Simulation Study

Problem

A simulation study was conducted, to evaluate the performance of the different introduced smoothing techniques in recovering non-linear state processes. Specifically, we were interested in the extend to which different types of non-linear state functions could be estimated under conditions that are realisitcally found in ESM research. For the present simulation we further limited considerations to univariate single subject desgins, for simplicity and to use each method in the form in which they are currently available in software. Thus, we manipulated two conditions, which were theorized to influence the state recovey. First, the type of non-linearity present in the state process, the quality of information about the state process.

Design

Based on theory underlying the considered smoothing techniques, we expect that all considered smoothing techniques will perform best if the non-linear state process is relatively smooth and follows a constant dynamic without regime switching. In order to test this, we operationalized smoothness as either low or high process noise. Since, we expect there will always be some process noise present in ESM data, we did not include a completely smooth condition without process noise. Further, we tested two types of regime switching. The first regime switching condition changes the dynamic of the simulated process, but not the attractor, whereas the second condition, switches the attractor of the state process, but not the dynamic.

Further, we expected that all considered methods will infer the state process with higher accuracy the better the quality of information is. To test this, we varied two factors. The first is to the signal-to-noise ratio of the data, which corresponds to the ratio of the variance of the centered process values and the variance of the measurement noise. Thus a larger signal-to-noise ratio corresplonds to more variation in the process itself relativ to the measurement uncertainty. The second factor corresponds to the sample size or the number of observations that are taken over a given time interval.

Therefore, the simulation consisted of five factors that were crossed in a complete factorial desgin:

1. Process noise (small vs. large)
2. Dynamic regime switching (absent vs. present)
3. Attractor regime switching (absent vs. present)
4. Signal-to-noise ratio (insert here common ESM effect sizes)
5. Sample size (insert here common ESM sample sizes)

Procedure

In all conditions, the data was generated from a univariate non-linear state space model, with a state function described by:

$$Y_{t+1} = f_j(Y_t) + \eta_{t+1}$$

$$\eta_{t+1} \sim N(0, \sigma^2_{Process})$$

Where f_j is either a single continous non-constant function in the single regime condition or switching between two different functions according to a Markov chain in the regime switching conditions. In the regime swtiching conditions these functions were selected in such a way that either the dynamic of the system, the attractor, or both change. In addition to that, the variance of the η_{t+1} was set according to the process noise condition and the time-points were determined by subdevidign the interval from zero to one by the desired sample size. After simulating the state values, we centered the state values and calculated the measurement error that would result in the required signal-to-noise ratio. Laslty, the measurement error was added to the simulated state values to generate observations.

N number of data sets were generated per cell of the factorial desgin. This number was determined by first running a small pilot simulation with 30 data sets per cell. Afterwards, a seperate simulation was conducted on the outcome models. To simulate from the outcome models we used the significant parameter estimates from the pilot simulation and set the insignificant parameters to zero. Based on these preliminary effect sizes N observations per cell resulted in selecting the correct outcome model 95% of the time.

Model estimation

General additive models

After simulating the deta, four methods were used to infer the state process for each data set. First, general additive models (GAM) were fit to each data set using the mgcv package. Specifically, a model with a single smooth function relating the state value to time was chosen and approximated using 50 thin-plate splines. The weight of the penalty term of the GAM was determined using the generalized cross-validation procedure that is incorporated in mgcv. The fitted model was then used to obtain state estiamtes at each time point as well as respective 95% Bayesian credible intervals

Local polynomial regression

Secondly, a local polynomial regression was fit to each data set using the nprobust package. We used first order local polynomials, resulting in a local linear regression. Further, we used a gaussian kernel and optimized the bandwidth of the kernel function using the build-in mean squared error optimization. Afterwards, bias corrected state estimates and robust standard error were obtained, to account for the inherrent bias of the local polynomial model. Lastly, we used these estiamtes and standard errors to construct 95% confidence bands for the state function.

Gaussian processes

In order to fit the Gaussian processes efficiently for a large number of data sets we used a Hilbert space approximate Gaussian process model in Stan. This approximation uses a linear combination of basis functions that are derived from the covariance matrix of the Gaussian process. The number of basis function was chosen in accordance to the length scale of the Gaussian process and boundary factor of the approximation. This was achieved by following the selection procedure detailed by Riutort-Mayol et al. (2023) To concur with how Gaussian processes are usually applied we selected a zero mean function together with a quadratic exponential covariance functions. This covariance function is both symmetric and stationary. Point estimates for the state value as well as 95% credible interval were obtained from posterior distribution of the Gaussian process.

State-space models

Lastly, a parametric dynamic model is fit to each data set using the dynr package. Here, the specific models were adapted to corrspond exactly to the data generation models of each respective data set. This mimics the situation in which prior theory about the parametric form of the dynamic model exists. Further, each model was estimated 10 times using different starting values to minimize the risk of obtaining local minimum solutions. Estimates for the smoothed state values and the smoothing error covariance were obtained directly and used to generate confidence intervals for the state inference.

Outcome measures

Root mean squared error

To compare the in-sample predictive accuracy of the state inference accross analysis methods and simulagtion conditions, we calculated the root mean squared error (RMSE) between the smoothed state inference and the simulated state values. The resulting RMSE values were then anaylzed using a repeated measures ANOVA, as each data set was analyzed with every method. Thus, the analysis method constitues a repeated-measuers effect, wheras the simulation conditions were entered into the model as fixed effects. To find evidence of both present and absent effects an exhaustative model selection based on AIC and BIC model weights was conducted.

Generalized cross-validation

To evaluate the out of sample interpolation predictive performance of each method we obtained generalized cross-validation scores (GCV). The GCV criterion is a computationally efficient and rotation invariant alternative to the ordinary cross-validation criterion that would be obtaiend from a leave-one-out cross-validation. Because the GAMs, local polynomail regression, and GPs are linear smoothers, the GCV can in these cases be obtained directly from the influence matrix of the model. In addition, for linear parametric state-space models there are approximate analytic results for ontaining the GCV from the innovation covariance matrix. However, since in this simulation non-linear and regime switching state-space models will be conidered, the GCV is instead calculated by performing a leave-one-out cross validation and adjust the resulting errors. The GCV values obtained for each analysis methods were anlyzed using the same repeated measures ANOVA based model selection procedure as was used for the RMSEs.

Confidence interval coverage

The last performance measure that was inestigated during the simulation is the confidence interval coverage probability. That is, the probability that the confidence generated by any of the analysis methods for the state inference, falls around the true simulated state value. To analyze this, confidence or Bayesian credible intervals were obtained for each of the state inferences and it was recorded whether the the confidence interval included the true state value or not. Afterwards, a multilevel logistic regression was used to predict the probability of a confidence interval being around the true state value using the simulation conditions and analysis methods as predictors. We again performed a exhaustative model search based on AIC and BIC model weights to evaluate the evidence for both present and absent effects.