



Rejoinder on: Hybrid semiparametric Bayesian networks

David Atienza¹ · Pedro Larrañaga¹ · Concha Bielza¹

Received: 22 May 2022 / Accepted: 23 May 2022 / Published online: 13 June 2022
© The Author(s) 2022

We thank all the discussants for their insightful remarks on our paper that broaden the possibilities of the new class of models we propose. Next we comment on the issues raised by them.

Marco Scutari suggests to add colors to the arcs that help define a single measure for the structural distance between two HSPBNs. Adding colors can give more expressiveness to the graph, although we think that shades and shapes that we use for the nodes (see Figure 1 in the paper) already capture the nature (discrete/continuous) and the conditional distribution type (Gaussian/KDE). Combining both, structural and type, Hamming distances into one measure may be interesting to simplify the performance output. However, by maintaining separate measures (as we also do with the likelihood, see Table 1), we can convey the specificities of each measure (goodness of fit, differential arcs, different conditional distribution types, etc.).

As Scutari points out, we agree on the need for updating the definition of equivalence classes for HSPBNs. This would be useful for building completed partially directed graphs, which is the output of constraint-based learning algorithms. Implementing those algorithms for HSPBNs is challenging, also raised by Scutari, because suitable conditional independence tests must be sought. They could be conditional mutual information-based tests for any possible combination of triplets of variables, including discrete, Gaussian or KDE. To the best of our knowledge, the state of the art only considers simpler scenarios. First, in semiparametric Bayesian networks (Atienza et al. 2022a), without discrete variables, we used nonparametric conditional independence tests, in particular, a permutation test based on the estimation of the mutual information

This rejoinder refers to the comments available at: <https://doi.org/10.1007/s11749-022-00815-0>, <https://doi.org/10.1007/s11749-022-00816-z>, <https://doi.org/10.1007/s11749-022-00817-y>, <https://doi.org/10.1007/s11749-022-00818-x>.

✉ David Atienza
datienza@fi.upm.es
Pedro Larrañaga
pedro.larranaga@fi.upm.es
Concha Bielza
mcbielza@fi.upm.es

¹ Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain

with K -nearest neighbors (CMIknn test) and a fast randomized conditional correlation test, version of the well-known kernel conditional independence test. Second, mutual information of discrete and continuous random variables, without any conditioning set, is computed with different methods, namely the nearest neighbor estimator, the kernel estimator and the orthogonal projection estimator (Beknazaryan et al. 2019).

Marco Scutari and Antonio Salmerón deal with the HSPBN constraint that continuous nodes cannot be parents of discrete nodes (also found in conditional linear Gaussian Bayesian networks) as a limitation of those models as causal models, since some arc directions are fixed and might not be in accordance with the cause–effect relationship. The solution of using a logistic regression or softmax functions to define the conditional model, mentioned by Salmerón, seems appropriate. Unfortunately, causal models are beyond the scope of our paper and requires further attention.

The complexity of learning HSPBNs may be an important issue, both in high-dimensional settings and/or in large sample problems, including streaming data. This concern has been raised by Salmerón. Eldest streaming data may be progressively deleted as new data arrive, as commonly done with sliding windows strategies. This involves discarding the corresponding terms in the summation of the KDE and adding the new terms of the incoming data, in an incremental and more local fashion. For static scenarios, our paper uses the `PyBNesian` library (Atienza et al. 2022b) for the experiments. This is an open-source Python package that implements KDEs with OpenCL to enable GPU acceleration and hence significant execution speeding up. Moreover, others authors have adapted KDEs to the streaming scenario (Kristan et al. 2011). As regards high dimensionality, the experiments in the paper can cope with tens of nodes and tens of thousands of instances.

Comparing HSPBNs against hybrid Bayesian networks with approximations of conditional probability densities based on mixtures of truncated basis functions (MoTBFs) (Langseth et al. 2012) (or mixtures of truncated exponentials and mixtures of polynomials, as particular cases) is interesting for future research, as Antonio Salmerón and Serafín Moral mention. Our previous results of inefficient evaluation time of KDEs when compared with mixtures of polynomials were restricted to supervised classification problems, and more specifically, to simple Bayesian classifiers as naive Bayes and tree augmented naive Bayes where conditional densities of predictor variables have at most two parents. We think this cannot be extrapolated to the more general HSPBNs, without any target class variable or restrictions on the number of parents. However, MoTBFs are closed under multiplication, addition and integration and this is an advantage over HSPBNs for probabilistic inference, that we circumvent with a sampling procedure from the KDEs. As highlighted by Moral, better sampling procedures when evidence (observed variables) is available would include likelihood weighting and Markov chain Monte Carlo methods.

Direct estimation of the joint conditional density, suggested by Moral, rather than using a quotient of the full joint density and the joint density of the conditioning variables (Equation (9)), as evidenced also in the expression we obtain for sampling (Equation (21)), is really appealing. The double kernel estimator, also known as the Nadaraya–Watson conditional density estimator, was accelerated in Holmes et al. (2007) to deal with conditional densities over multivariate variables, which was confined to bivariate variables so far, being an option in our context. Also, the density

ratio estimation (Sugiyama et al. 2012) seems an interesting alternative to explore in the future.

Our design of model selection via cross-validation with an extra validation set has been motivated to achieve a tradeoff between computational cost and overfitting avoidance. Moral comments on alternative schemes include repeated k -fold cross-validation or leaving one out in the training data set. In addition, other honest methods, as bootstrap, bolstered or jackknife estimations can be used for the training set, depending on its size and the statistical properties (mainly bias and variance) intended to be held by the score estimator. Furthermore, the partition training/validation might be also repeated, improving the stability of the learning process.

Stefan Sperlich comments on having a generalized LG, allowing interactions of continuous parents in the mean of the Gaussian distribution, to relax CLGs but being less flexible than KDEs (a compromise between these two extremes). This raises many problematic issues: how the graph would encode those interactions via the g_j functions is not clear, how to design a structure learning algorithm (apart from the parameter learning methods, for β 's and g_j parameters), how to determine s and g 's and how to develop (exact or approximate) inference approaches for these new CgLG densities. Furthermore, we still assume Gaussianity, whose violation in the data has been the motivation of our HSPBN models.

We were not aware of the smooth handling of discrete variables suggested by Sperlich, which requires further inspection in the future. Getting rid of conditioning on each value of the discrete variables is certainly interesting especially when each resulting subsample is not sufficiently large, although Bayesian estimation is typically used to correct for small observed samples. Anyway treating discrete variables as continuous hinders model interpretability.

While it is true that the Gaussian kernel is not the most asymptotically efficient kernel, it inherits many interesting theoretical properties from the Gaussian distribution. For example, Equation (21) can be calculated with a closed-form expression because the conditional distributions of a Gaussian are known. Also, some bandwidth estimation procedures (such as biased cross-validation and plug-in) require computing derivatives, and we know that the Gaussian is infinitely derivable (Chacón and Duong 2015); hence, the use of Gaussian kernels can be suitable.

In Section 3.5, we describe an analysis of the computational asymptotic complexity of the proposal. This section could have been alternatively entitled “Asymptotic Complexity Analysis” to avoid confusions with the asymptotic statistical theory, as in the case of Sperlich. The asymptotic statistical properties are beyond the scope of this paper. However, much work has been done on the statistical behavior of KDE, for example about its consistency (Wied and Weißbach 2012). A discussion about consistency is also found in interesting works on Bayesian network classifiers (John and Langley 1995; Pérez et al. 2009). Unlike these works, our proposal can make parametric assumptions, and studying how the consistency (and other statistical properties) of the joint distribution estimator changes with the mix of parametric and nonparametric distributions remains as an open issue.

Acknowledgements We thank Ana M. Aguilera for her kind invitation to submit our paper to TEST. We also greatly appreciate the suggestions of two anonymous reviewers, which helped to improve the paper.

This work has been partially supported by the Ministry of Education, Culture and Sport through the grant FPU16/00921, by the Spanish Ministry of Science and Innovation through the PID2019-109247GB-I00 and RTC2019-006871-7 projects, and by the BBVA Foundation (2019 Call) through the “Score-based nonstationary temporal Bayesian networks. Applications in climate and neuroscience” (BAYES-CLIMA-NEURO) project.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Atienza D, Bielza C, Larrañaga P (2022) Semiparametric Bayesian networks. *Inf Sci* 584:564–582
- Atienza D, Bielza C, Larrañaga P (2022b) `PyBNesian`: an extensible Python package for Bayesian networks. *Neurocomputing* Under review
- Beknazaryan A, Dang X, Sang H (2019) On mutual information estimation for mixed-pair random variables. *Stat Probab Lett* 148:9–16
- Chacón JE, Duong T (2015) Efficient recursive algorithms for functionals based on higher order derivatives of the multivariate Gaussian density. *Stat Comput* 25(5):959–974
- Holmes MP, Gray AG, Isbell CL (2007) Fast nonparametric conditional density estimation. In: *Proceedings of the twenty-third conference on uncertainty in artificial intelligence*. AUAI Press, pp 175–182
- John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the eleventh conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers, pp 338–345
- Kristan M, Leonardis A, Skočaj D (2011) Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recogn* 44(10–11):2630–2642
- Langseth H, Nielsen TD, Rumí R, Salmerón A (2012) Mixtures of truncated basis functions. *Int J Approx Reason* 53(2):212–227
- Pérez A, Larrañaga P, Inza I (2009) Bayesian classifiers based on kernel density estimation: flexible classifiers. *Int J Approx Reason* 50(2):341–362
- Sugiyama M, Suzuki T, Kanamori T (2012) *Density ratio estimation in machine learning*. Cambridge University Press, Cambridge
- Wied D, Weißbach R (2012) Consistency of the kernel density estimator: a survey. *Stat Pap* 53(1):1–21

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.