Partition modeling with arbitrary data noise: a 2015 research highlight

M. Sambridge, R. Hawkins, K. Lambeck, E. Rohling, K. Grant.

During the year a new research question emerged out of ongoing studies in geophysical data inference concerning signal reconstruction and change-point detection in time series data sets with arbitrary errors in both measured and independent variables, i.e. y and t. We choose to highlight this for 2015 as it is an example where a practical research need drove some new theoretical insights in Bayesian regression and resulted in a workable algorithm that extends capabilities.

In recent years an area of research within the group has been application of partition modeling to geophysical change-point detection. A change-point is a discontinuity, or sudden change in either the amplitude of a time signal and or its derivative. The last ten years has seen much progress for this problem in Bayesian statistics, where probabilistic information can now be extracted from observations about the location (in time) and magnitude (e.g. in sea level) of events. Detection of changepoints is of interest in many areas of the geophysics, either to remove instrumentation or processing errors, e.g. in Geodetic measurements, or to find sudden changes in behavior as in geochemical proxies of environmental studies. These methods are also of use in situation where changepoints are not expected but one seeks reliable ways of reconstructing (y, x) or (y, t) signals from unevenly distributed noisy observations, with quantifiable uncertainty, e.g. in Sea-level reconstruction (Lambeck et al. 2014).

Previous work with the group in collaboration with colleagues in London and Rennes, was responsible for pioneering the use of partition modeling in the geosciences (Gallagher, et al. 2011). Local interest in such regression problems is an off-shoot of our broader studies on trans-dimensional inversion, where we have used similar concepts as the basis of novel Earth imaging methodologies (Sambridge et al. 2013). The new study arose from separate discussions with Profs. Lambeck, Rohling and Dr. Grant on the robustness of their past sea-level estimation studies from noisy time series data. An earlier collaboration with Prof. Lambeck and co-workers led to our first application of partition modeling to this problem which was restricted to the case where errors are assumed only in the dependent variable, e.g. sea-level height. Normal practice is to expand uncertainty in observed data y values to account for proportionately smaller errors in observed x or t values. The more general problem is to account for arbitrary large errors in both y and t observations including correlation between noise values. This has now been done by deriving a new Likelihood function for the general noise case and combing this with our existing partition modeling algorithms from probabilistic reconstruction of time signals. While regression has been studied for more than 100 years in statistics and more recently from a Bayesian sampling viewpoint, the general case of (un)correlated noise in both variables has not previously been possible within the partition modeling methodology. Our extension does this and results in a workable algorithm. This work will be expanded on in 2016 with applications to sea-level and other types of regression data.

The Figure shows an illustrative (toy) example where the new algorithm probabilistic regression algorithm has been applied. Here 20 noisy (x,y) observations are synthetically generated from a four partition linear segment curve (red) and we wish to recover the red curve as best we can including realistic estimates of uncertainty. Panel a) shows the original data with known standard deviations of its noise in both variables as error bars, together with the true curve.

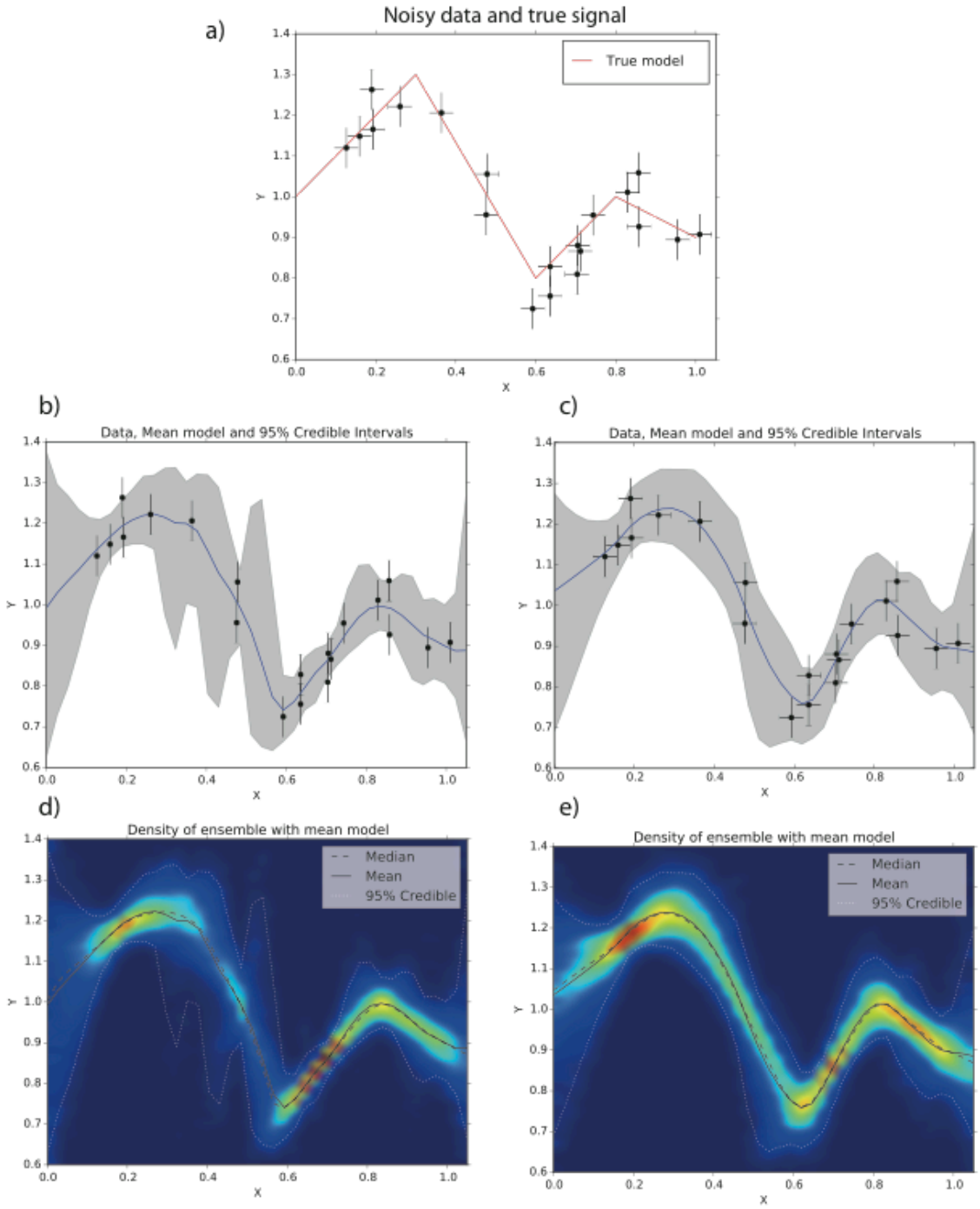

Figure 1. Example of partition modeling with the new methodology on a 20 data problem.

In a Bayesian sampling framework the solution to the inference problem is a probability distribution on the location of the true (red) curve. Typically then we generate an ensemble of curves as a representation of the solution rather than choose a single 'best fit' curve. The variable of the ensemble contains all information of what we learn from the data and the spread of the ensemble gives a measure of uncertainty. Panels b) and c) show the results of partition

modeling to estimate the red curve with uncertainty. The blue curve and the gray region show the mean and 95% uncertainty (credible) interval of 50,000 curves generated by Bayesian partition modeling whose location density is constrained by the quality of fit to the observations measured by a Likelihood function. In the left hand panel (b) the x errors are ignored and standard partition modeling is applied. In the right panel (c) the new Likelihood function is used and both x and y errors are accounted for. The curve in (c) is a closer recovery of the truth (a) but particularly the uncertainty estimation is more stable and realistic with the new approach. In both cases the smoothness, or complexity, of the recovered signal is driven by the data using trans-dimensional McMC. Figures (d) and (e) show the full density of the 50,000 solution curves generated in (b) and (c) respectively, as a colour image. Warm colours indicate high density for the location of the true curve.

References

Gallagher, K., Charvin, K., Nielsen, S.,Sambridge, M. and Stephenson, J., 2009. Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems, *Marine and Petroleum Geology*, **26**, 525-535, doi:10.1016/j.marpetgeo.2009.01.003.

Lambeck, K., Rouby, H., Purcell, A., Sun, Y., and Sambridge, M., 2014.
Sea level and global ice volumes from the Last Glacial Maximum to the Holocene, *Proc. Nat. Acad. Sci.*, **111**, no. 43, 15296-15303, doi:10.1073/pnas.1411762111.

Sambridge, M., Bodin, T., Gallagher, K. and Tkalčić, H. 2013.
Transdimensional inference in the geosciences,
Phil. Trans. R. Soc. A., 20110547, http://dx/doi.org/10.1098/rsta.2011.0547