Noname manuscript No.
(will be inserted by the editor)

Supplementary material. Composite likelihood methods for parsimonious model-based clustering of mixed-type data

Received: date / Accepted: date

1 A further comparison

As requested by an anonymous reviewer during the revision process, we conducted a further comparison between our model and the BD model included in the `clustMD` R package. However, since there is no theoretical detail about the data generation process or the estimation method used to get the parameter estimates, we were not able to make reasonable comments. We have just shown the numeric results without giving a possible explanation on what we have observed. We considered the same settings seen in the manuscript, apart from that one with a higher number of ordinal variables due to the high computational times of the BD model.
We focused our attention on the settings summarized in section 5.3 of the paper, that is 250 samples generated from the following design. The number of observations is $N = 500$, the number of groups is $G = 2, 3$ and the number of latent factors is $K = 4$. The number of variables is 8 of which 4 are ordinal, with 5 and 10 categories. The other model parameters are generated as described in the paper under the model $(SC)UU$. In this setting, we evaluate the effect of a higher number of categories for the ordinal variables. We compare the clustering performances, in terms of ARI, of the proposal with the naive case (i.e. our model treating all variables as they were continuous) and the models available in the package `clustMD`.
Looking at table 1, BD model shows performances in line with our proposal; in some cases it performs better. In the simpler settings (for example when $G = 2$ and the number of categories is equal to 5) it performs slightly better, even if we do not know if the model is still parsimonious as our proposal. However, in all cases the BD model seems to be slower than composite likelihood. Furthermore, the computational time increases with a higher number of groups and a higher number of categories. To analyze more in dept this aspect, we focused our attention on the simplest setting, i.e. when the number of groups is equal to 2 and the number of categories is equal to 5 (this is the case where the BD model performs better than composite likelihood). The computational time ratio is reported in table 2.

Address(es) of author(s) should be given

Table 1 Quartiles, mean and standard deviation (in brackets) of the ARI for the estimated partition obtained by different estimators over 250 samples generated from the model $(SC)UU$ with $G = 2, 3$, $K = 4$, $N = 500$, ordinal variables with 5 or 10 categories.

	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Composite	0.8327	0.9428	0.9834	0.8642 (0.1998)	0.7621	0.8591	0.9285	0.7949 (0.1892)
Naive	0.5431	0.8881	0.9362	0.7053 (0.3151)	0.4019	0.4806	0.7551	0.5583 (0.2375)
EII	0.7709	0.8888	0.9692	0.8435 (0.1579)	0.4761	0.6000	0.7357	0.6020 (0.1757)
VII	0.7583	0.8890	0.9673	0.8337 (0.1728)	0.4641	0.5865	0.7229	0.5963 (0.17749)
EEI	0.7148	0.8814	0.9753	0.7836 (0.2579)	0.4179	0.5976	0.7594	0.5955 (0.2104)
VEI	0.6515	0.8816	0.9754	0.7714 (0.2651)	0.4323	0.6064	0.7636	0.5944 (0.2161)
EVI	0.6688	0.8934	0.9754	0.7763 (0.2602)	0.4289	0.5860	0.7553	0.5911 (0.2124)
VVI	0.5356	0.8724	0.9754	0.7440 (0.2818)	0.4237	0.5888	0.7569	0.5874 (0.2126)
BD	0.9234	0.9756	1.0000	0.9080 (0.1701)	0.6914	0.8486	0.9262	0.7917 (0.1686)
	$G = 2$				$G = 3$			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
Composite	0.9196	0.9754	1.0000	0.8706 (0.2560)	0.7280	0.8480	0.9224	0.7788 (0.1977)
Naive	0.1168	0.6546	0.9837	0.5531 (0.4190)	0.2458	0.4294	0.6472	0.4611 (0.2486)
EII	0.7625	0.8889	0.9754	0.8433 (0.1572)	0.4051	0.5531	0.6613	0.5605 (0.1660)
VII	0.7499	0.9043	0.9751	0.8356 (0.1742)	0.4223	0.5441	0.7089	0.5528 (0.1833)
EEI	0.7122	0.8872	0.9754	0.7860 (0.2601)	0.3972	0.5768	0.7429	0.5629 (0.2038)
VEI	0.6151	0.8945	0.9755	0.7712 (0.2732)	0.3623	0.5872	0.7422	0.5567 (0.2211)
EVI	0.6716	0.8921	0.9757	0.7777 (0.2637)	0.3767	0.5212	0.7170	0.5455 (0.2099)
VVI	0.5269	0.8749	0.9757	0.7357 (0.2914)	0.3585	0.5461	0.7243	0.5399 (0.2195)
BD	0.9222	0.9754	1.0000	0.8670 (0.2471)	0.6040	0.8404	0.9364	0.7683 (0.2128)

Table 2 Quartiles, mean and standard deviation (in brackets) for the distribution of the ratios of the computational times by maximizing the composite likelihood and the BD model over 250 samples generated from the model $(SC)UU$ with $G = 2$, $K = 4$, $N = 500$; each ordinal variable has 5 categories.

	Computational Time Ratios			
	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$\bar{x}$ (std)
$\frac{\text{BD}}{\text{Composite}}$	2.5608	3.3571	4.0631	3.5626 (1.4716)

BD model is more than 3 times in median and in mean slower than the composite likelihood. It follows that the slightly lower clustering performance of the composite likelihood is well paid in terms of computational time.
For computational reasons, we do not report the results for the setting with 10 ordinal variables.