RESPONSE TO THE REVIEWER

We would like to thank the reviewer for the excellent report. We have addressed all the comments and suggestions, which have significantly improved the manuscript with respect to the original submission.

Here is a detailed response. The reviewer's text is in boldface, and our response is the indented text:

The paper describes the R package MoTBFs, which implements a family of Bayesian networks that uses mixtures of truncated basis functions. Compared to more classic distributional assumptions, this allows to include both discrete and continuous variables in the same network without restrictions on the arcs (as would be the case for conditional Gaussian Bayesian networks).

The paper covers the common use-case of having a mixed set of continuous and discrete variables to learn a Bayesian network from. When the aim is to learn a causal networks, restricting arc directions is problematic; the fact that MoTBFs has no such restriction makes it suitable for this kind of analysis, unlike the other packages referenced in the paper.

The point raised by the reviewer about causal networks is indeed relevant. We have mentioned it in the introduction of the revised manuscript.

My only general complaint is that the use of the English language could definitely be improved by some proof-reading from a native speaker. Currently the tone is too informal in places, and idioms and turns of phrase clearly come from a different language. Some examples are in the minor comments below. The English in the package documentation also needs improvement, as I have spotted several typos and syntactically-incorrect sentences in different manual pages.

We agree that the use of the English language could be significantly improved. We have thoroughly checked it both in the manuscript and in the package documentation. We have uploaded to CRAN a new version of the package (version 1.4) with the new documentation.

MAJOR COMMENTS

**** page 1: the authors reference various packages implementing similar Bayesian network models.***

However, they do not mention the abn package for additive Bayesian networks. Unlike other Bayesian network packages that are likewise not mentioned (pcalg, deal, etc.), abn allows modelling both discrete and continuous variables in the same network without any restrictions on the arcs by modelling local distributions as GLMs. In terms of features it is the closest to MoTBFs. Hence it should be discussed and compared to MoTBFs, possibly including it in the example as well.

Thank you for the comment. We were not aware of the abn package, which is certainly relevant. We have included a new paragraph in the introduction commenting on the difference with respect to the MoTBFs package. The main difference between both packages is the kind of models they support. MoTBFs do not belong to the exponential family, unlike the models supported by abn, and in that sense both packages are complementary. We have also cited pcalg and deal in the revised manuscript.

**** page 5: the authors mention that learningHC() is a modified version of the hc() function from bnlearn, without mentioning how it has been modified. A quick read of the R code for that function shows that it calls hc() with either score = "bic" or score = "loglik", but the logic of the code is not clear to me. I suggest that it should be better explained, both in the paper and in the package documentation.***

LearningHC() converts non-numeric columns in the dataset into factors before calling hc() in bnlearn. It can also be used to discretize the data before calling hc(), using equal width intervals, where the number of intervals is given as an argument. We have updated the documentation as well as the manuscript in order to clarify this.

**** page 7 and later: the summaries of the models are very difficult to read because of the large number of digits reported in the printouts, which make the summaries very long and the model formulas difficult to parse. I have the impression that keeping only the first 3-4 significant digicts could easily shave 30% of a page from the paper.***

We agree. We have kept only the 4 most significant digits in the revised version of the manuscript.

**** page 13: the authors show an example of how to include prior knowledge in the learning process by augmenting the data with new samples generated from generateNormalPriorData.***

I think the authors should provide some guidelines on how to choose the combination of (sample size, means, standard deviations) to pass to this function; I would imagine that the overall weight of the information in the prior compared to that in the sample will be determined by the interaction of these three quantities.

The natural way to incorporate prior knowledge in Bayesian inference is to define prior distributions on the parameters. However, in the case of MoTBF distributions, the parameters do not have a meaning in general. Therefore, there is no clear way in which a practitioner could provide prior information on any of the parameters, but still some information could be specified. For instance, assuming a random variable representing the body temperature of the patients in a given population, a practitioner could choose not to give prior information on any parameter, but instead provide the full distribution of the variable that he or she would consider if no data is available. For instance, such a prior

information could be that the distribution for the body temperature is normal with mean 37 and standard deviation 0.5. This is the approach followed by Pérez-Bernabé et al. (2016), where the prior knowledge is encoded as a distribution on the random variable which is encoded as an MoTBF that is later combined with the MoTBF density learnt from data. More precisely, the estimated density is the result of a linear combination between the density corresponding to the prior knowledge and the one estimated from the data. The coefficients of the linear combination are computed so that they reflect how accurately each density describes the observed data, in terms of likelihood. This is the procedure implemented in the MoTBFs package. We have explained this in more detail in the revised version of the manuscript, moving this discussion to pages 4 and 5, when we explain the procedure for incorporating prior knowledge.

**** page 15: incorporating prior information produces better parameter estimates when data are scarce, but only if the prior information is correct. I wonder 1) how much data are needed to make the contribution of the prior irrelevant and 2) if weakly and non-informative priors could also be of use?***

We have now specified in the revised version that the conclusion about the better parameter estimates was referred to the example in the paper. Regarding question 1) rather than the amount of data, the method described by Perez-Bernabé et al. (2016) determines the contribution of the prior information through the weight it is assigned in the linear pool where the density corresponding to the prior information and the density estimated from the data are combined. Each density receives a weight that is equal to the difference between its log-likelihood (i.e. the data log-likelihood computed using that density) and the expected log-likelihood of an MoTBF density whose parameters are selected at random. Regarding 2), that would be an interesting idea to explore but is a bit outside the theory of MoTBFs developed so far, and would probably require an important amount of research effort to explore, but it is indeed a way to proceed. In fact, the assignment of weights mentioned above, in some sense replicated the idea of measuring how far the prior information is from something that could be resembled as a non-informative prior, since the expected loglikelihood is computed with respect to MoTBFs whose parameters are chosen uniformly at random.

MINOR COMMENTS

**** page 1: "have previously been studied and covers" -> "have previously been studied and cover".***

Fixed.

**** page 4: "incorporating prior knowledge to the estimation process" -> "incorporating prior knowledge in the estimation process".***

Fixed.

**** page 4: "is the task consisting of computing" -> "consists in".***

Fixed.

**** page 5: "It can always be done since the sampling order is top-down" -> "Sampling is always possible since variables are sampled following the topological ordering of the network."***

Fixed.

**** page 5: "S3 oriented objects" -> "S3 objects".***

Fixed.

**** page 5: "The functions developed in the packages" -> "The functions provided by the packages".***

Fixed.

**** page 6: I assume "quadratic error" means "squared error"?***

Yes, we have fixed that.

**** page 8: "through a series of methods" -> "with a collection of methods".***

Fixed.

**** page 11: "The final piece of the process" -> "The last step is".***

Fixed.

**** page 15: there is no need to cite (Henrion, 1988) again as it was cited earlier in the paper.***

We have removed the citation.