Discussion of *150 years of finite mixture analysis how statisticians reveal hidden structures in complex data* by Sylvia Frühwirth-Schnatter

Luca Tardella
Sapienza Università di Roma

Statistical methods and models for complex data
800 years of research to understand a complex world
Padova 21-23 September 2022

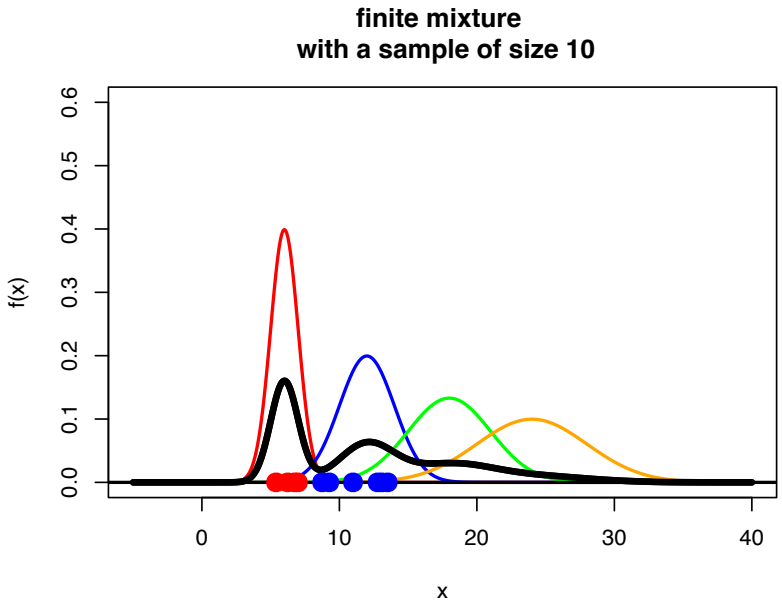
- Finite mixture modelling

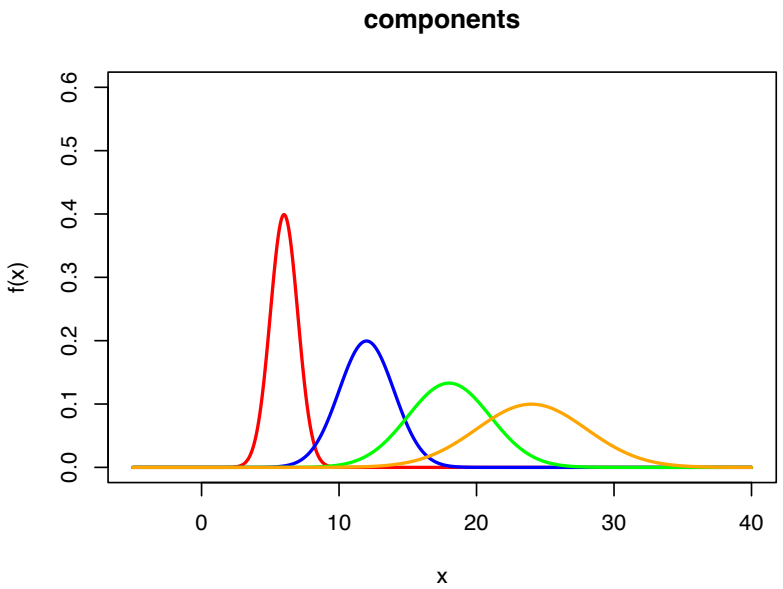
$$p(\boldsymbol{y}_i|\vartheta_K, K) = \sum_{k=1}^{K} \eta_k f_{\mathcal{T}}(\boldsymbol{y}_i|\theta_K)$$

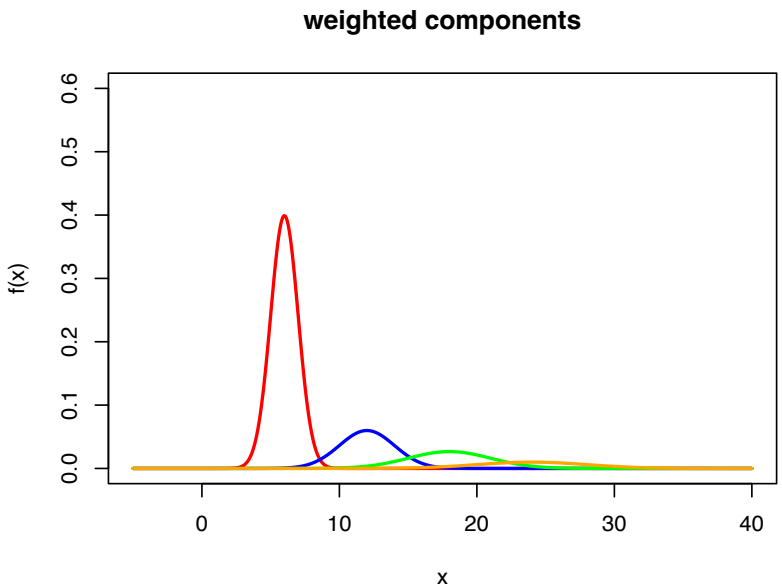
  with fixed K parametric components
- Mixture of finite mixtures modelling

$$p(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = \sum_{K=1}^{\infty} p(\boldsymbol{y}_i|\vartheta_K, K)p(K)$$

- Generalized mixture of finite mixtures by means of data-augmented hierarchical modelling


finite mixture
with a sample of size 10
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
x


components
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
x


weighted components
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
x

Generalized mixture of finite mixtures (MFM) model

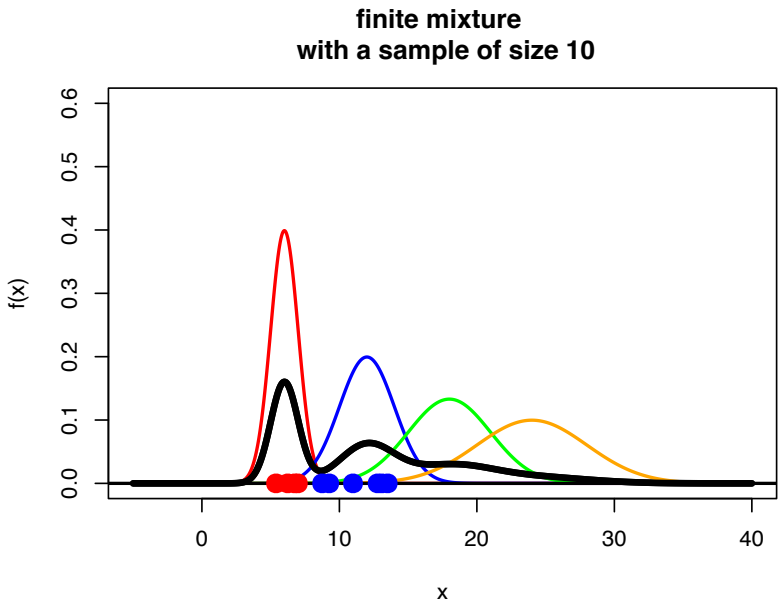
- A fully Bayesian mixture model is defined in a hierarchical way:

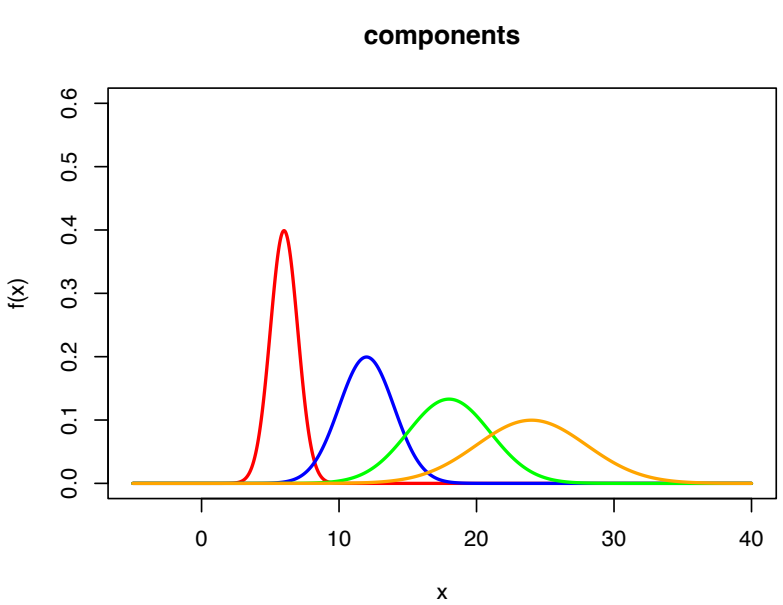
$$
\begin{aligned}
K &\sim p(K), \\
\eta_1, \ldots, \eta_K | K, \gamma_K &\sim \mathcal{D}_K(\gamma_K), \\
S_i | K, \eta_1, \ldots, \eta_K &\sim \mathcal{M}(1; \eta_1, \ldots, \eta_K), \text{ independently for } i = 1, \ldots, N, \\
\phi &\sim p(\phi), \\
\boldsymbol{\theta}_k | \phi &\sim p(\boldsymbol{\theta}_k | \phi), \text{ independently for } k = 1, \ldots, K, \\
\mathbf{y}_i | K, S_i = k, \boldsymbol{\theta}_k &\sim f_{\mathcal{T}}(\mathbf{y}_i | \boldsymbol{\theta}_k), \text{ independently for } i = 1, \ldots, N.
\end{aligned}
$$

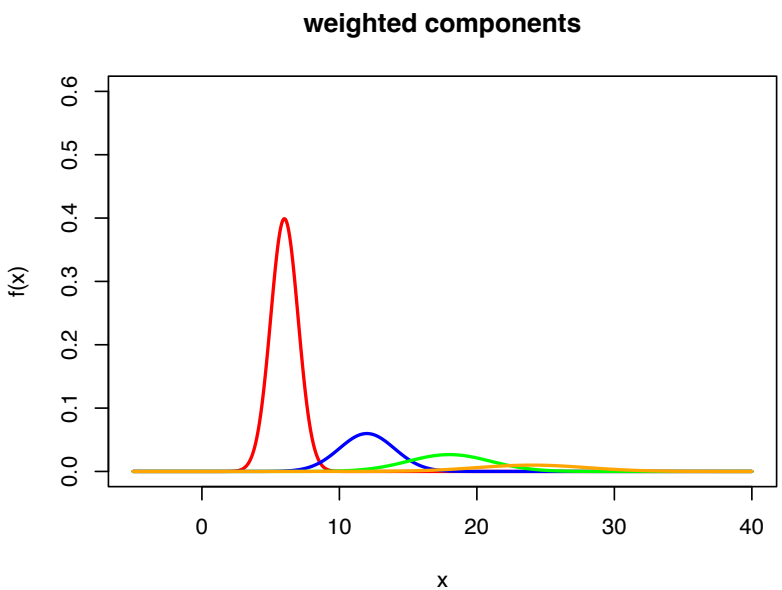
- Generic framework with no specific restrictions on
  - $f_{\mathcal{T}}(\cdot | \boldsymbol{\theta}_k)$ (parametric family),
  - observations $\mathbf{y}_i$ can be univariate or multivariate, continuous, discrete-valued, mixed-type, time series data, outcomes of a regression model, ...
  - the prior $p(K)$ (e.g., parametric pmf, $\delta_{\{K_{fix}\}}$, $\delta_{\{\infty\}}$, ...),
  - **and ...**

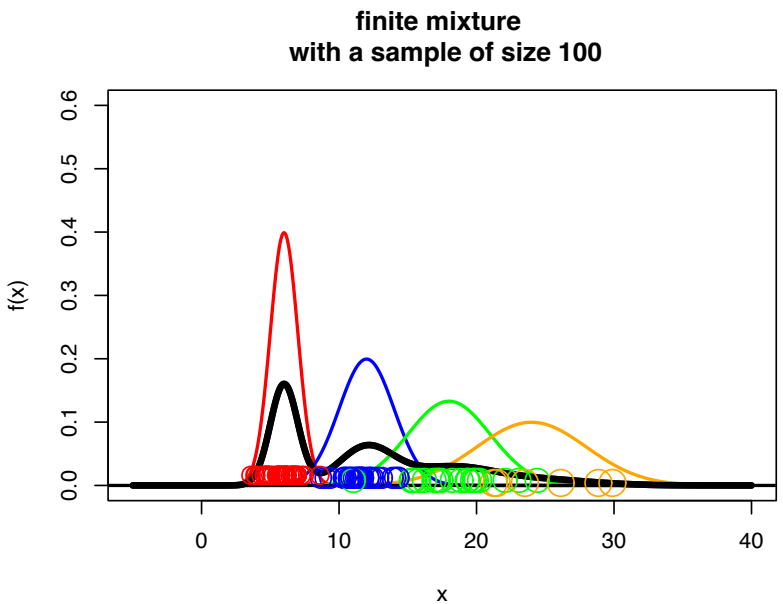
Key point: leveraging on the distinction between K and K_+

- K is the number of components in the finite mixture model
- K_+ is the random number of latent labels representing each components in the sample of size N. It is indeed a $K_+(N) \leq K$
- the difference $K - K_+(N)$ depends on N and the magnitude of the weights and it can be a priori regulated by means of the sparsity of the Dirichlet(γ_K) with $\gamma_K = \frac{\alpha}{K}$ as in the Dynamic MFMs or more generally with an extra layer of a priori randomness on α
- the difference $K - K_+(N)$ in the species sampling sequence of latent labels is the number of unseen species
- there is an ingenious use of this feature in the EPPF which allows to develop and exploit a new Telescopic Sampler as well as relate GMFM with the theory of EPPF within BNP (DPM, Gibbs-type, etc.)


finite mixture
with a sample of size 10
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40


components
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40


weighted components
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40


finite mixture
with a sample of size 100
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40

My Comments/questions

1. Inferring the (unknown) number of groups in the observed sample of size N or the number of groups in the underlying population?
2. How can we cope with the impact of the prior on the posterior conclusions?
3. Possible model/component mispecification: how it hurts and how one can deliver diagnostics (especially with complex components) for global/local lack of fit

K_+ or K?

- parsimonious approach for removing the components which are difficult to discern in a finite sample (bias? minorities?)
- K_+ and K will agree with $N \to \infty$
- in some setting of MFM K can be consistently estimated (Guha et al. 2019)
- should we be worried if the posteriors of K and K_+ do not merge with $N = 11922$?

Impact of the prior on the posterior conclusions

- single out a principled default prior?
- sensitivity of posterior masses?
- what about BF for hypothesis testing?