Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2017-52-RC1, 2017
© Author(s) 2017. This work is distributed under
the Creative Commons Attribution 4.0 License.

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Interactive comment on "Feature-based data assimilation in geophysics" *by* Matthias Morzfeld et al.

Anonymous Referee #1

Received and published: 30 November 2017

GENERAL COMMENTS

This paper is about parameter estimation problems in data assimilation. Conventional data assimilation techniques attempt to estimate parameters of a dynamical model by matching the model output to observations of the system. In some situations this is infeasible. The authors give the examples of a low-dimensional model of a complex system and a chaotic model over a long time-scale. In these situations the observations contain more information than can be matched. The authors show how (at least in some cases) this problem can be solved by extracting low-dimensional features from the observations and trying to match the model to these instead. A potential obstacle to this approach is the need to pass the observation noise model through a nonlinear mapping from observation space to feature space. The authors circumvent this obsta-

Printer-friendly version

Discussion paper

CC BY

cle by the ingenious expedient of noting that observation noise models are often based on ad hoc additive Gaussian assumptions with little physical basis, and we are just as entitled to model the feature noise directly using such assumptions instead of trying to convert observation noise to feature noise.

The techniques described by the authors deserve to be better known. However, there are a number of ambiguities and minor errors in the exposition and experimental methodology. These should be corrected before publication in NPG.

SPECIFIC COMMENTS

p1, L24. The 'or vice versa' part of the sentence implies that data assimilation may be infeasible when the data have a lower dimension than the model. I don't understand how this could be. Indeed, this is one of the situations in which it has just been said (L22-23) that conventional data assimilation is not required.

p4, L2; p6, L27. When the likelihood is expressed in terms of the noise pdf, the noise pdf should be conditioned on \theta too. Thus the first case should be written l(z|\theta) = p_\epsilon(z-\mathcal(M)(\theta)|\theta).

p4, L4-5. I dispute that all types of prior information can be represented by a probability distribution. In particular, I disagree that knowledge of just lower or upper bounds (or both) can be so represented. However, I agree that in some cases prior information can be represented by a probability distribution, and in those cases the framework of the paper makes sense.

Following on from the previous point, the paper lacks justification for its choice of priors. This doesn't matter too much in the artificial examples (Examples 1 and 4), but in the examples with real data (Examples 2 and 3) the priors need to be justified if the results are to tell us anything about reality.

p4, L30-31. On my first reading I didn't understand how F(\theta) could be a random variable. It would help the reader to include an brief explanation of how this can arise.

CC BY

p7, L7-8. R_f is not the sample covariance unless f is the sample mean.

SPECIFIC COMMENTS ON EXAMPLE 1

There is no statement about the initial conditions of the experiment. What were they?

p10. There is inconsistency over whether there is a data point at time 0. The formulae at L11 and L13 imply that there is not, as does L9 in referring to M as the number of data points. On the other hand, a data point at time 0 is shown in L12. On p11, Figure 1 shows a data point at time 0, but L1 of the caption implies that there is not one.

p10, L11. Is the variance of v_i really 1? It looks much smaller in Figure 1.

p11, Figure 1, caption L1-2. The curves plotted are trajectories, not samples from the prior distribution of the parameters p(\theta|z). This affects L4 too.

p11, L9. What is varied randomly in each experiment? The parameters? The initial conditions? Observation noise? Anything else?

p11, L9 says there were 100 experiments, but the caption of Figure 1 says there were 1000.

Figure 1 shows the mean of the KL divergence, but what about the spread? Conclusions such as p12, L2-3 would be ruined if the spread were much larger at M=100 than at M=500.

SPECIFIC COMMENTS ON EXAMPLE 2

As well as the aforementioned need to justify the prior, the choice of a unit covariance matrix in the observation noise (p13, L23) and the choice of covariance matrix in initialising the ensemble of walkers (p13, L30) need to be justified. Without these justifications it's not clear how (if at all) the results are connected to reality

SPECIFIC COMMENTS ON EXAMPLE 3

p15, L18 to p16, L1 and Figure 4. Is the MCD for B13 really the same as observed

CC BY

over the past 30 Myrs? It looks much longer to me.

p16, L17. Are there really 150 values of MCD? If there are, the averaging windows must have been truncated at one end or the other. How was this done? It looks from Figure 5 that there are only 140 values (stopping 140 Myr ago).

p17, L4-5. What is the origin of t? It cannot be the present (as in Figure 5) if k in f_k is to be positive. Does \theta_k apply to the interval before or the interval after k.1Myr?

p17, L11. I assume that \theta_k is kept constant throughout the simulation. It would be useful to mention this here. What are the initial conditions of the simulation?

p17, L11-12. The part of the sentence after 'and' seems wrong. My understanding is that a single MCD is computed for the simulation, but this part implies that a sequence of values is calculated using a sliding window.

p18, L3-4. 100 simulations for each grid point or in total? (And if the latter, how were the simulations distributed about grid points?)

p18, Figure 6. Why is the \theta grid different for the two curves? What do the small circles on the orange curve represent?

p18, L15-16. The standard deviation should be a function of \theta_k, not of f_k. The relevant pdf of \eta_k is the one conditioned on \theta_k (see note on p4, L2; p6, L27). Substituting a standard deviation that is a function of f_k into the formula for a Gaussian pdf gives something that is the pdf of a Gaussian random variable when f_k is held fixed, but when \theta_k is held fixed it might not even be a valid pdf.

p18, L17-18; p19, L5-6. What are the justifications for these priors? They will need to be justified if the intention is to draw conclusions about physical reality.

SPECIFIC COMMENTS ON EXAMPLE 4

p20, L3-4. How is \theta determined for these integrations of the KS equation?

CC BY

p20, L4. Over what range are the Fourier coefficients uniformly distributed?

p21, L1-2. How is \theta determined for these integrations of the KS equation?

p22, L17. Are these snapshots obtained in the same way as at p21, L1-4 or in some other way?

p23, L15. Even when it is restricted to be diagonal, R is not merely a scaling factor. The ratio of the diagonal elements determines which of the two components of the feature has to be matched most closely, and this can have a large effect on \theta.

p23, L16. As in my comments on Example 3, R should be a function of \theta, not of f.

p23, L20 onwards. Like, I suspect, most of the potential readership I'm unfamiliar with global Bayesian optimization and need a few more details to understand, even in outline, what is going on. I list specific points below. It would also be useful if the paper gave additional references for the method to increase the chance that at least one of them is in the library.

p24, L1. What is the log marginal likelihood and how does it allow us to estimate four hyperparameters from three function evaluations?

p24, L4-5. How is the updating of the GP done?

p24, L5. I assume that the \mu here is the mean function rather than the parameter. To avoid confusion it would be better to write \mu(\theta) here or adopt another notation for the parameter (such as \mu_0).

p24, L9. How were the extra function evaluations used to improve the GP model?

TECHNICAL CORRECTIONS

p4, L9. What is v? Is it a typo for \epsilon or something different?

p5, L14; p13, L24; p24, L3. 'Matlab' should be 'MATLAB' (at least that's how MathWorks spells it).

CC BY

p5, L14; p13, L25. Grinsted lacks a year.

p5, (4). N_e hasn't been defined.

p5, L27. 'the the' should be 'the inverse'.

p5, L31 to p6, L1. 'm dimensional' should be hyphenated.

p6, L4. p_f(\theta) should be p_f(\theta|f).

p6, L11. 'are' should be 'is' (to agree with 'exception').

p6, L12. 'posterior distribution' should be 'likelihood'.

p6, L15. Insert comma after 'however'.

p6, L28. 'feature based' should be hyphenated.

p7, L4. Insert 'the' before 'EnKF'.

p7, L16; p8, L8. This is the third meaning of k. In Section 2.2 it was a sample index, and at the start of Section 3 (p5, L31) it was the dimension of data space. Could some other notation be found?

p7, L24. 'likelihoods' should be 'likelihood'.

p8, L8. The second 'values' should be 'vectors'.

p8, L18. Comma after 'see' should be before (and perhaps stronger).

p8, L20. Insert commas around 'however'.

p9, L3. Insert comma after 'hence'.

p9, L5. Insert comma before 'however'.

p9, L11. Hyphenate 'feature based'.

p9, L15. Delete comma after 'field'.

CC BY

p9, L25-28. It would be useful to remind the reader which of the three types of section 3.3 example 4 is (as was done for the other three examples).

p10, L13. T should be t.

p10, L26. ‘form’ should be ‘from’.

p10, L28. Period should be outside final parenthesis.

p11, Figure 1, top right. x-axis is duration, not number of data points.

p11, L1. ‘suggest’ should be ‘suggests’.

p11, L2. ‘were’ should be ‘we’.

p11, L4. Delete second ‘posterior distributions’.

p11, L4. p_{450} should be p_{500} and not followed by a comma.

p11, (8). Drop subscripts on z_M for consistency with notation on p10.

p12, L9. ‘features’ should be ‘feature’.

p12, L10-11. This sentence lacks a main verb.

p12, L10-14. For consistency with the notation of Section 3.1, R should be R_f.

p12, L14-15. For consistency with the notation of Section 3.1, \mathcal{F}_\mathcal{M} should be mathcal{M}_\mathcal{F}.

p13, L15-17. The curious reader will appreciate being told at this point that the data can be seen in Figure 3.

p13, L20-21. On first reading the first sentence I couldn’t understand how singular values and vectors had been got from the data. By the end of the second sentence it was clear that the data had been arranged in a 2x11 matrix. It would be good to mention this first.

CC BY

p13, L24. ‘feature based’ should be hyphenated.

p14, Figure 2. The notation used for the axis labels is a mixture of subscripted \theta-components and parameter names. It would be neater to change the first four labels to \alpha, \beta, \gamma, \delta.

p15, L12. \theta(t) should not be here. It hasn’t been introduced yet.

p15, L18; p18, L7; p23, L7, L18; p24, L12. Stronger punctuation is needed before ‘however’ (and perhaps a comma afterwards). With just a comma before ‘however’ I was expecting a sentence of the form ‘The EnKF will not work with these data, however much the covariance in inflated’. Instead, what we have is ‘The EnKF will not work with these data; however, the WEnKF will’.

p16, L4. Delete space before comma.

p18, Figure 6, caption, L2. ‘bards’ should be ‘bands’.

p18, L4. ‘reversal’ should be ‘reversals’.

p18, L8. ‘upper’ and ‘lower’ should be swapped for consistency with the rest of the sentence.

p19, L8-9. ‘right’ should be ‘left’ and vice versa.

p19, L28. The domain in Figure 9 is [0,10]\times[0,10].

p20, L2. Delete first ‘different’.

p22, L2. ‘features’ should be ‘feature’.

p22, L13. ‘feature based’ should be hyphenated.

p23, L6. Insert ‘to’ after ‘due’.

p23, L11. ‘issue’ should be ‘issues’.

p23, L30. Based on the table of contents of Frazier and Wang, it looks like ‘chapter 3.5’

CC BY

should be ‘section 3.3.5’.

p24, L12. ’lead’ should be ‘led’ or ‘leads’.

p24, L15. ‘generated’ should be ‘generate’.

p27, L18. Volume and page numbers should be ‘198, 597-608’.

p28, L12, L14. The title of the same journal is given in two different ways.

p28, L26. This is a chapter in a monograph but has been formatted as though it were a paper in a journal.

p29, L4. The second ‘reversals’ should be ‘field’.

p29, L7. Unlike the titles of other journal papers, this one has been capitalised.

p29, L12. ‘4144’ should be ‘4114’.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-52, 2017.

Printer-friendly version

Discussion paper

CC BY