Annals of the Institute of Statistical Mathematics (2024) 76:43–46
https://doi.org/10.1007/s10463-023-00887-1

INVITED ARTICLE: FOURTH AKAIKE MEMORIAL LECTURE

Rejoinder of "Identifiability of latent-variable and structural-equation models: from linear to nonlinear"

Aapo Hyvärinen[1]

Received: 6 September 2023 / Accepted: 15 September 2023 / Published online: 1 November 2023
© The Institute of Statistical Mathematics, Tokyo 2023

I'm very grateful that Dr. Morioka and Dr. Matsuda have kindly discussed my memorial lecture, making very interesting points about the history of the topic, its connections to other research, and also proposing explicit discussion points. In the following, I will make some comments on their Discussion.

In his first discussion point, Dr. Morioka asks about whether, and how, ICA might be implemented in the brain. This is of course a very deep and interesting question. I tend to approach it from the viewpoint of separating different levels of modelling the brain. David Marr proposed a well-known division to three levels, but here I would like to consider levels that have clear counterparts in statistical theory and machine learning, as I have done previously (Hyvärinen et al., Ch. 18). On a very abstract level, we have the "statistical modelling", i.e. specifying a statistical model to be estimated. After (or below) that, there is the "objective function" level where one specifies an objective function for the estimation, or a similar concrete goal of computation. Then, there is the "algorithmic" level that specifies a computational algorithm, typically to optimize the objective function.

Suppose we define the model as nonlinear ICA, the objective as the likelihood, and the algorithm as gradient ascent. Thus, we see all the three levels. This is of course not a universal division, and there are a lot of details on how the algorithm is implemented in a computer hardware, but let it suffice for this discussion.

Now, I would be optimistic about modelling the brain on the statistical and objective levels. Regarding linear ICA, there is a large body of work that has looked at the features learned by linear ICA and compared them to the features of the primary visual cortex (Olshausen and Field 1996; van Hateren and van der Schaaf 1998; Hyvärinen et al. 2009). There seems to be quite a good match at least on the level of individual basis vectors. Going to the nonlinear domain, one can train a

The Related Articles are https://doi.org/10.1007/s10463-023-00884-4; https://doi.org/10.1007/s10463-023-00885-3; https://doi.org/10.1007/s10463-023-00886-2.

✉ Aapo Hyvärinen
aapo.hyvarinen@helsinki.fi

[1] Department of Computer Science, University of Helsinki, Pietari Kalmin katu 5, 00560 Helsinki, Finland

deep neural network based on different statistical criteria and/or objective functions and compare the function of the resulting network with measurements from human or animal brains, or experiments related to behaviour of humans or animals. In particular, the internal representations learned in the hidden units, as well as the final outputs of the networks can be correlated with measurements of such neuroscience experiments. In fact, a number of attempts have been made to do this regarding the visual system of the brain (Zhuang et al. 2021). Such studies have used objective functions similar to nonlinear ICA, and I hope, such studies will be conducted in the future precisely using nonlinear ICA as well. While it is a bit too early to say anything definitive, those studies seem to indicate that some results of unsupervised deep learning (although usually not related to rigorous statistical modelling) seem to have meaningful similarities with the computations in the brain.

On the other hand, I'm afraid modelling the algorithmic level in the brain is much more challenging. It is very difficult to measure the computations of single neurons, let alone groups of neurons, and measuring their learning (plasticity) is perhaps the most difficult of all. So, there is less hope, at least in light of current neuroscience, that we could to say very much about the algorithmic level. One thing that is clear is that it seems to be very different from what a digital computer would do. Yet, the work mentioned by Dr. Morioka in his Discussion is courageously attempting to investigate how a linear ICA algorithm could be a realistic model on the "algorithmic" level and even on a further level of physical implementation. I hope future research will shed more light on how realistic such models may be, and whether algorithms performing nonlinear ICA could be modelled in the same way.

In his second discussion point, Dr. Morioka points out the possibility of reverse engineering the brain to find better, in particular robust algorithms for nonlinear ICA. This is a very fascinating possibility. But personally, I have a bit the impression that machine learning has already incorporated a lot of such knowledge from neuroscience, ever since the pioneering neural network models of the 1960s and 1970s. It should be noted that even some of the earliest work on ICA by Jutten, Hérault, and Ans in the 1980s was about modelling neural processing; the same is true about the original work on sparse coding or dictionary learning by Olshausen and Field (1996). But now, in 2023, it is not clear to me if a lot of new insights can still be obtained from that direction, especially given the painstakingly slow progress in neuroscience. Admittedly, this may be only due to lack of imagination from my part. Nevertheless, I would be more optimistic regarding some other learning paradigms, in particular reinforcement learning, which is highly complex and is still constantly absorbing new ideas of how humans and other biological organisms learn and behave. Obviously, the information flow could go in the other direction as well: AI can teach us something about the human brain or mind, which was the topic of my recent book (Hyvärinen 2022).

Dr. Matsuda, in his Discussion, points out the interpretation of nonlinear ICA as defining an exponential model, where the sufficient statistics are given by some point-wise functions of the independent components. I agree that this is a very fundamental interpretation which, I think, applies to a large variety of neural network learning paradigms, and has great potential. Related to this, in our work on nonlinear ICA, we have actually sometimes made the simplifying assumption of an

exponential family model of the components (Hyvärinen and Morioka 2016; Khemakhemet al. 2020a) and sometimes not (Hyvärin and Morioka 2017; Hälvä et al. 2021). Using the exponential family already in the model specification tends to make the identifiability conditions much simpler, e.g. reducing them to a full-rank condition of a matrix collecting the parameters of the exponential family. On the other hand, it can be seen as restricting the generality of the model, and our most general results for time series are given without any such assumption (Hälvä et al. 2021). Khemakhemet al. (2020b) propose a kind of generalization of exponential families in the context of nonlinear ICA. Whether such an assumption is made in the model or not, it is always possible to take the independent components and define an exponential family using the components. Indeed, Matsuda and Hyvärinen (2019) proposed an application of such a framework in clustering, and many different applications in such "transfer" learning are conceivable.

Dr. Matsuda further discusses noise-contrastive estimation and bridge sampling as examples of "learning by classification". These are instances of the class of methods called "self-supervised learning", which are currently extremely popular in machine learning. The basic idea is that when given just a multidimensional data vector $\mathbf{x}$, one defines an artificial classification or regression problem based on prior knowledge of the data domain. Typically, this is done in the context of images, which of course have a rich structure. One can, for example, pretend that some pixels are missing, and learn to predict them from the surroundings. The hope is that if this is done with a deep neural network, the neural network will learn an internal representation of the data that is useful for some other real tasks. The reason why such self-supervised approaches are popular is that no new algorithms need to be developed, since well-known regression algorithms can be readily used.

Typical self-supervised learning is very heuristic and often has no theoretical basis, statistical, or otherwise. However, some methods can be given a solid justification by statistical theory. Among such theoretically principled self-supervised learning, I would argue the main paradigms are based on either classification or autoencoders. Classification is indeed the basis of noise-contrastive estimation (NCE), which can be shown to be able to estimate a statistical parametric model, and even if it is unnormalized. Its connection to bridge sampling has been considered very recently by Chehab et al. (2023). Autoencoders have a long history going back to the 1980s, where they have been used for nonlinear dimension reduction, i.e. nonlinear PCA, while a probabilistic version was more recently proposed as the variational autoencoder (Kingma and Welling 2014).

Going back to the topic of nonlinear ICA, indeed, many of our estimation algorithms are self-supervised, as reviewed in our companion review paper (Hyvärinen et al. 2023). Crucially, these algorithms are not purely heuristic; they implement estimators of the statistical nonlinear ICA models, and they are usually shown to be consistent (even if not statistically efficient). However, we have also proposed several estimators based on maximum likelihood (Khemakhemet et al. 2020a; Hälvä et al. 2021; Gresele et al. 2020). Obviously, a statistical model and its estimator are two different things, and a single model can have many estimators, but in the self-supervised learning literature, this distinction is not always very clear.

To conclude this rejoinder, I would argue that nonlinear ICA is a complicated statistical model in many ways: identifiability is not guaranteed, estimation may need some

other objective function than likelihood to be practical, and computation is always a problem. Indeed, the brain may have solved these problems during millions of year of evolution, but we are trying to do that in a matter of years, which is a challenge.

References

Chehab, O., Hyvärinen, A., Risteski, A. (2023). Provable benefits of annealing for estimating normalizing constants. Submitted manuscript.

Gresele, L., Fissore, G., Javaloy, A., Schölkopf, B., and Hyvärinen, A. (2020). Relative gradient optimization of the Jacobian term in unsupervised deep learning. *Advances in Neural Information Processing Systems (NeurIPS2020),* Virtual.

Hälvä, H., Corff, S. L., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., Hyvärinen, A. (2021). Disentangling identifiable features from noisy data with structured nonlinear ICA. *Advances in Neural Information Processing Systems (NeurIPS2021), Virtual*.

Hyvärinen, A. (2022). Painful intelligence: What AI can tell us about human suffering. arXiv:2205.15409.

Hyvärinen, A. Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems (NIPS2016)*, Barcelona, Spain.

Hyvärinen, A. Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS2017)*, Fort Lauderdale, Florida.

Hyvärinen, A., Hurri, A. J., Hoyer, P. O. (2009). *Natural Image Statistics*. Springer-Verlag.

Hyvärinen, A., Khemakhem, I., Morioka, H. (2023). Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns* (in press).

Khemakhem, I., Kingma, D. P., Monti, R. P., Hyvärinen, A. (2020a). Variational autoencoders and nonlinear ICA: A unifying framework. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS2020)*.

Khemakhem, I., Monti, R. P., Kingma, D. P., Hyvärinen, A. (2020b). ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems (NeurIPS2020), Virtual*.

Kingma, D. P. Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the International Conference on Learning Representations (ICLR2014)*, Banff, Canada.

Matsuda, T. Hyvärinen, A. (2019). Estimation of non-normalized mixture models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS2019)*, Okinawa, Japan.

Olshausen, B. A. Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607–609.

van Hateren, J. H. van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society, Series. B*, 265, 359–366.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), e2014196118.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.