Reponse to Reviews – TOSN-2010-0006

Real-Time Information Processing of Environmental Sensor Network Data using Bayesian Gaussian Processes

M. A. Osborne and S. J. Roberts, A. Rogers and N. R. Jennings

6th September 2010

Thank you for comments on our submitted manuscript. We would like to thank all the reviewers for their insightful and useful comments. We believe that the changes we have made in response have significantly improved the quality of the paper.

In summary, we have made the following broad changes:

1. We have included a detailed algorithm description (Algorithms 1) to Section 3 that provides a roadmap of how to apply our approach to a particular dataset.
2. To show that the approach that we describe does indeed scale up to much larger datasets we have included new empirical evaluations to Section 4 using the UK Met Office MIDAS land surface weather station dataset. In doing so, we apply our approach to a dataset comprising 670 sensors, and present a covariance function parameterised by the longitude and latitude of the sensors.
3. For each of the empirical evaluations, we now present the covariance function used and discuss in detail the reasons for this choice.
4. We have added a more detailed description of the state-space alternative to Section 4 and also discussed its limitations more fully in Section 7. In addition, we have represented the empirical evaluations both as RMSE and also as SNR expressed as a normalised mean square error with units of dB, and presented this comparison in the new Tables I-IV.
5. We have extended the presentation of computation time in Section 6 to show the number of samples and the number of hyperparameters used in each of the comparison (Table VI) so that the relationship between these numbers and the size of the sensor network is clear. In addition, we have compared the computation times with the efficient update rules detailed in our paper with the equivalent times without such rules in Tables V(a) and V(b), and show a four-fold increase in computation speed.

In the following pages, we describe in detail the changes that we have made to address the specific points of each of the reviewers.

Referee: 1

Recommendation:

Needs Major Revision (revised Paper Will Be Sent Out For Re-review)

Comments:

The modeling ideas proposed in this work are mathematically sound and interesting. The main concern I have is with the evaluation of the practical benefits of the proposed approach. The authors outline several goals informed by a need for real-time data processing. These include operating with incomplete information, uncertainty and doing so with relatively low complexity algorithms. To further strengthen the case for the proposed techniques I would suggest the following:

- Real datasets are used in the empirical evaluation. While the results are convincing (and realistic) they are also somewhat limited to just two deployments, with small numbers of sensors and specific characteristics in terms of correlation. Do these results generalize to larger deployments? to cases where there is less correlation between sensor measurements? based on the results provided it is hard to determine whether similarly good results can be achieved in general. As an example, even for a relatively modest size deployment (16 sensors) in the second real dataset, the authors already need to make simplifications in their model (p. 20). Potentially a broader experimental evaluation with simulated data might help to make this case. A more general discussion of the kinds of networks sizes for which these approaches are suited would also be helpful (can they be used for networks of 100s of nodes?).

> The approach that we describe does indeed scale up to much larger datasets. To show this have included new empirical evaluations to Section 4 using the UK Metrological Office MIDAS land surface weather station dataset. In doing so, we apply our approach to a dataset comprising 670 sensors, and present a covariance function parameterised by the longitude and latitude of the sensors.

- The evaluation of complexity is also unconvincing. First, there are no comparisons with other methods, so we don't know how much savings the proposed method leads to. Also, the table that is provided gives actual computation values in seconds in an off the shelf computer using Matlab. The authors should explain whether this is a realistic setting. From the paper it is not clear what is required in order to operate in "real-time". Is it possible to describe the kinds of compute resources, delay constraints, etc. that would apply in a desirable "low cost" deployment?

> We have extended the presentation of computation time in Section 6 to show the number of samples and the number of hyperparameters used in each of the comparison (Table VI) so that the relationship between these numbers and the size of the sensor network is clear. In addition, we have compared the computation time with our approach, with the equivalent time without the efficient update rules detailed in our paper in Tables V(a) and V(b), and show a four-fold increase in computation speed over the existing highly optimised LAPACK routines that are used as an alternative.
>
> We would note that the information processing described here is intended to be performed at the base station or at the data portal, and thus, the computational requirements does not impact on the sensors themselves. We have added a footnote to Section 6 to this effect. In our discussion of computation time in Section 6, we now show that for all the information

processing tasks presented, it is feasible to perform them in real-time (assuming data collection intervals of > 1 minute), and that the extra computation time can be used to increase the number of data points and/or the number of hyperparameter samples used.

Referee: 2

Recommendation:

Needs Major Revision (revised Paper Will Be Sent Out For Re-review)

Comments:

The implementation and evaluation sections could be expanded, for completeness and also to make the technique more accessible to an estimation/learning non-specialist. Some specific suggestions are as follows:

1. Consider providing a 'roadmap' for going from the theory and equations to an implementation i.e. starting with observations, outline the sequence of computations leading to the final results.

> We have included a detailed algorithm description (Algorithm 1) to Section 3 that provides this roadmap.

2. Since the limitations of and improvement over state-space models provides a key motivation for the new technique, this aspect could be brought out more strongly. In addition to the discussion in Section 7, highlighting the same in a specific manner for the implementation and evaluation in Section 5 will be useful. Provide more detail on the Kalman filter implementation to convince the reader that a fair comparison is being made. Consider including a table with results for the comparative evaluation for several different instances/variables/experiments. Also, use SNR for performance evaluation, in addition to RMSE.

> As suggested, we have added a more detailed description of the state-space alternative to Section 4 and also discussed its limitations more fully in Section 7. In addition, we have represented the empirical evaluations both as RMSE and also as SNR expressed as a normalised mean square error with units of dB, and presented this data in Tables I-IV.

3. Add detail to empirical evaluation. To help connect with the theory in Section 3, for each environmental variable in Section 5,

- identify the hyperparameters

- show choice of covariance functions

> For each of the empirical evaluations, we now present the covariance function used and discuss the reasons for this choice.

Typos:

Pg 4: "large quantities of data timestamped data."

> Fixed.

Pg 13: "parses and stores is using Jena"

> Fixed.

Pg 15: "Air temperature was chosen since they exhibit"

> Fixed.