Response to peer review: Gaussian mixtures in R

by Bastien CHASSAGNOL and Antoine Bichat and Cheima Boudjeniba
and Pierre-Henri Wuillemin and Mickaël Guedj and Gregory Nuel and Etienne Becht

April 28, 2023

Dear editor and authors, we would like to thank you for the evaluation of the revised version of our manuscript and for your constructive suggestions for improving it. We have revised the second version of our manuscript by performing a comparison of the packages using additional simulations in a multivariate (10-dimensional) framework and by commenting on the relevant theoretical aspects of high-dimensional Gaussian mixture models.

– Editor review –

Please separate the appendix as a separate document, and include the name of this file in the supplementary material.

In this reviewed version, we have separated the appendix from the main part of the article and made it available in the supplementary material, in folder supplementary material along with the associated datasets and figures. In addition to this final, ready-to-published version, we provide an easy to review version of our article (that merges together the main text with the appendix, in order to keep hyperlinks) with highlighted changes in blue colour, available publicly on personal Github account, with respectively the HTML version: proofreading-version-chassagnol-becht-nuel-benchmark-of-Gaussian-mixtures.html and pdf version: proofreading-version-chassagnol-becht-nuel-benchmark-of-Gaussian-mixtures.pdf.

– AE review –

The revision has made some substantial changes and addressed all comments raised by the two reviewers. There still exists one problem: the authors extended their examples from the univariate setting to a bivariate scenario, but not to the multidimensional scenario as the reviewer suggested. This update is not enough for R users to deal with real-world data in a practical sense. Multivariate framework and examples (more than two variables) should be covered by the paper.

We thank the associated editor for acknowledging the additional simulations established to address the comments and extensions suggested by the reviewers. As advised, we extended our benchmark of Gaussian mixture models to a higher dimensional configuration, comparing the performance of several packages, including two additional ones dedicated to high-dimensional setting, namely *EMMIXmfa* and *HDclassif* and initialisation methods across several configurations of a multivariate GMM spanning over a 10-dimensional space. We have notably taken advantage of the simulation features implemented by the *MixSim* package to test a large set of mixture configurations with differing imbalances in component proportions, covariance structures and overlap between clusters. The whole set of configurations tested are shown in Table 1 below:

ID	OVL	Number of observations	Proportions	Spherical
HD1a	1e-04	200	0.5 / 0.5	☑
HD1b	1e-04	2000	0.5 / 0.5	☑
HD2a	1e-04	200	0.19 / 0.81	☑
HD2b	1e-04	2000	0.19 / 0.81	☑
HD3a	1e-04	200	0.5 / 0.5	☒
HD3b	1e-04	2000	0.5 / 0.5	☒
HD4a	1e-04	200	0.21 / 0.79	☒
HD4b	1e-04	2000	0.21 / 0.79	☒
HD5a	2e-01	200	0.5 / 0.5	☑
HD5b	2e-01	2000	0.5 / 0.5	☑
HD6a	2e-01	200	0.15 / 0.85	☑
HD6b	2e-01	2000	0.15 / 0.85	☑
HD7a	2e-01	200	0.5 / 0.5	☒
HD7b	2e-01	2000	0.5 / 0.5	☒
HD8a	2e-01	200	0.69 / 0.31	☒
HD8b	2e-01	2000	0.69 / 0.31	☒

Figure 1: The 16 parameter configurations tested to generate the samples in a high dimensional context. The first digit of each ID index refers to an unique parameter configuration, identified by its level of overlap, entropy and topological structure (either circular or ellipsoidal), while the lowercase letter depicts the number of observations, a) with $n = 200$ and b) with $n = 2000$.

and the corresponding performance results are all reported in the appendix and summarised in the main text of our article.

The main conclusion from our extended benchmark of our extended benchmark is that most of the observations and recommendations on the use of packages with respect to the characteristics of the mixture model in a bivariate setting hold in a higher dimensional setting. Namely, the criterion threshold and the implementation choices for dealing with numerical underflow significantly impact the performance of the EM algorithm, especially with strongly overlapping and highly-unbalanced components. Additionally, we keep on observing a significant dichotomy between the benchmarked packages that is closely related to the choice of the criterion threshold (either absolute or relative), as illustrated below on Figure 2

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	rebmix	0.5611 // 0.0058	0.2364 // 0.0028	0.3035 // 0.0026	0.019 // 0.00071	0.372 // 0.018	0.915 // 0.04	98 // 100
mclust / flexmix / GMKMcharlie	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0095 // 0.00071	0.053 // 0.018	0.14 // 0.04	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0095 // 0.00071	0.053 // 0.018	0.14 // 0.04	100 // 100
	rebmix	0.3134 // 0.0058	0.1305 // 0.0029	0.1729 // 0.0026	0.0039 // 0.0022	0.2 // 0.02	0.537 // 0.0044	88 // 83
bgmm	hc	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0098 // 0.00071	0.053 // 0.018	0.139 // 0.041	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0097 // 0.00071	0.053 // 0.018	0.139 // 0.04	100 // 100
	rebmix	0.7437 // 0.1977	0.3409 // 0.0028	0.3895 // 0.1946	0.02 // 0.00034	0.308 // 0.017	1.602 // 0.926	97 // 99
EMCluster	hc	0.0577 // 0.0058	0.0289 // 0.0028	0.0264 // 0.0026	0.0093 // 0.00063	0.054 // 0.018	0.139 // 0.041	100 // 100
	kmeans	0.0577 // 0.0058	0.0288 // 0.0028	0.0264 // 0.0026	0.0092 // 0.00039	0.054 // 0.017	0.139 // 0.04	100 // 100
	rebmix	0.6887 // 0.3391	0.2466 // 0.1276	0.4201 // 0.1969	0.019 // 0.048	0.519 // 0.289	1.721 // 0.875	87 // 83
HDclassif	hc	11.3787 // 11.3322	11.3499 // 11.3293	0.0264 // 0.0026	0.0094 // 0.00071	11.166 // 11.179	0.139 // 0.04	100 // 100
	kmeans	11.3833 // 11.3303	11.3543 // 11.3271	0.0264 // 0.0026	0.0094 // 0.00068	11.162 // 11.183	0.139 // 0.04	100 // 100
	rebmix	11.5949 // 11.6085	11.5167 // 11.6055	0.072 // 0.0026	0.0024 // 0.0022	11.227 // 11.369	0.27 // 0.044	87 // 83
EMMIXmfa	hc	5.9739 // 5.9149	4.0397 // 3.9727	1.8288 // 1.8228	0.32 // 0.47	6.999 // 7.042	4.25 // 4.296	100 // 100
	kmeans	5.972 // 5.8863	4.0431 // 3.9596	1.8283 // 1.8259	0.33 // 0.43	6.997 // 7.051	4.255 // 4.34	100 // 100
	rebmix	5.9835 // 5.9078	4.0477 // 3.9671	1.8257 // 1.8244	0.37 // 0.46	6.994 // 7.045	4.232 // 4.301	87 // 83

(a) MSE and Bias associated to scenarios HD1a) and HD1b) in Table 1.

Package	Initialisation Method	Global MSE p	Global MSE μ	Global MSE σ	Global Bias p	Global Bias μ	Global Bias σ	% Success
mixtools / Rmixmod / RGMMBench	hc	18.6085 // 0.6735	3.566 // 0.0536	14.9495 // 0.6193	0.23 // 0.0017	3.327 // 0.107	14.475 // 0.649	100 // 100
	kmeans	16.7452 // 0.6735	2.9005 // 0.0536	13.7962 // 0.6193	0.2 // 0.0016	2.859 // 0.107	12.552 // 0.649	100 // 100
	rebmix	22.2986 // 0.6738	4.3127 // 0.0536	17.8418 // 0.6196	0.25 // 0.0021	3.768 // 0.108	16.249 // 0.648	95 // 100
mclust / flexmix / GMKMcharlie	hc	20.6916 // 0.728	4.5072 // 0.0696	16.0328 // 0.6557	0.28 // 0.004	4.381 // 0.459	18.656 // 2.07	100 // 100
	kmeans	17.9622 // 0.7169	3.7547 // 0.0671	14.1405 // 0.6474	0.27 // 0.062	3.88 // 0.465	16.802 // 2.049	100 // 100
	rebmix	22.4636 // 0.7553	4.7502 // 0.0678	17.5784 // 0.6853	0.26 // 0.0054	4.165 // 0.158	17.735 // 0.725	94 // 98
bgmm	hc	35.6085 // 13.8411	12.8826 // 3.6502	22.3428 // 10.0718	0.29 // 0.46	6.212 // 5.661	26.812 // 23.753	100 // 100
	kmeans	33.8007 // 12.5545	11.7236 // 3.1419	21.7292 // 9.2934	0.28 // 0.47	6.348 // 5.654	26.546 // 23.141	100 // 100
	rebmix	35.3167 // 13.106	12.2374 // 3.3747	22.6635 // 9.6213	0.37 // 0.42	6.007 // 5.273	26.287 // 23.02	96 // 100
EMCluster	hc	23.4472 // 16.4192	6.2451 // 4.9191	17.0777 // 11.3124	0.35 // 0.51	5.469 // 6.279	23.503 // 22.823	99 // 100
	kmeans	21.0058 // 14.6852	5.9051 // 4.3684	14.9329 // 10.4293	0.38 // 0.42	5.604 // 6.592	24.628 // 24.706	100 // 100
	rebmix	23.0408 // 19.7372	6.4923 // ?	16.3824 // 12.5099	0.36 // 0.35	5.272 // 5.454	23.419 // 23.9	93 // 98
HDclassif	hc	36.5924 // 33.4077	16.108 // 14.4307	20.3809 // 18.8638	0.3 // 0.44	12.736 // 13.085	25.393 // 29.363	100 // 100
	kmeans	34.7935 // 30.1935	15.4326 // 13.78	19.2529 // 16.2856	0.4 // 0.41	12.756 // 12.756	24.908 // 25.151	100 // 100
	rebmix	38.9707 // 24.0327	16.134 // 12.9266	22.6961 // 11.0138	0.25 // 0.21	12.811 // 12.275	25.79 // 15.996	95 // 98

(b) MSE and Bias associated to scenarios HD8a) and HD8b), in Table 1.

Figure 3: We delimitate using doubled backslashes the scores with $n = 200$ and $n = 2000$ observations for each input of the summary table.

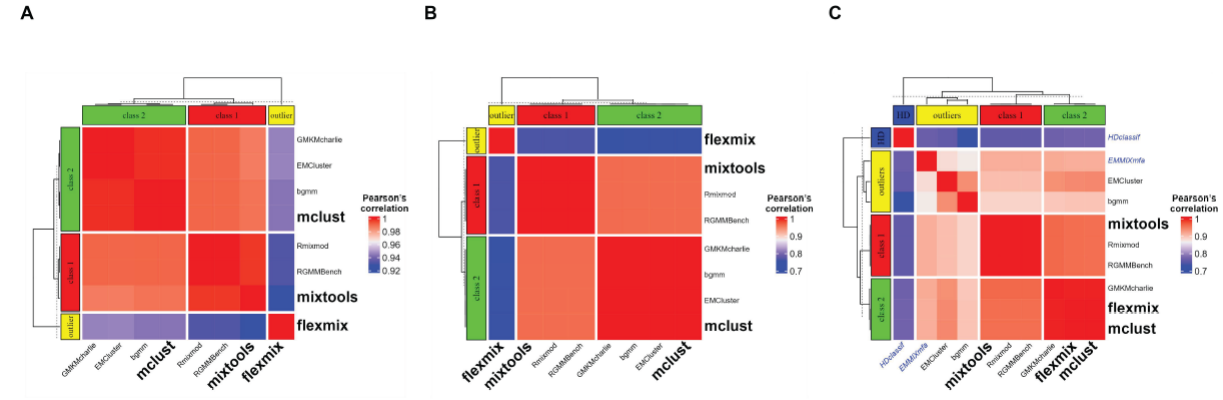

Figure 2: Panels A, B and C show respectively the heatmap of the Pearson correlation in the univariate, bivariate and high-dimensional framework of the parameters estimated by the benchmarked packages, evaluated for each of them at the most complex scenario (overlapping and unbalanced clusters), with k-means initialisation. We can note significant differences between a first class of packages, composed on the one hand of *mixtools*, *Rmixmod* and our custom R implementation *RGMMBench*, and on the other hand of *flexmix*, *GMKMCharlie* and *mclust*, joined at least in a bivariate framework by *bgmm* and *EMCluster*.

On the contrary, the packages dedicated to high-dimension tend to underperform with intermediate large datasets. From this, we assume that projection into a lower-dimensional space is only beneficial in a really high-dimensional setting, for example when the number of dimensions exceeds the number of observations. However, a larger sample space revealed that the packages *bgmm* and *EMCluster* display more biased and noisy parameters compared to the other packages benchmarked and that their performance does not improve with a higher number of simulated data points. Finally, observations established from this higher dimensional framework strongly suggest that the rebmix initialisation method is not tailored for high-dimension in relation with to *Curse of Dimensionality in Distance Function*. Both behaviours are highlighted with several visualisations and summary metrics tables throughout additional section *Supplementary Figures and Tables in the HD simulation*, and notably by noting the same distributional pattern in summary Table 3a, displaying mixtures that are easy to deconvolve, and Table 3b, displaying highly-overlapping components.

About theoretical aspects, we have covered thoroughly the challenges induced by GMM estimation in a high-dimensional context and reviewed some strategies developed to overcome this issue, through parsimonious parametrisations of Gaussian mixtures (see Appendix section **Parsimonious parametrisation of multivariate GMMs**) or projection into a smaller subspace, enlarging the scope of application of *factorial analysis*, see Appendix *Parameters estimation in a high-dimensional context*.