Erratum

Date: 5 August 2021

Multivariate Density Estimation by Neural Networks

After publication of an earlier version of this paper, we received feedback that there were several incorrect references to related methods in the literature. These errors are corrected in the current version. We are grateful to professor E. Trentin from University of Siena for his valuable feedback.

Introduction section:

Page 4: original text in the footnote
We note that it is also possible to estimate the PDF by neural networks straight away, see Trentin et al. (2018). However, it is more natural to construct labels for the CDF values instead for PDF values . In addition, this is less sensitive to statistical error, since the integral to construct a PDF from a histogram can be regarded as a regularizer, as Magdon-Ismail and Atiya (2002) argue amongst other reasons.

Page 4: new text
We note that it is also possible to estimate the PDF by neural networks straight away, see Trentin et al. (2018). However, it is more natural to construct labels for the CDF values instead for PDF values, since these values range from 0 to 1 and are monotonic. More importantly, constructing PDF values is more prone to statistical error than CDF values as PDF values are sensitive to the bandwidth while CDF values do not require bandwidth selection.

This footnote on page 4 is altered and placed in main text. Moreover, the text has been changed to explain more clearly the advantage of constructing CDF labels over PDF labels.

Related Literature section:

Page 6: original text
Disadvantages of the KDE are the poor performance on high dimensions and the high dependence on the specification of bandwidth.

Page 6: new text
Disadvantages of the KDE are the expensive and cumbersome computation on high dimensions and the high dependence on the specification of bandwidth, see Bishop (2006).

Page 6: original text in the footnote

Rules of thumb are created for this inconvenience, such as the Silverman's rule, see Silverman (2018). However, these rules are usually only implied for univariate cases.

This footnote on page 6 is removed, since there are rules of thumbs for bandwidth specification on higher dimensions (of which we did not know of before).

Page 7: original text
It attempts to obtain the best of both worlds: no assumptions are made a priori and the extension to multiple dimensions is straightforward without contemplating with issues as memory and time.

Page 7: new text
It attempts to obtain the best of both worlds: no assumptions are made a priori and their extension to multiple dimensions is theoretically straightforward, even though computational issues such as time and memory still need to be considered.

Page 7: original text
Specifically, Multilayer Perceptrons (MLPs) are used, feeding forward from the input layer to the output layer 6) .

Page 7: new text
Specifically, Multilayer Perceptrons (MLPs) are used, feeding forward from the input layer to the output layer 4), see (Bishop et al., 1995, Chapter 4)

Page 7: original text
There are two distinct methods within this approach. Trentin et al. (2018) directly estimate the PDF by neural networks. The second method is to estimate the CDF by neural networks and from this to derive the PDF as Magdon-Ismail and Atiya (2002) and Zhang (2018) do. This paper follows the non-parametric estimation method of Magdon-Ismail and Atiya (2002). Zhang (2018) uses a similar technique. He however, argues that also the PDF of non-smooth distributions, next to continuous distributions, can be estimated by imposing a different activation function. He also shows the analytical derivative of a simple case. The proposed method extends this literature by showing the analytical derivatives of any model. That is, the MLP can consist of as many hidden layers as desired, as opposed to Magdon-Ismail and Atiya (2002), Zhang (2018) and Trentin et al. (2018), who use only one hidden layer. Next to the number of hidden layers, the number of hidden neurons complete the specification of the MLP or in other words the model. Model selection can be executed in several ways such as random search or a certain MLP can be chosen at random. Magdon-Ismail and Atiya (2002), Trentin et al. (2018) and Zhang (2018) fix the number of hidden layers to 1, such that only the number of hidden neurons need to be specified. Note however, when applying a nonlinear activation function, using more hidden layers increases the ability to estimate a highly nonlinear distribution of the underlying DGP. Magdon-Ismail and Atiya (2002) and Zhang (2018) do not explain why they use a certain structure. Trentin et al. (2018) train a network several times with a different number of neurons, then they pick one amongst these according to a log-likelihood criterion. This slows down the process of model selection heavily though. In this paper a novel model selection procedure, which may be computationally more efficient is proposed as an alternative. 5) Also referred to as neural networks. 6) When referring to the neural networks model itself, either the term model or MLP is used. CBS | Discussion paper | May 12, 2021 7By using neural networks, the input variables need to be labeled such that it can be used for training. Trentin et al. (2018) construct these labels like the KDE does with some alterations on bandwidth selection and bias prevention. They show that their method, called

Parzen Neural Networks (PNN), is less sensitive to bandwidth specification than KDE is. In addition, they overcome other shortcomings of KDE such as memory issues. However, it is more natural to construct labels for the CDF values instead for PDF values. More importantly, this is less sensitive to statistical error, since the integral to construct a PDF from a histogram can be regarded as a regularizer, as Magdon-Ismail and Atiya (2002) argue amongst other reasons. Moreover, they show density estimation in the multivariate setting as well. Note that as more variables N are used, the more hyper parameters need to be specified for the bandwidth for Trentin et al. (2018)'s approach, the $N \times N$ matrix H and the initial bandwidth. However, as this bandwidth needs to be selected empirically, choosing this value in a multivariate setting is even more difficult. More importantly, Trentin et al. (2018) assume that the underlying N variables need to be independent. The method proposed for density estimation in this paper, does not imply any assumptions on these relationships. This paper extends the literature of CDF estimation by neural networks. Instead of numerically deriving the PDF from the obtained CDF, analytical derivations are shown for any specified model. That is, the MLP can consist of as many hidden layers as desired, contrary to Magdon-Ismail and Atiya (2002), Zhang (2018) and Trentin et al. (2018). A novel approach for model selection is shown which is computational time efficient opposed to Trentin et al. (2018). Moreover, the PDF estimation works for any correlated variables in the multivariate setting. Trentin et al. (2018) also performs multivariate density estimation, conditioned that the variables are independent though.

Page 7: new text

There are two distinct methods within this approach. Trentin et al. (2018) directly estimate the PDF by neural networks. The second method is to estimate the CDF by neural networks and from this to derive the PDF as Magdon-Ismail and Atiya (2002) and Zhang (2018) do. By using neural networks, the input variables need to be labeled such that it can be used for training. Trentin et al. (2018) construct these labels like the KDE does with some alterations on bandwidth selection and bias prevention. They show that their method, called Parzen Neural Networks (PNN), is less sensitive to bandwidth specification than KDE is. However, it is more natural to construct labels for the CDF values instead for PDF values, since these values range from 0 to 1 and are monotonic. More importantly, constructing PDF values is more prone to statistical error than CDF values as PDF values are sensitive to the bandwidth while CDF values do not require bandwidth selection. This paper follows the non-parametric estimation method of Magdon-Ismail and Atiya (2002). Zhang (2018) uses a similar technique. He however, argues that also the PDF of non-smooth distributions, next to continuous distributions, can be estimated by imposing a different activation function. He also shows the analytical derivatives of a neural network with a single hidden layer. The proposed method extends this literature by showing the analytical derivatives of any model. That is, the MLP or model can consist of as many hidden layers and hidden neurons as desired. This is useful since when applying a nonlinear activation function, using more hidden layers instead of only one increases the ability to estimate a complex distribution of the underlying DGP. The proposed method also differs from the above mentioned literature in four other 3) Also referred to as neural networks. 4) When referring to the neural networks model itself, either the term model or MLP is used. CBS | Discussion paper | July 30, 2021 7ways. First, we choose the optimal MLP structure from a set of MLPs empirically in one estimation. This is different from Magdon-Ismail and Atiya (2002) and Zhang (2018) which use a pre-specified MLP structure, and computationally more efficient from Trentin et al. (2018) who train a network several times with a different number of neurons. Second, unlike Magdon-Ismail and Atiya (2002) and Trentin et al. (2018) the proposed method can be applied to discrete distributions. Third, our method does not require to select an initial bandwidth empirically as opposed to Trentin et al. (2018). We conjecture that choosing this bandwidth can be cumbersome especially in a multivariate setting, thus highlight this as an advantage of our method. Fourth, as an extension to the literature, such as Trentin et al. (2018), we explicitly consider dependent input variables in our simulation cases. We show our method deals with dependency as such by the inter-connectivity of the neural network. Hence ANNs are able to estimate the PDF regardless its underlying DGP and whether it is correlated or not. This paper extends the literature of CDF estimation by neural networks. Instead of numerically deriving the PDF from the obtained CDF, analytical derivations are shown for any specified model. A novel approach for model selection is shown which is computational time efficient opposed to Trentin et al. (2018). Furthermore, the PDF estimation can be applied to continuous as well as discrete distributions. Moreover, the PDF estimation works for any correlated variables in the multivariate setting.

Results section:

The following sentence is removed as the PNN is able to be applied to dependent variables.

Page 35: original text
For these simulations, we only compare our results to KDEs, as the other baseline method Trentin et al. (2018) is not applicable to correlated multivariate data.

The following sentence is added to page 35.

Page 35: new text
It is also possible to do this by using sigmoids with adaptive amplitude on the output layer as is described in Trentin (2001).