Nonlin. Processes Geophys. Discuss., referee comment RC2
https://doi.org/10.5194/npg-2021-2-RC2, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on npg-2021-2

Maarten Ambaum (Referee)

Referee comment on "The blessing of dimensionality for the analysis of climate data" by Bo Christiansen, Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2021-2-RC2, 2021

This paper has a strong didactic focus; much of it is a review of convergence theorems and properties when a high number of independent variables are available, i.e. high dimensionality. A lot of it is reasonably well-known, but the way it has been put together in this paper seems quite valuable. I am particularly pleased to see that basic ideas that existed for a long time in statistical mechanics (thermodynamic limits, essentially) are here being highlighted as potentially applicable to climate data. The main result seems to be that a selection of climate data is shown to be behaving as if it is drawn from a (moderately) high dimensional space. This has important repercussions for much literature that has not been highlighted in this paper, in my opinion, particularly around regime behaviour in geophysical data - NB: I am not suggesting you should now include such a discussion, but it may be something to think about as a further application.

In my view the paper is well written and can be published pretty much as is, except that I invite the author to address a few minor points/questions first. I describe them in order of apperance, below.

l.85 & l.92: "a constant" I know this phraseology is used in related literature, but I do not like it very much. I would have prefered to explicitly say something like "... a constant for differnt realizations of the sampling process," or something similar. Essentially this is a frequentists statistical argument, suggesting there is a fixed distribution mean (the "constant") which can be approximated by a sample mean.

l.116: "sub-Gaussian" I am happy with this word, but it may be useful to include a one-sentence definition of it (I am not sure how widely known the word is.)

Section 3: the discussion of effective dimensionality is a well-worn topic in geophysics, and it remains an important topic. It reminds me of a paper I wrote some years ago (doi:10.1175/2007jas2298.1) on how suggested mulimodality in a wave amplitude index for the atmosphertic circulation possibly is a statistical fluke: it hinges on exactly the high dimensionality argument you are discussing here, but interpreted slightly differently. In effect, standard statsitics (moments) from a high dimensional data set are always expected to exhibit these "blessing of dimensionality" properties, and more fancy, non-linear properties, such as multimodality, are more-or-less by definition excluded. I think this is an important application of the high dimensionality property, and I wonder whether you care to comment on it.

Section 4.2: I thought this section, despite its simplicity, was really thoughtprovoking. I tried to interpret this in light of the well-known "signal-to-noise paradox" (https://doi.org/10.1038/s41612-018-0038-4), as it seems to be germane to that problem. Does the author agree that his discsusion here may shed light in the signal-to-noise problem? It would be a rather important application.

l.333 and Abstract: Perhaps I did not catch it but the dimensionality of 25-100 seems to fall a little bit from thin air. Can you please highlight where this estimate is based on? Furthermore: Figs 2 & 3 show empirical distributions of \phi, which have a given shape for independent data (Eq. 5). I would have thought that you can fit a Gaussian to those distributions and thus estimate a value of N. Did you do this? Does it give you the 25-100 estimate?