase: el diagnóstico: normal, químicamente diabético y abiertamente diabético. Excluido del análisis de cluster.

glucosa: respuesta de glucosa en plasma a glucosa oral

insulina: respuesta de insulina en plasma a glucosa oral

sspg: glucosa en plasma en estado estable (mide la resistencia a la insulina)

# Model-Based Clustering

**Ejemplo:**

```
library(mclust)
df <- scale(diabetes[, -1])
mc <- Mclust(df) # Model-based-clustering
summary(mc)
```

*Para estos datos, se puede observar que la agrupación basada en modelos seleccionó un modelo con tres clusters.*

*El nombre óptimo del modelo seleccionado es el modelo VVV. Con los tres componentes elipsoidales con volumen, forma y orientación variables.*

*El resumen contiene también la tabla de conglomerados que especifica el número de observaciones en cada cluster.*

# Model-Based Clustering

**Visualizar clustering basado en modelos**

- Usaremos la función fviz_mclust () [en el paquete factoextra]
- Cuando los datos contienen más de dos variables, fviz_mclust () usa un análisis de componentes principales para reducir la dimensionalidad de los datos.
- Las primeras dos componentes principales se utilizan para producir un diagrama de dispersión de los datos.
- Sin embargo, si desea trazar los datos usando solo dos variables de interés, por ejemplo, c ("insulina", "sspg"), puede especificarlo en la función fviz_mclust () utilizando el argumento choose.vars = c ("insulina", "sspg").

# Model-Based Clustering

**library**(factoextra)

*# Criterio BIC para seleccionar el número de clusters*

**fviz_mclust**(mc, "BIC", palette = "jco")

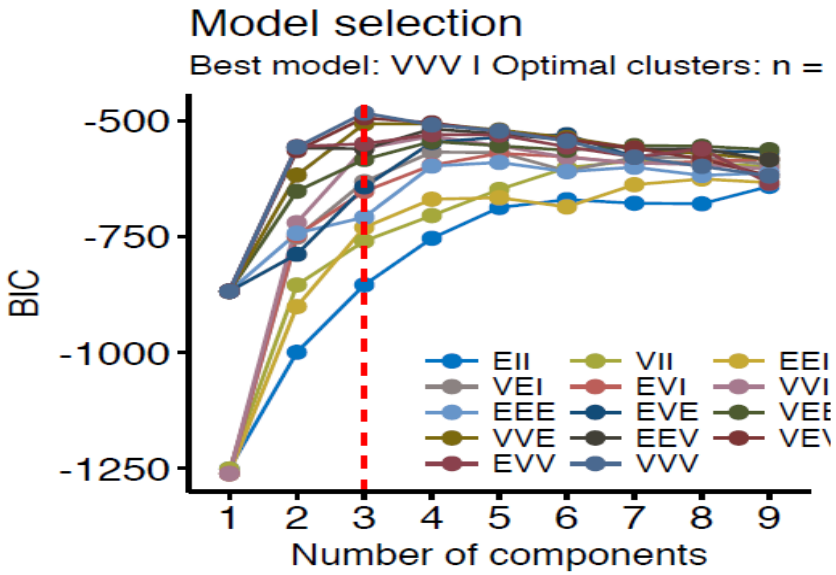
*# Gráfica de los clusters*

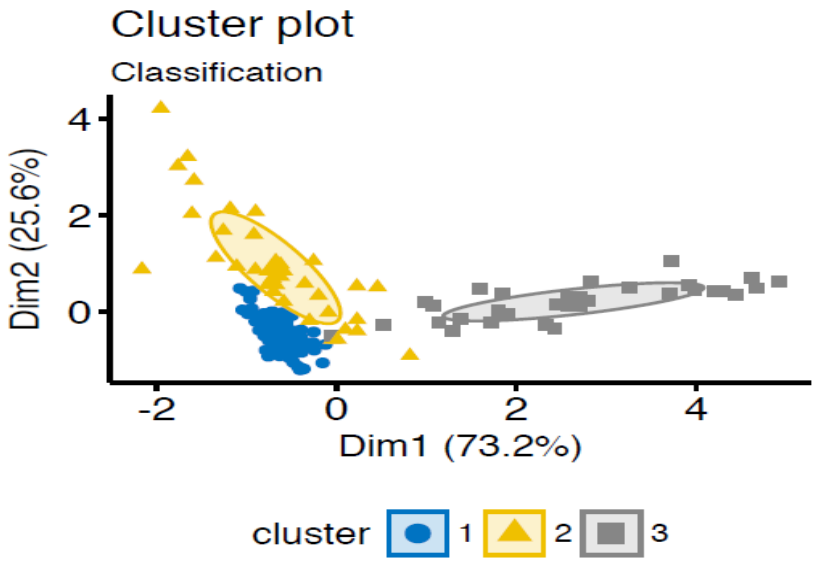
**fviz_mclust**(mc, "classification", geom = "point", pointsize = 1.5, palette = "jco")

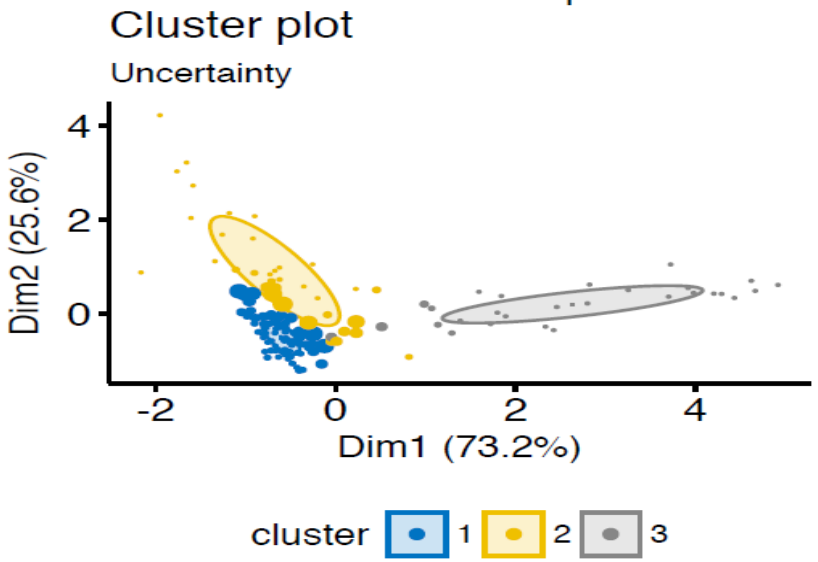
*# Datos con incertidumbre en la clasificación*

- **fviz_mclust**(mc, "uncertainty", palette = "jco")

# Model-Based Clustering




Tenga en cuenta que, en el gráfico de incertidumbre, los símbolos más grandes indican que es más incierta la clasificación.