Finite Mixtures of SMSN 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

Univariate Mixtures of SMSN Distributions

The *finite mixture of SMSN distributions model (FM-SMSN)* is defined by a random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ from

$$f(y|\boldsymbol{\theta}) = \sum_{j=1}^{G} p_j g(y|\boldsymbol{\theta}_j), \; p_j \geq 0, \; \sum_{j=1}^{G} p_j = 1, \quad j = 1, \ldots, G,$$

where

- $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j^\top)^\top$ is the specific vector of parameters for the component j;
- $g(\cdot|\boldsymbol{\theta}_j)$ is the SMSN($\boldsymbol{\theta}_j$) density;
- $p_1, \ldots, p_G$ are the mixing probabilities and
- $\boldsymbol{\theta} = ((p_1, \ldots, p_G)^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$ is the vector with all parameters;
- For computational convenience we assume that $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$.

Hierarchical Representation of the FM-SMSN Model

$$
\begin{aligned}
Y_i|U_i = u_i, T_i = t_i, Z_{ij} = 1 &\sim N(\mu_j + \Delta_j t_i, u_i^{-1}\Gamma_j), \\
T_i|U_i = u_i, Z_{ij} = 1 &\sim HN(0, u_i^{-1}), \\
U_i|Z_{ij} = 1 &\sim H(u_i; \boldsymbol{\nu}) \text{ and} \\
\mathbf{Z}_i &\sim \text{Multinomial}(1; p_1 \ldots, p_G) \qquad i = 1, \ldots, n, \;\; j = 1, \ldots, G,
\end{aligned}
$$

where

$$
\Gamma_j = (1 - \delta_j^2)\sigma_j^2, \quad \Delta_j = \sigma_j \delta_j, \quad \text{and} \;\; \delta_j = \frac{\lambda_j}{\sqrt{1 + \lambda_j^2}}.
$$

Maximum likelihood estimation via EM algorithm

The complete-data log-likelihood function is

$$\ell_c(\boldsymbol{\theta}) = c + \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\left(\log p_j - \frac{1}{2}\log|\Gamma_j| - \frac{u_i}{2\Gamma_j}(y_i - \mu_j - \Delta_j t_i)^2 + \log(h(u_i;\boldsymbol{\nu}))\right),$$

where c is a constant that is independent of the parameter vector $\boldsymbol{\theta}$ and $h(\cdot;\boldsymbol{\nu})$ is the density of U_i. Let us define

$$\begin{aligned} \widehat{z}_{ij} &= \mathrm{E}[Z_{ij}|\widehat{\boldsymbol{\theta}}, y_i], \quad \widehat{zu}_{ij} = \mathrm{E}[Z_{ij}U_i|\widehat{\boldsymbol{\theta}}, y_i], \quad \widehat{zut}_{ij} = \mathrm{E}[Z_{ij}U_i T_i|\widehat{\boldsymbol{\theta}}, y_i], \\ \widehat{zut^2}_{ij} &= \mathrm{E}[Z_{ij}U_i T_i^2|\widehat{\boldsymbol{\theta}}, y_i]. \end{aligned}$$

Maximum likelihood estimation via EM algorithm

and using known properties of conditional expectation, we obtain

$$\begin{aligned}
\widehat{z}_{ij} &= \frac{\widehat{p}_j g(y_i|\widehat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^{G} \widehat{p}_j g(y_i|\widehat{\boldsymbol{\theta}}_j)}, \\
\widehat{zu}_{ij} &= \widehat{z}_{ij}\widehat{u}_{ij}, \quad \widehat{zut}_{ij} = \widehat{z}_{ij}(\widehat{u}_{ij}\widehat{m}_{ij} + \widehat{M}_j\widehat{\eta}_{ij}), \\
\widehat{zut^2}_{ij} &= \widehat{z}_{ij}(\widehat{u}_{ij}\widehat{m}_{ij}^2 + \widehat{M}_j^2 + \widehat{M}_j\widehat{m}_{ij}\widehat{\eta}_{ij}),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\eta}_{ij} &= E\left[U_i^{1/2} W_{\Phi_1}\left(\frac{U_i^{1/2}\widehat{m}_{ij}}{\widehat{M}_j}\right) \mid \widehat{\boldsymbol{\theta}}, y_i, Z_{ij} = 1\right], \\
\widehat{M}_j^2 &= \frac{\widehat{\Gamma}_j}{\widehat{\Gamma}_j + \widehat{\Delta}_j^2}, \ \widehat{m}_{ij} = \frac{\widehat{\Delta}_j}{\widehat{\Gamma}_j + \widehat{\Delta}_j^2}(y_i - \widehat{\mu}_j)
\end{aligned}$$

and

$$\widehat{u}_{ij} = E[U_i|\widehat{\boldsymbol{\theta}}, y_i, Z_{ij} = 1], \qquad i = 1, \ldots, n, \quad j = 1, \ldots, G.$$

For the computation of $\widehat{\eta}_{ij}$ and $\widehat{u}_{ij}$, see Section 2.3.1.

Maximum likelihood estimation via EM algorithm

The $Q-$function is given by

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \quad = \quad & c + \sum_{i=1}^{n}\sum_{j=1}^{G}\left(\widehat{z}_{ij}^{(k+1)}(\log(p_j) - \frac{1}{2}\log|\Gamma_j|) - \frac{1}{2\Gamma_j}\left(\widehat{u}_{ij}^{(k+1)}(y_i-\mu_j)^2\right.\right. \\
& \left.\left. -2(y_i-\mu_j)\Delta_j\widehat{zut}_{ij}^{(k+1)} + \Delta_j^2\widehat{zut^2}_{ij}^{(k+1)}\right) + E[Z_{ij}\log(h(U_i;\boldsymbol{\nu}))|\widehat{\boldsymbol{\theta}}^{(k)}, y_i]\right).
\end{aligned}
$$

The Algorithm

E-step: Given a current estimate $\widehat{\boldsymbol{\theta}}^{(k)}$, compute $\widehat{z}_{ij}$, $\widehat{zu}_{ij}$, $\widehat{zut}_{ij}$, $\widehat{zut^2}_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, G$.

CM-steps: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[\ell_c(\boldsymbol{\theta})|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}]$ over $\boldsymbol{\theta}$, obtaining

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

$$\widehat{\mu}_j^{(k+1)} = \sum_{i=1}^{n} (\widehat{zu}_{ij}^{(k+1)} y_i - \widehat{\Delta}_j^{(k)} \widehat{zut}_{ij}^{(k+1)}) / \sum_{i=1}^{n} \widehat{zu}_{ij}^{(k+1)},$$

$$\widehat{\Delta}_j^{(k+1)} = \sum_{i=1}^{n} (y_i - \widehat{\mu}_j^{(k+1)}) \widehat{zut}_{ij}^{(k+1)} / \sum_{i=1}^{n} \widehat{zut^2}_{ij}^{(k+1)}, \quad \text{and}$$

$$\widehat{\Gamma}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \left(\widehat{zu}_{ij}^{(k+1)} (y_i - \widehat{\mu}_j^{(k+1)})^2 - 2(y_i - \widehat{\mu}_j^{(k+1)}) \widehat{\Delta}_j^{(k+1)} \widehat{zut}_{ij}^{(k+1)} + (\widehat{\Delta}_j^2)^{(k+1)} \widehat{zut^2}_{ij}^{(k+1)} \right)}{\sum_{i=1}^{n} \widehat{z}_{ij}^{(k+1)}}$$

The Algorithm

CML-step: Update $\widehat{\boldsymbol{\nu}}^{(k)}$ by maximizing the actual marginal log-likelihood function, obtaining

$$\widehat{\boldsymbol{\nu}}^{(k+1)} = \text{argmax}_{\boldsymbol{\nu}} \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{G} \widehat{p}_j^{(k+1)} g(y_i|\widehat{\mu}_j^{(k+1)}, \widehat{\sigma}_j^{2(k+1)}, \widehat{\lambda}_j^{(k+1)}, \boldsymbol{\nu})\Big).$$

This process is iterated until a suitable convergence rule is satisfied, e.g.

- if $||\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)}||$ is sufficiently small;
- or until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta})$, like $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})||$ or $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1||$, is small enough.

Simulation studies

- Study 1: Investigating the ability of the FM-SMSN models in clustering observations;
- We generated 500 samples from a mixture of two SMSN densities and, for each sample, proceeded clustering ignoring the known true classification;
- The FM-SMSN models were fitted using the algorithm described earlier; The estimate of the posterior probability that an observation y_i belongs to the jth component of the mixture is given by

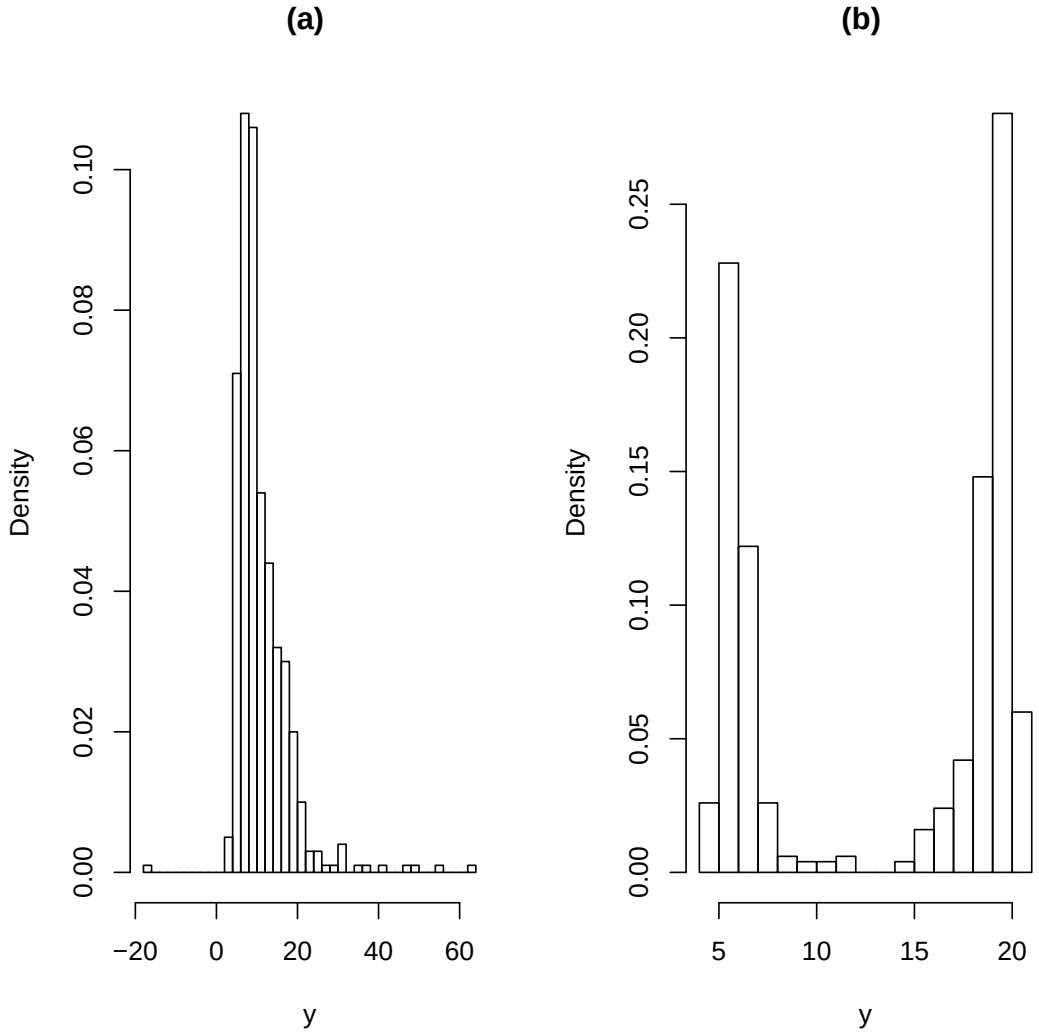
$$\widehat{z}_{ij} = \frac{\widehat{p}_j g(y_i|\widehat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^{G} \widehat{p}_j g(y_i|\widehat{\boldsymbol{\theta}}_j)};$$

- Then, the threshold value 0.5 was used to allocate the observation to some specific component.

Simulation studies

- For sample l, $l = 1, \cdots, 500$, we computed the rate r_l, the number of correct allocations divided by the sample size n;
- When fitting the FM-SSL model, the parameter ν was considered known and we fixed $\nu = 2$.
- We fixed the parameter values at $\mu_1 = 15$, $\mu_2 = 20$, $\sigma_1^2 = 20$, $\sigma_2^2 = 16$, $\lambda_1 = 6$, $\lambda_2 = -4$, $p_1 = 0.8$ and $\nu = 3$. For the SCN case we fixed $(\nu_1, \gamma_1) = (\nu_2, \gamma_2) = (0.2, 0.2)$;
- The sample sizes considered were $n = 100$, 500, 1000.

a) poorly separated; (b) well separated



00) with two components: (a)

Mean right allocations rates

True model	Sample size	Fitted model				
		FM-NOR	FM-SN	FM-ST	FM-SCN	FM-SSL
FM-ST	100	0.4102	0.6825	0.7872	0.7879	0.7705
	500	0.3067	0.7521	0.8369	0.8340	0.8329
	1000	0.2942	0.7834	0.8381	0.8375	0.8361
FM-SCN	100	0.5601	0.6967	0.7783	0.7778	0.7686
	500	0.6072	0.7904	0.8324	0.8340	0.8323
	1000	0.6406	0.8139	0.8349	0.8346	0.8358
FM-SSL	100	0.5765	0.7602	0.7755	0.7669	0.7562
	500	0.6162	0.8216	0.8336	0.8341	0.8324
	1000	0.6287	0.8340	0.8336	0.8327	0.8341

Study 2: Asymptotic properties

- The main focus are the evaluations of bias and mean square error;
- We consider only the FM-ST model and the following sets of true parameter values:
  1. The same set used earlier (poorly separated components);
  2. Changing the true values of the scale and mixing proportion parameters, now using $\sigma_1 = \sigma_2 = 1$ and $p_1 = 0.4$ (well separated components). Values for the remaining parameters are the same as before.
- Sample sizes were fixed as n =100, 500, 1000, 5000 and 10000. For each combination of parameters and sample size, 500 samples from the FM-ST model were artificially generated;

Study 2: Asymptotic properties

- Then we compute the bias and mean squared error (MSE) over all samples. For $\mu_j,\ j = 1, 2$, they are defined as

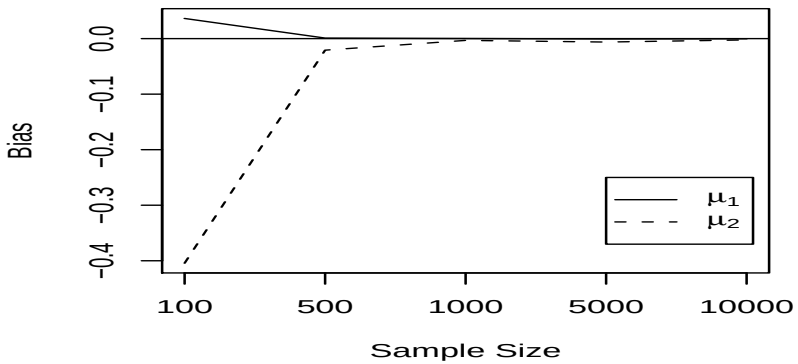
$$\text{bias} = \frac{1}{500}\sum_{i=1}^{500}\widehat{\mu}_j^{(i)} - \mu_j \quad \text{and} \quad \text{MSE} = \frac{1}{500}\sum_{i=1}^{500}(\widehat{\mu}_j^{(i)} - \mu_j)^2,$$

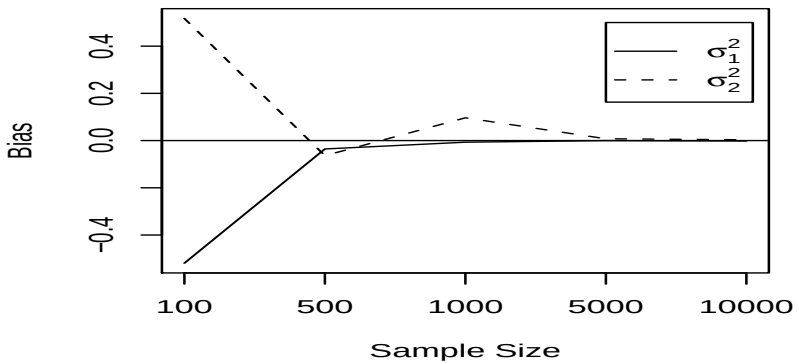
  respectively, where $\widehat{\mu}_j^{(i)}$ is the ECME estimate of μ_j when the data is sample i;
- Definitions for the other parameters are obtained by analogy.

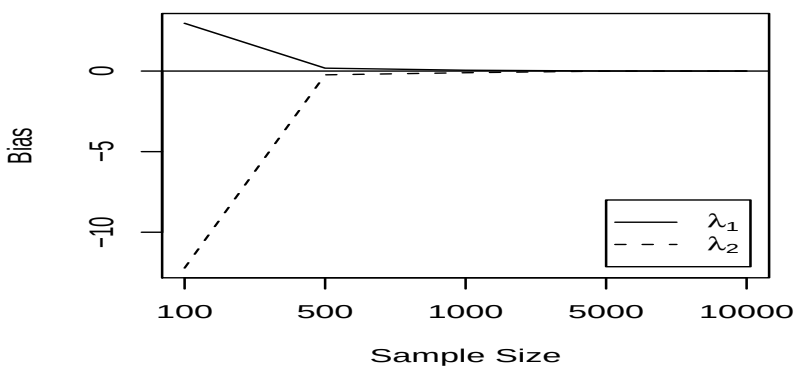
Bias and MSE – poorly separated case

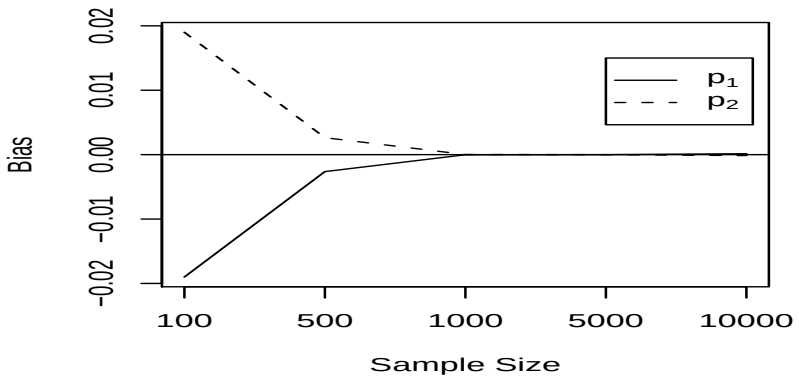
Measure	Parameter	Sample size 100	500	1000	5000	10000
bias	μ_1	3.626685e-02	9.553629e-04	2.251980e-04	-6.480655e-04	-3.259887e-04
	μ_2	-4.043036e-01	-2.082871e-02	-3.225519e-03	-6.226229e-03	-2.37579e-03
	σ_1^2	-5.192698e-01	-3.552685e-02	-6.928553e-03	2.437042e-05	-1.77739e-03
	σ_2^2	5.172681e-01	-6.393652e-02	9.639911e-02	7.684399e-03	3.255335e-03
	λ_1	2.962971e+00	1.799639e-01	5.203251e-02	4.408876e-03	4.129488e-04
	λ_2	-1.221872e+01	-2.345765e-01	-1.041270e-01	-2.643574e-03	-3.593367e-03
	ν	1.178223e+02	1.442717e-01	5.591489e-02	1.192845e-02	4.705524e-03
	p_1	-1.899612e-02	-2.633289e-03	-1.884522e-05	3.049117e-07	1.450857e-04
MSE	μ_1	1.151169e-01	9.618573e-03	3.961588e-03	2.836060e-04	7.941788e-05
	μ_2	2.057845e+00	1.306358e-01	4.897538e-02	3.154556e-03	9.266971e-05
	σ_1^2	1.833444e+01	1.107149e+00	2.085420e-01	3.284945e-03	5.445568e-04
	σ_2^2	1.862787e+02	1.571325e+01	4.580214e+00	6.833315e-02	1.346153e-02
	λ_1	1.157501e+02	9.825899e-01	2.416234e-01	8.612938e-03	1.380906e-03
	λ_2	7.212604e+02	2.469349e+00	5.175716e-01	8.273690e-03	1.710742e-03
	ν	1.014661e+03	5.657795e-01	1.021304e-01	1.472001e-02	6.089940e-03
	p_1	7.401136e-03	7.574217e-04	3.783682e-04	6.581538e-05	2.512335e-05

Bias – poorly separated case





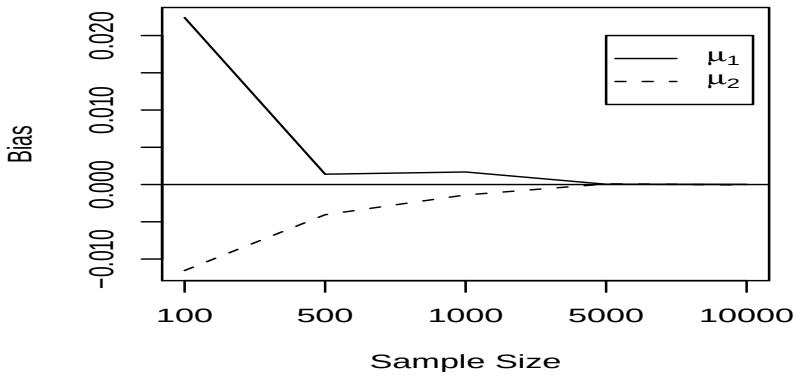


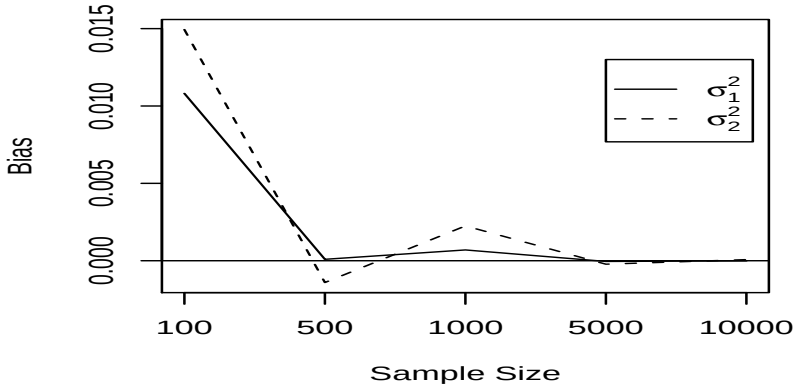


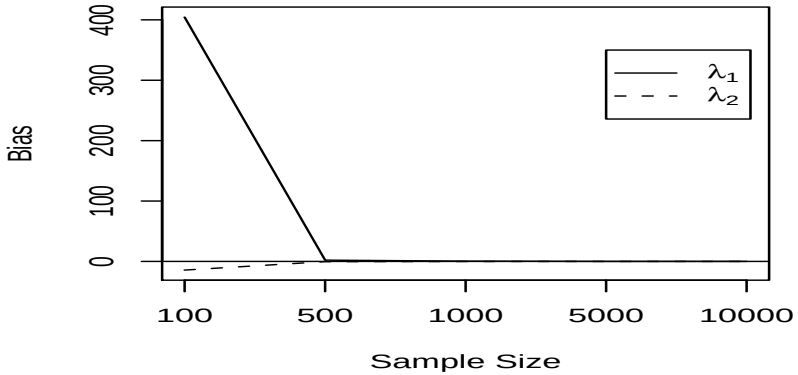
Bias and MSE – well separated case

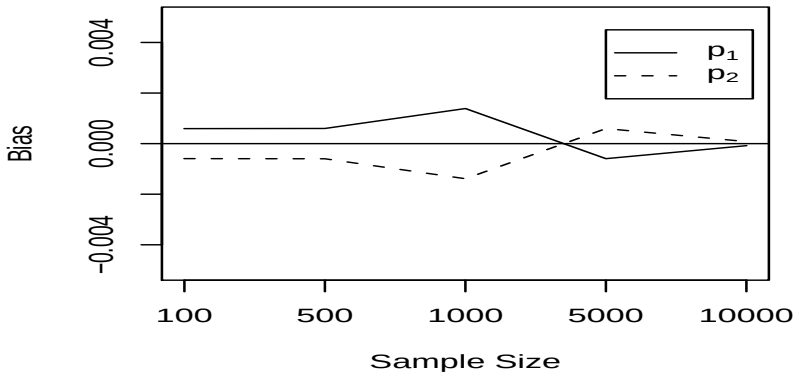
Measure	Parameter	Sample size				
		100	500	1000	5000	10000
Bias	μ_1	2.240432e-02	1.387814e-03	1.685946e-03	4.182902e-05	2.235176e-05
	μ_2	-1.153799e-02	-4.052004e-03	-1.395336e-03	7.981414e-05	-4.667282e-05
	σ_1^2	1.080980e-02	9.245055e-05	6.969676e-04	-4.585860e-05	-9.148042e-06
	σ_2^2	1.493537e-02	-1.407618e-03	2.249074e-03	-2.232164e-04	7.266806e-05
	λ_1	4.044328e+02	2.150470e+00	4.582539e-01	6.167612e-03	1.653220e-03
	λ_2	-1.431243e+01	-3.589517e-01	-1.137305e-01	-1.626848e-02	2.566557e-03
	ν	2.527001e+00	1.765753e-01	4.866826e-02	1.145722e-02	-1.965416e-03
	p_1	5.931726e-04	5.989383e-04	1.387312e-03	-5.946862e-04	-8.035652e-05
MSE	μ_1	5.235930e-03	4.353205e-04	1.729061e-04	7.793829e-06	1.072331e-06
	μ_2	6.361194e-03	7.572480e-04	2.695497e-04	1.874958e-05	5.167974e-06
	σ_1^2	3.956735e-02	1.528215e-03	1.997454e-04	1.720467e-06	1.546036e-07
	σ_2^2	9.320955e-02	5.710687e-03	1.208212e-03	1.520323e-05	2.486816e-06
	λ_1	4.231106e+03	5.384741e+01	7.080843e+00	1.445613e-01	1.548913e-02
	λ_2	1.402899e+03	4.138201e+00	8.713269e-01	3.303539e-02	4.560911e-03
	ν	1.420523e+02	3.322731e-01	8.918025e-02	1.585834e-02	8.78927e-03
	p_1	3.077124e-03	5.033739e-04	2.744831e-04	4.742073e-05	2.37786e-05

Bias – well separated case









Study 3: Model Selection

- We compared the ability of some classical procedures in choosing between the underlying FM-SMSN models;
- We fixed the number of components ($G = 2$), sample size ($n = 1000$) and parameter values ($\mu_1 = 20$, $\mu_2 = 30$, $\sigma_1^2 = 15$, $\sigma_2^2 = 40$, $\lambda_1 = 2$, $\lambda_2 = 10$ and $p_1 = 0.6$);
- Several values for the degrees of freedom parameter ν (equal to 3, 6, 10, 15 and 30) were taken into account;
- For each combination value of the parameters, 500 samples from a mixture of skew-t densities were artificially generated and, for each sample, we fitted the FM-SN and the FM-ST models;
- For each fitted model, we computed the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Efficient Determination Criterion (EDC) and the Integrated Completed Likelihood Criterion;

Study 3: Model Selection

- AIC, BIC and EDC have the form

$$-2\ell(\widehat{\boldsymbol{\theta}}) + \gamma c_n,$$

where $\ell(\cdot)$ is the actual log-likelihood, γ is the number of free parameters that have to be estimated under the model and the penalty term c_n is a convenient sequence of positive numbers;

- We have $c_n = 2$ for AIC and $c_n = \log(n)$ for BIC. For the EDC criterion, c_n is chosen so that it satisfies the conditions $c_n/n \to 0$ and $c_n/(\log\log n) \to 0$ when $n \to \infty$. Here we use $c_n = 0.2\sqrt{n}$;

Study 3: Model Selection

The ICL is defined as

$$-2\ell^*(\widehat{\boldsymbol{\theta}}) + \gamma \log(n),$$

where $\ell^*(\cdot)$ is the integrated log-likelihood of the sample and the indicator latent variables, given by

$$\ell^*(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{g} \sum_{j \in \mathcal{C}_i} \log(\widehat{p}_i g(y_j|\widehat{\boldsymbol{\theta}}_i)),$$

where $\mathcal{C}_i$ is a set of indexes defined as: j belongs to $\mathcal{C}_i$ if, and only if, the observation y_j is allocated to component i by the clustering method defined ealier.

Number of times (out of 500) the true model is chosen using different criteria

	Degrees of freedom				
Criterion	3	6	10	15	30
AIC	500	488	424	330	139
BIC	500	469	279	135	22
EDC	500	475	296	145	33
ICL	438	93	7	1	0

Application with real data: BMI for men aged between 18 to 80 years

- The data set comes from the National Health and Nutrition Examination Survey, made by the National Center for Health Statistics (NCHS) of the Center for Disease Control (CDC) in the USA;
- The problem of obesity has attracted attention in the last few years due to its strong relationship with many chronic diseases;
- Body mass index (BMI, kg/m^2) has become the standard measure for overweight and obesity;
- BMI is defined as the ratio of body weight in kilograms and body height in meters squared.

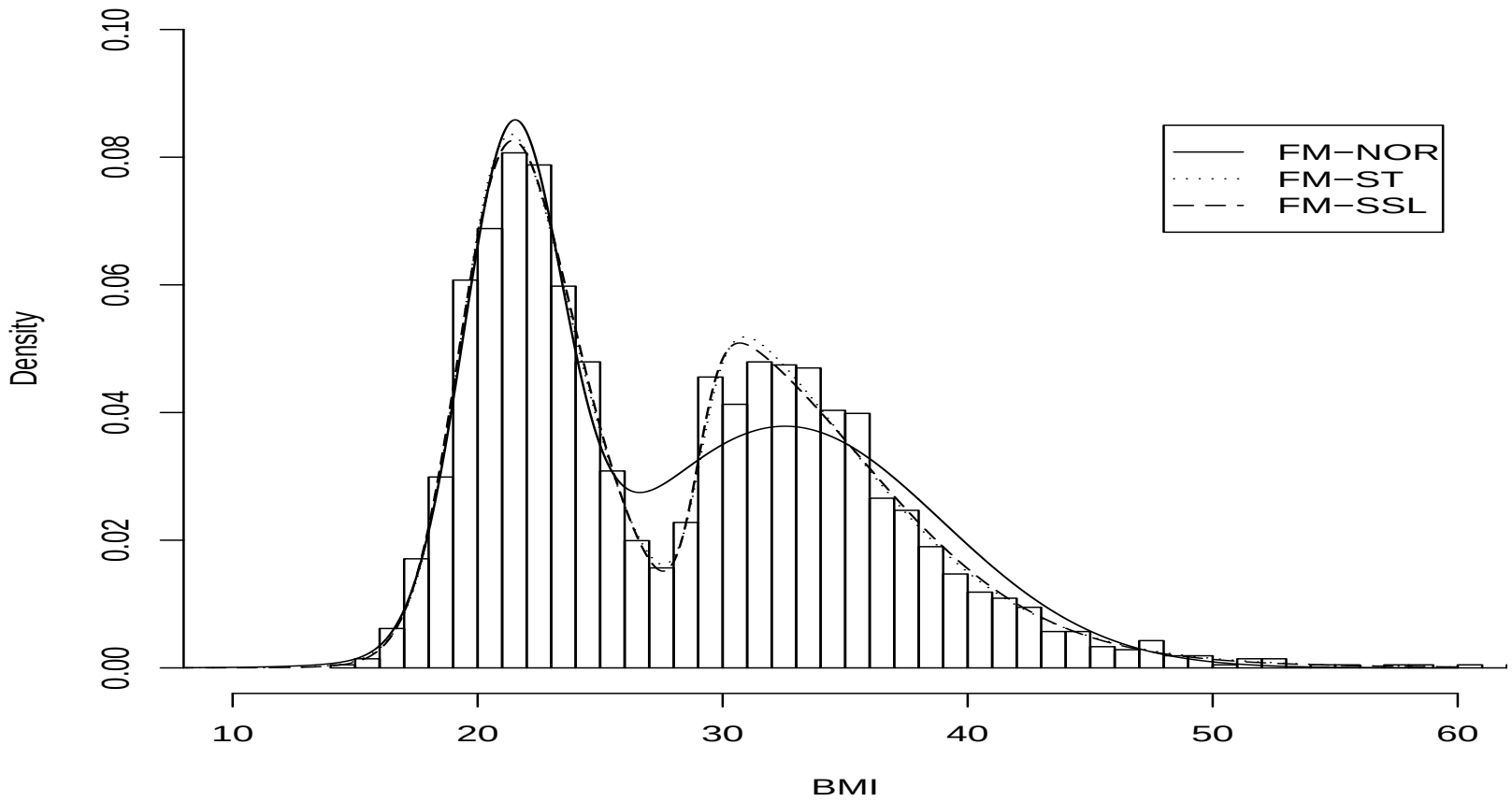
MLE results for fitting various mixture models to the BMI data

Parameter	FM-NOR	FM-SN	FM-ST	FM-SCN	FM-SSL
p_1	0.391 (0.0188)	0.528 (0.0125)	0.538 (0.0142)	0.538 (0.0140)	0.536 (0.0135)
μ_1	21.412 (0.0936)	19.500 (0.2429)	19.572 (0.2432)	19.487 (0.2363)	19.512 (0.2372)
μ_2	32.548 (0.3681)	28.760 (0.1456)	29.100 (0.1652)	29.023 (0.165)1	28.972 (0.1544)
σ_1^2	4.071 (0.0873)	14.365 (0.2841)	12.916 (0.3072)	12.864 (0.3022)	9.896 (0.2636)
σ_2^2	41.191 (0.1578)	63.217 (0.1580)	45.841 (0.3100)	44.543 (0.4440)	36.115(0.3414)
λ_1	-	1.902 (0.3446)	1.900 (0.3723)	2.003 (0.3973)	1.955 (0.3636)
λ_2	-	10.588 (2.7408)	7.131 (1.8474)	7.656 (2.1135)	8.330 (2.1402)
ν	-	-	8.759 (2.1238)	0.141 (0.061)	2.421 (0.4169)
γ	-	-	-	0.284 (0.069)	-

Model Selection Criteria – BMI Data

Criterion	FM-NOR	FM-SN	FM-ST	FM-SCN	FM-SSL
AIC	13833.35	13750.89	13726.67	13726.73	13726.56
BIC	13961.61	13790.46	13771.89	13777.61	13771.78
EDC	13869.25	13801.44	13784.12	13791.36	13784.00
ICL	14280.49	13865.16	13898.14	13902.63	13884.41

Histogram of BMI data



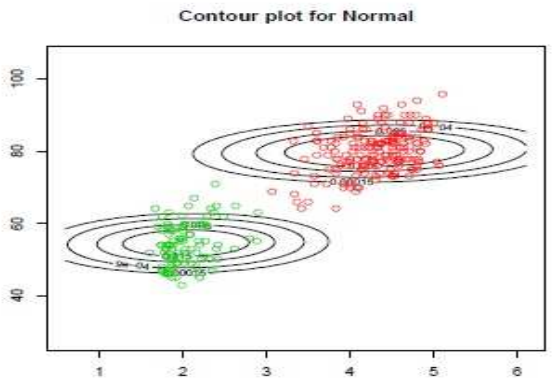
Multivariate application: Old Faithful geyser

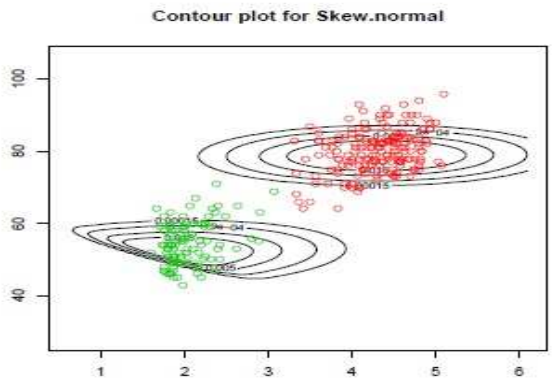
- The data set comes from the Yellowstone National Park, which was created in 1872 and was the first America's national park
- The data consists of 272 pairs of measurements, referring to the time interval between the starts of successive eruptions and the duration of the subsequent eruption;

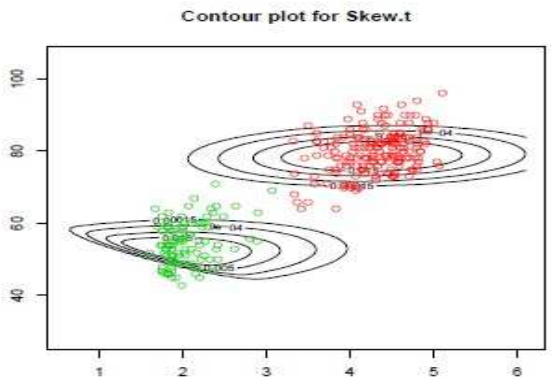
Model Selection Criteria – Old Faithful Data

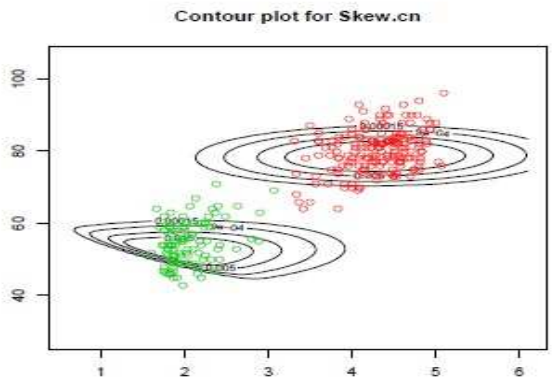
Criterion	FM-NOR	FM-SN	FM-ST	FM-SCN
AIC	2292.53	2265.55	2265.09	2264.98
BIC	2350.22	2323.24	2322.78	2322.68
EDC	2313.31	2286.32	2285.86	2285.76
ICL	2350.72	2325.18	2324.87	2324.56

Contours: Old Faithful Data









Thank you for your attention!