FINITE MIXTURE OF SKEWED DISTRIBUTIONS

Víctor Hugo Lachos Dávila

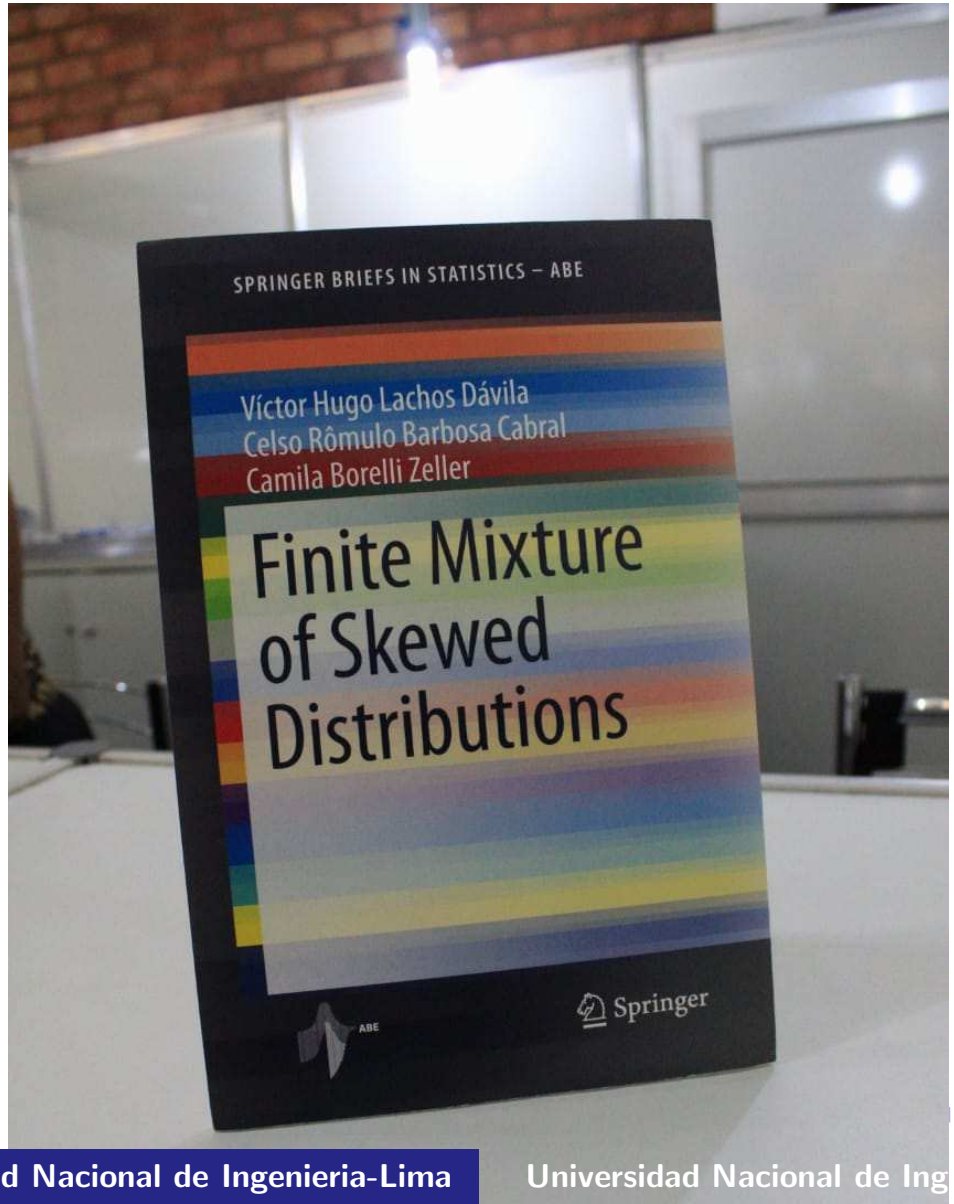
Department of Statistics
University of Connecticut, U.S.A.

Joint work with Celso R. Cabral and Camila B. Zeller

Second International Conference in **Stochastic Processes and Random Phenomena and Their Applications 2018**: In Tribute to the 65th birthday of Professor Dipak K. Dey

October 03-06, Lima-Peru

Recent book (Springer)



Motivation

- Modeling based on finite mixture distributions is a rapidly developing area with an exploding range of applications in areas such as biology, biometrics, genetics, medicine and marketing.
- Statistical models which are based on finite mixture distributions capture many specific properties of real data such as multimodality, skewness, kurtosis, and unobserved heterogeneity
- The importance of mixtures can be noted from the large number of books (McLachlan and Peel (2000), Fruhwirth-Schnatter (2006) and Mengersen et al. (2011)), including the FOURTH special edition in CSDA (2019).

An example: Human height data

- height is typically modeled as a normal distribution for each gender with a mean of approximately 5'10" for males and 5'5" for females. In this case gender is the source of "the latent heterogeneity"
- Given only the height data and not the gender assignments for each data point, the distribution of all heights would follow the sum of two scaled (different variance) and shifted (different mean) normal distributions.
- A model making this assumption is an example of a Finite Mixture (FM) of two Gaussian models.
- In general a FM model may have more than two components.

Fishery data

- 256 observations of fish lengths;
- Specialists agree that the source of the latent heterogeneity can be the age groups, which is a variable very hard to observe directly.
- Length at age is a useful metric for evaluating growth.
- Data freely available through the `R` package `bayesmix`;

An Example – Fishery data

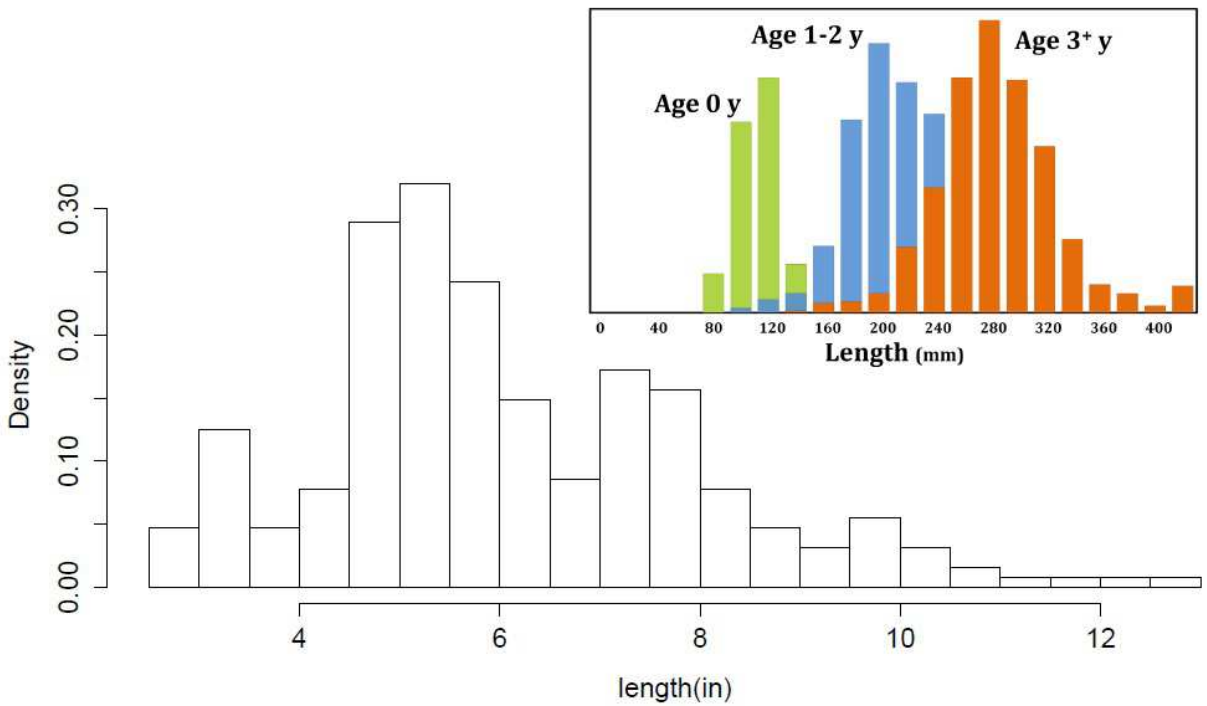


Figure: Histogram of the Fishery data

Finite Mixtures

Definiçao

Given densities g_j and weights $p_j \geq 0$, $j = 1, \ldots, G$, such that $\sum_{j=1}^{G} p_j = 1$, a finite mixture of densities is the density

$$f(y) = \sum_{j=1}^{G} p_j g_j(y).$$

- The density g_j is named the *jth component of the mixture*.

Finite Mixtures

- Finite mixtures are useful to model population heterogeneity, when we know that observations belong to G distinct subpopulations, but we do not know how to discriminate between them;
- *Finite mixture model representation*: suppose a non observable random variable W such that

$$P(W = j) = p_j, \quad \text{with} \quad Y|W = j \sim g_j.$$

  Then the distribution of Y is a finite mixture with *jth* component g_j and weight p_j, $j = 1, \ldots, G$.
- In general, $g_j(\cdot) = g(\cdot|\boldsymbol{\theta}_j)$, $j = 1, \ldots, G$. That is, the components g_j belong to the same parametric family.
- The research on estimation of the parameters in finite mixture models was more focused in the normal components case

Great Flexibility!

- Modeling with finite mixtures makes possible to capture many features, such as multimodality, skewness, and kurtosis.

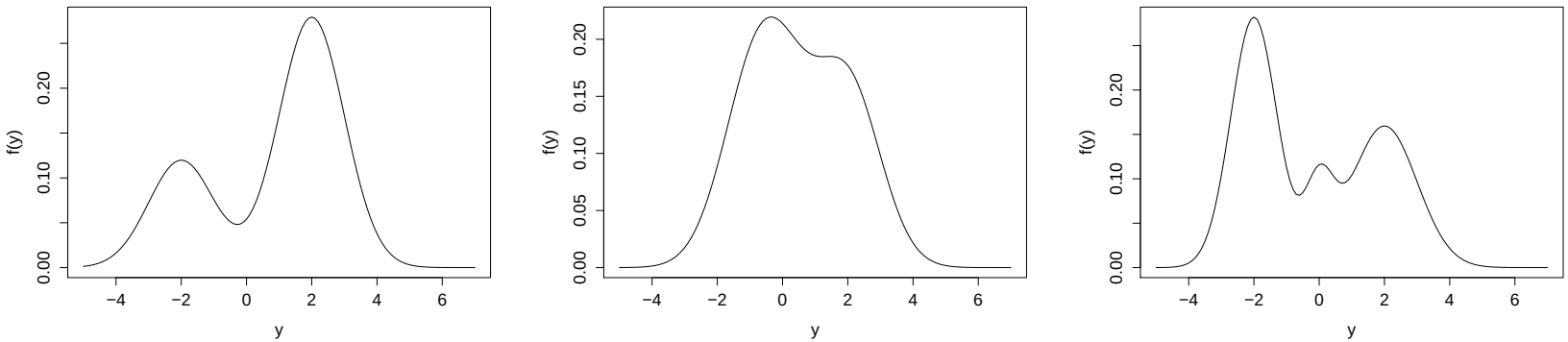


Figure: Densities of mixtures of univariate normal distributions. Left: $\mu_1 = -2$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 1$, $p_1 = 0.3$. Middle: $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $p_1 = p_2 = 0.3$. Right: $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.2$, $\sigma_3^2 = 1$, $p_1 = 0.5$, $p_2 = 0.1$.

Mathematical motivation for using mixtures

- The following theorem shows that any continuous density can be approximated by a proper finite mixture of any (not necessarily normal) continuous densities. See DasGupta (2008, Theorem 33.2) for more details.

Theorem

Let $f(\cdot)$ and $g(\cdot)$ be continuous densities on $\mathbb{R}^q$. Given $\epsilon > 0$ and a compact set $C \subset \mathbb{R}^q$, there exists a finite mixture of the form $h(x) = \sum_{j=1}^{G} p_j \sigma_j^{-q} g\left((x - \mu_j)/\sigma_j\right)$ such that $\sup_{x \in C} |f(x) - h(x)| < \epsilon$.

Maximum Likelihood Estimation

- We consider a random sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from the mixture model

$$f(\mathbf{y}|\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j g(\mathbf{y}|\boldsymbol{\theta}_j), \quad \mathbf{y} \in \mathbb{R}^q$$

  where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top, p_1, \ldots, p_G)^\top$;
- The likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{G} p_j g(\mathbf{y}_i|\boldsymbol{\theta}_j),$$

  where $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$;
- Direct maximization of $L(\cdot|\mathbf{y})$ can lead to a difficult and unstable numerical problem;
- Instead, the best option is to use the EM algorithm.

Data Augmentation

For each $i = 1, \ldots, n$, let us define

$$P(W_i = j) = p_j, \quad \text{with} \quad \mathbf{Y}_i | W_i = j \sim g(\cdot|\boldsymbol{\theta}_j)$$

and the random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})$, such that

$$Z_{ij} = 1 \ \text{ if and only if } \ W_i = j.$$

- If $Z_{ij} = 1$ then $Z_{ik} = 0$ for $k \neq j$;
- The distribution of the vector $\mathbf{Z}_i$ is multinomial with one trial and probabilities $p_1, \ldots, p_G$;
- We use the notation $\mathbf{Z}_i \sim \text{Multinomial}(1; p_1, \ldots, p_G)$.

Data Augmentation

Let $\mathbf{Z} = (\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_n^\top)^\top$. The *complete-data likelihood* is the likelihood function obtained as if $(\mathbf{y}^\top, \mathbf{Z}^\top)^\top$ were observable, that is,

$$\ell_c(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{Z}) = \prod_{i=1}^{n} \pi(\mathbf{y}_i|\mathbf{Z}_i)\pi(\mathbf{Z}_i) = \prod_{i=1}^{n}\prod_{j=1}^{G} g(\mathbf{y}_i|\boldsymbol{\theta}_j)^{z_{ij}} p_j^{z_{ij}}.$$

- The E-step consists in taking the conditional expectation

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}\left[\log \ell_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}\right],$$

  where the expectation is being effected using $\widehat{\boldsymbol{\theta}}^{(k)}$ for $\boldsymbol{\theta}$.

The EM Algorithm

Now, observe that

$$\mathrm{E}\left[\log \ell_c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{G} Z_{ij}\left[\log g(\mathbf{y}_i|\boldsymbol{\theta}_j) + \log p_j\right]|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}\right\}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{G} \mathrm{E}(Z_{ij}|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)})\left[\log g(\mathbf{y}_i|\boldsymbol{\theta}_j) + \log p_j\right].$$

- It is easy to compute $\mathrm{E}(Z_{ij}|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)})$, because the distribution of Z_{ij} is Bernoulli with probability of success p_j:

$$\widehat{z}_{ij}^{(k+1)} = \mathrm{E}(Z_{ij}|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}) = P(Z_{ij}=1|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}) \propto g(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_j^{(k)})\widehat{p}_j^{(k)}$$
$$\Rightarrow z_{ij}^{(k+1)} = \frac{g(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_j^{(k)})\widehat{p}_j^{(k)}}{\sum_{m=1}^{G} g(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}_j^{(m)})\widehat{p}_j^{(m)}}, \quad j=1,\ldots,G.$$

The EM Algorithm

The next step is to maximize the Q-function. Regarding the parameters p_j, $j = 1, \ldots, G$, this is equivalent to maximize the function

$$\sum_{i=1}^{n} \sum_{j=1}^{G} \widehat{z}_{ij}^{(k+1)} \log p_j$$

with respect to p_j, $j = 1, \ldots, G$. Then, it is straightforward to prove that

$$\widehat{p}_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{z}_{ij}^{(k+1)}.$$

The EM Algorithm

- Let us assume that we have a mixture density with *jth* component $\mathrm{N}_q(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with weight p_j, $j = 1, \ldots, G$;
- The M-step is equivalent to maximize the function

$$\sum_{i=1}^{n} \sum_{j=1}^{G} \widehat{z}_{ij}^{(k+1)} \log \phi_q(\mathbf{y}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

with respect to $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, where $\phi_q(\mathbf{y}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is the $q-$variate normal density with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$;

- The solution is

$$\widehat{\boldsymbol{\mu}}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ij}^{(k+1)} \mathbf{y}_i}{\sum_{i=1}^{n} \widehat{z}_{ij}^{(k+1)}}, \quad \text{and}$$

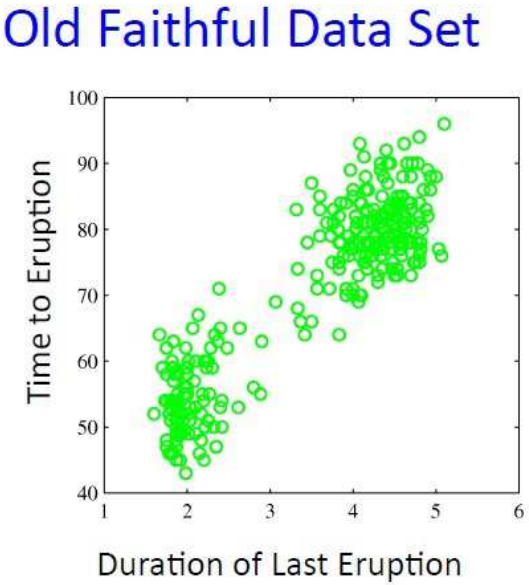
$$\widehat{\boldsymbol{\Sigma}}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ij}^{(k+1)} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i^{(k+1)})(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i^{(k+1)})^\top}{\sum_{i=1}^{n} \widehat{z}_{ij}^{(k)}}, \quad j = 1, \ldots, G.$$

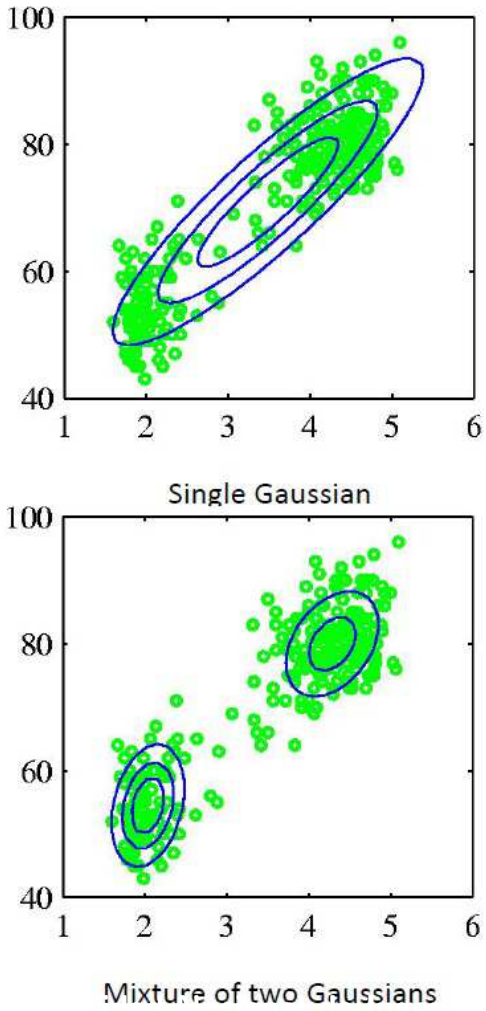
- stopping criterion

$$||\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)}|| < 10^{-6}$$

Inference using FM models

1.-Density estimation





Inference using FM models

2.-Clustering.

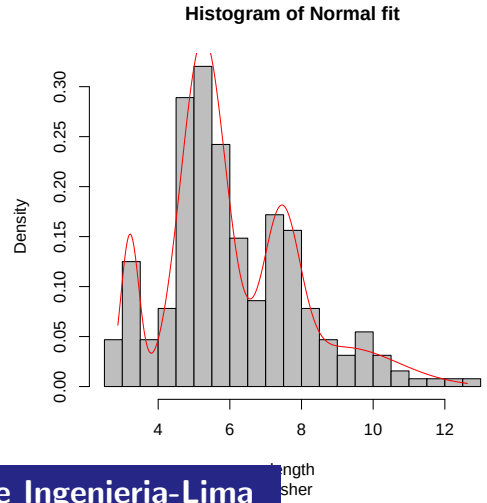
Given a univariate normal model's parameters, the probability that a data point belongs to component C_i is calculated using Bayes' theorem:

$$\begin{aligned} P(C_i|x) &= \frac{P(C_i,x)}{p(x)} = \frac{P(C_i)P(x|C_i)}{\sum_{i=1}^{K} P(C_i)P(x|C_i)} \\ &= = \frac{p_i\phi(x|\mu_j,\sigma_j^2)}{\sum_{i=1}^{K} p_i\phi(x|\mu_j,\sigma_j^2)} \end{aligned}$$

Package '*mixsmsn*'

Application: Fishery data

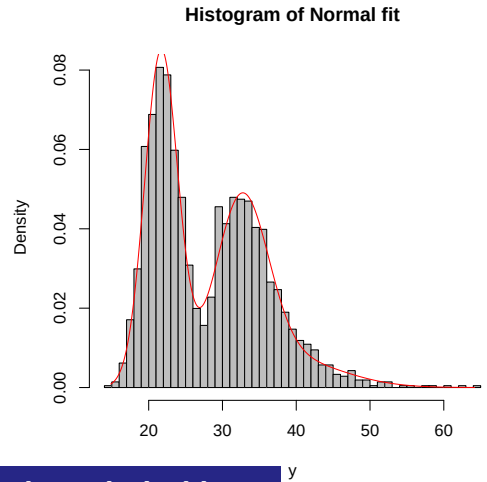
- library(mixsmsn)
- data(fish)
- hist(fish, breaks = 40, main = "Histogram of fishery data", xlab = "length")
- fish.analysis = smsn.search(fish, nu = 3, g.min = 1, g.max = 6, family = "Normal", criteria = "aic")
- fish.fit = smsn.mix(fish, nu = 3, g = 3, get.init = TRUE, criteria = TRUE, group = TRUE, family = "Normal", calc.im = FALSE)
- mix.hist(fish, fish.fit). Best model FOUR components.



Package '*mixsmsn*'

Application: BMI data

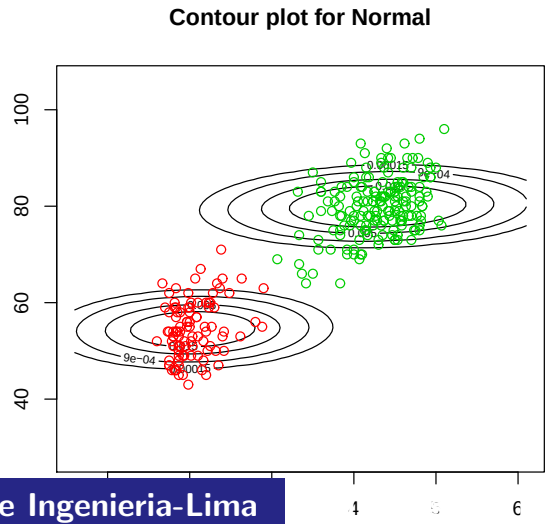
- library(mixsmsn)
- data(bmi)
- hist(bmi, breaks = 40, main = "Histogram of BMI", xlab = "bmi")
- bmi.analysis = smsn.search(bmi, nu = 3, g.min = 1, g.max = 5, family = "Normal", criteria = "aic")
- bmi.fit = smsn.mix(bmi, nu = 3, g = 3, get.init = TRUE, criteria = TRUE, group = TRUE, family = "Normal", calc.im = FALSE)
- mix.hist(bmi, bmi.fit) Best model: THREE components.



Package '*mixsmsn*'

Application: Old Faithful geyser

- library(mixsmsn)
- data("faithful")
- faith.analysis = smsn.search(faithful, nu = 3, g.min = 1, g.max = 5, family = "Normal")
- faith.fit ¡- smsn.mmix(faithful, nu = 3,g = 2, get.init = TRUE, group = TRUE, family = "Normal", calc.im = FALSE)
- mix.contour(faithful, faith.fit, x.min = 1, x.max = 1, y.min = 15, y.max = 10, levels = c(0.1, 0.015, 0.005, 0.0009, 0.00015)) Best model: TWO components.



Drawbacks of the Gaussian assumption

- In the framework of FM models, the subpopulations are routinely assumed to be normal for mathematical convenience.
- However, deviations from the normal assumption among the subpopulations such as, strong asymmetry or heavy tails, are not uncommon.
- Sometimes these departures are not well captured by a finite mixture of normal distributions, or even by a finite mixture of a more robust symmetric distribution, like the Student-t.

Table: BMI data. BIC criterion for several mixture models. The number in parenthesis denotes the number of components.

Model	BIC	AIC
Normal (2)	13861.78	13833.51
Normal (3)	13787.05	13741.83
Student-t (2)	13814.37	13786.10
Student-t (3)	13787.51	13742.29
Skew-normal (2)	13803.36	13763.79
Skew-t (2)	13777.13	13737.56

References

- Peel, D. and McLachlan, 2000 . Robust mixture modeling using the t distribution distributions. Statistics & Computing 10, 339–348.
- Lin, T-I., Lee, J.C. and Hsieh, 2007. Robust mixture modeling using the skew-t distribution distributions. Statistics & Computing 17, 81–92.
- Basso, R. M., Lachos, V. H., Cabral, C. R. B., Ghosh, P., 2010. Robust mixture modeling based on scale mixtures of skew-normal distributions. Computational Statistics & Data Analysis 54 (12), 2926–2941.
- Cabral, C. R. B., Lachos, V. H., Prates, M. O., 2012. Multivariate mixture modeling using skew-normal independent distributions. Computational Statistics & Data Analysis 56 (1), 126–142.
- Galimberti, G., Soffritti, G., 2014. A multivariate linear regression analysis using finite mixtures of t distributions. Computational Statistics & Data Analysis 71, 138–150.
- Prates, M. O., Cabral, C. R. B. and Lachos, V. H, 2012. mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. Journal of Statistical Software 54 (12).
- Soffritti, G., Galimberti, G., 2011. Multivariate linear regression with non-normal errors: a solution based on mixture models. Statistics and Computing, 21, 523–536.
- Zeller, C. B., Cabral, C. R., Lachos, V. H., 2016. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, 25, 375-396.

Thank you!