Atmos. Meas. Tech. Discuss., 6, C381–C384, 2013
www.atmos-meas-tech-discuss.net/6/C381/2013/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Dimensionality reduction in Bayesian estimation algorithms" *by* G. W. Petty

F. Aires (Referee)

filipe.aires@lmd.jussieu.fr

Received and published: 19 March 2013

Review of paper "Dimensionalyty reduction in Bayesian estimation algorithms" By G.W. Petty

This paper proposes a pre-processing step to reduce the dimension of the input data before an empirical Bayesian retrieval (focussing on the retrival of rain from MW observations). This is tested on synthetic data. The pre-processing is based on two PCA (one in the non-raining data, the other on "anomalies"). The impact of the pre-processing is very important in this database for the forthcoming retrieval step.

General comments:

* This topic is very important because it can have important practical consequences for the retrieval of precipitation for example.

* The experiments are conducted on a synthetic dataset built on purpose for this experiment. I understand that some techniques can be much better illustrated in synthetic cases, but it is frustrating to not see the impact of such a technique on real data considering that the authors has the possibility to easily make test on real data.

* What is the content of the cited paper "Petty and Li" (in review)? Isn't it applying this technique to real data? Is there a true interest in this case in having a new paper on data that are so synthetic?

* It is clear that reducing the size of the input data before the retrieval can have a strong impact: it reduces the number of parameters in the retrieval scheme which regularizes the inverse problem, it can extract important features and suppress part of the noise. However, the paper doesn't do a good job in citing the other approaches and even compare to them. There is a lot of litterature on this subject: - PCA-regression (PCA before a regression), - PCA as a pre-processing of a NN network, - projected-PCA to perform a PCA to facilitate the retrieval of a particular quantity, - "intelligent" distances (such as the Mahalanobis distance) before the Bayesian search step in the retrieval - ... Since you are really focusing in this paper on the technical aspects of PCA, I believe that you really need to spend more time comparing with the existing approaches otherwise the reader cannot judge on the potential of your method. For example, work has been done using a PCA to compress and extract the cloudy signal on IASI data. I believe that the PCA was used to cloud-clear the data (dimension of data is 8461 in this case, the number of channels in IASI). I believe that this type of experiments is very close to what you are doing so a discussion, even comparison would be interesting.

* I find the synthetic database a disturbing factor. First, you should make a 3D figure with the 3D ellipsoid of the inputs space, and then represent the direction of the rain signature. Maybe a 3D plot of the raining and non-raining data. Second, the approach you describe here is based on the fact that you can extract linearly a raining signature from the input data. This is true in your synthetic case, and therefore the technique

as to work. I have two questions: (1) Is it true that the raining signature is just a linear signature in the input data in real cases? (2) Would your approach work if this is not true? (3) Since your raining signature is linear in the input data, would a linear regression work as fine a your approach? A test on (3) would be very easy and is necessary here I believe.

* (1) The study on the minimal distance for the MC integration doesn't seem to include any uncertainty in the input parameters. When you perform these test, I think that adding a random error in the inputs of the data should bring a more realistic behavior. With this input noise, you optimal distance should behave slightly differently, especially when tested in the training dataset. (2) Using a small distance gives your a un-biased estimator but with high variance, using a higher distance provides you a biased estimator but with lower variance (this is the bias-variance dilemma). Such a discussion can be find for example in k-Nearest Neighboors discussion, varying the k like you vary the minimal distance.

* I find the discussion about the stage 2 not detailed enough. I don't understand exactly what is been done here. I believe that you could illustrate and explain what this stage 2 does geometrically.

Specific comments:

* page 2330, Eq. (1): I find the description of the "empirical" Bayesian scheme not good enough for readers not familiarized with this technique. Eq. (1) is the traditional Bayes formula. If we had information on the involved PDF, the expression could be used analytically, but this is not the case in a lot of cases. As a consequence, the integration of the Eq. (1) is performed empirically with the search and weighting procedure the authors describes, in a Monte-Carlo way. I believe you could improve this part and that this would be beneficial for new-comers in this field.

* page 2330, line 20: supress "only"

* page 2330, line 25: a way to accelerate this search is to order in some way the dataset so that the search is more efficient.

* Page 2331, line 4: Could you explain why RTE simulations cannot preserve the physical correlation structure among the TBs?

* Page 2331, line 10-15: I don't understand the numbers you provide (10^4 and 10^12). From where they come from?

* Page 2334, line 4: "as as".

* Page 2335, Eq. (4): please, descrive better the MC integration of Eq. (1).

* Fig 1: It seems that there is no difference on distribution between rainy and non rainy data. This cannot be true otherwise the retrieval couldn't work. Could you try to find a representation (maybe 3D) that could illustrate the distribution of non-rainy points and the raining signature?

Interactive comment on Atmos. Meas. Tech. Discuss., 6, 2327, 2013.