Recovering the function for the whole region of interest, through a limited number of observed locations

Ziang Zhang

23/09/2021

Simulation Setting

In this simulation study, we aim to compare the performance of RW2 method with ARIMA method to yield the inference of some unknown effect function $g(x)$.

We consider two different types of settings for simulations. In the first type of setting, we fix the region of interest to $[0, 100]$, and vary the number of observations n between $\{10, 50, 100\}$ in that fixed interval. In the second type of setting, we fix the number of observations to $n = 50$, but vary the length of the region of interest. For simplicity, we consider the spacing between locations to be equal in all the simulation study.

The simulated data set has the form of $\{(x_i, y_i) : i \in [n]\}$, where x_i denotes the i-th (observed) covariate value and y_i denotes its corresponding observation. The inferential target is not just to know the posterior distribution of the effect function at the observed locations $g(\boldsymbol{x})$, but also to infer the shape of the function $g(.)$ at the whole region of interest. To do that, we take a high resolution equally spaced grids $\{z_i : i \in [M]\}$ where $M \in \mathbb{N}$ is much larger relative to the sample size n. Since M is large, we assume the function $g(.)$ can be well approximated by the step function $\tilde{g}(.) = \sum_{i=1}^{M} \mathbb{I}(z_i \leq . < z_{i+1}) g(z_i)$ where $z_{M+1} := +\infty$.

To obtain samples of the unobserved values $g(\boldsymbol{z})$, we first draw samples $\tilde{g}(\boldsymbol{x})$ from the posterior of $g(\boldsymbol{x})$, then sample from the conditional distribution of $g(\boldsymbol{z})|\tilde{g}(\boldsymbol{x})$ given by the prior distribution.

For the true function $g(.)$, we consider it being the function

$$g(x) = 5\sin(0.1x),$$

observed at $x \in [0, 100]$. We assume the observation level model is

$$y_i = g(x_i) + \epsilon_i,$$

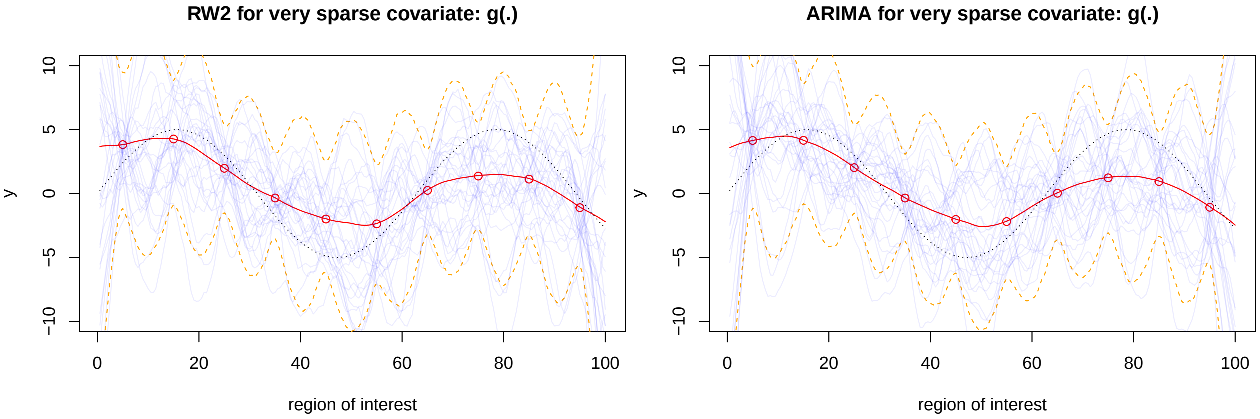
with $\epsilon_i \sim N(0, 3)$.

The performance between $\tilde{g}_{\mathrm{RW2}}(.)$ and $\tilde{g}_{\mathrm{ARIMA}}(.)$ will be compared in terms of *root integrated absolute error*(RIAE) and *mean credible interval width*(MCI). The RIAE is defined as

$$\mathrm{RIAE}(\tilde{g}) = \sqrt{\int_0^{100} |\tilde{g}(t) - g(t)| dt},$$

where the point estimate is defined using the posterior mean. These measures are computed from 100 independent replications at fixed set of observed locations.

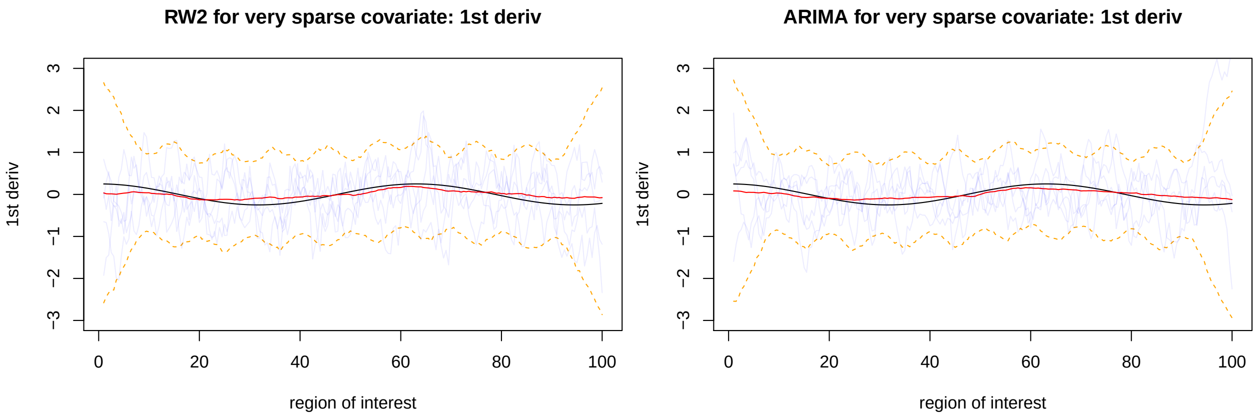
Sample Size being 10



The inferred results using RW2/ARIMA method are summarized at above. For the resolution value, we used $M = 200$. The red line represents the posterior mean of $g(\boldsymbol{z})$, the orange lines represent its 95 % credible interval, and the light blue lines are 10 sample paths simulated from the posterior distribution of $g(\boldsymbol{z})$. The black dotted line is the true function. The red points are the posterior mean values at the observed locations $g(\boldsymbol{x})$.

Just looking at the posterior mean and the posterior credible intervals, the two methods yield similar inference for $g(.)$. The corresponding RIAE and MCI are 1.266 and 13.689 for RW2 method; 1.296 and 13.651 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ that are observed.

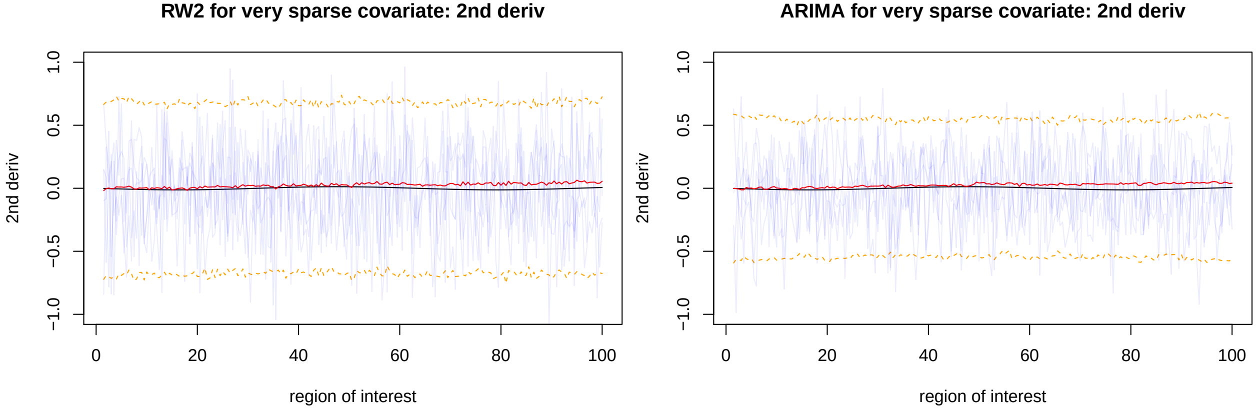
To better understand the difference between these methods, we also look at the posterior samples obtained using each method for different functional/operator on the function $g(.)$. We first consider applying first/second order derivative operator on the functions obtained using each of the method, which can be approximated using first/second order differences of the discrete vector $g(\boldsymbol{z})$ due to its high resolution.


RW2 for very sparse covariate: 1st deriv

ARIMA for very sparse covariate: 1st deriv

For the first order derivative, there seems to be no big difference between the two approaches as well. The corresponding RIAE and MCI are 0.317 and 2.335 for RW2 method; 0.315 and 2.26 for ARIMA method.

Then we will apply the same procedure to see the second order derivatives yielded by each method.


RW2 for very sparse covariate: 2nd deriv

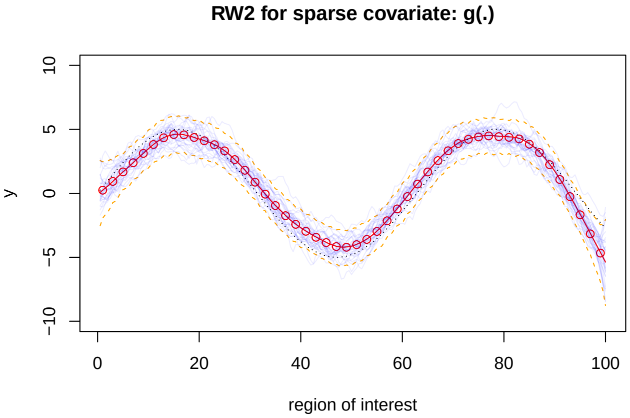
ARIMA for very sparse covariate: 2nd deriv

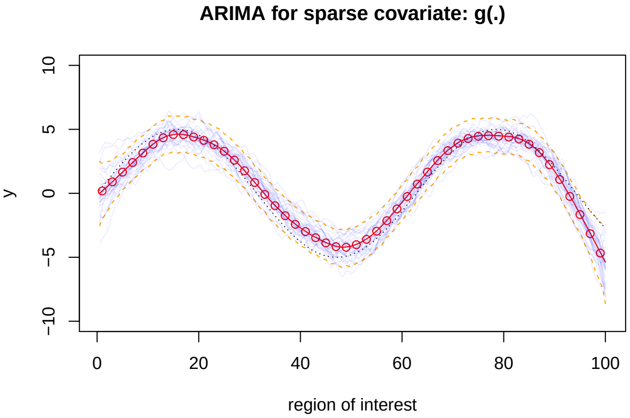
For the second order derivative, the corresponding RIAE and MCI are 0.165 and 1.359 for RW2 method; 0.163 and 1.097 for ARIMA method.

Sample Size being 50

We can see that at sample size of 10, neither methods provide satisfactory inferences for $g(.)$. We then consider the same setting for the case when sample size is 50.

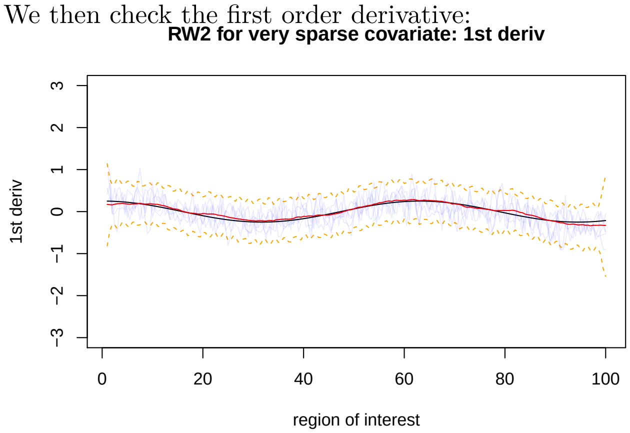
Case 1: The locations are all unique

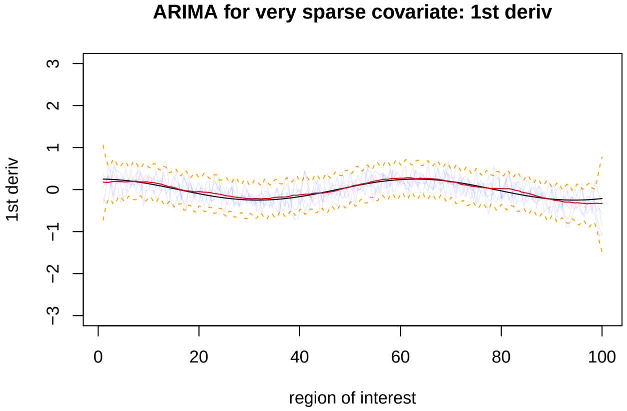




The corresponding RIAE and MCI are 0.799 and 2.93 for RW2 method; 0.795 and 2.898 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ that are observed. The conclusion is similar to what we observed for the case when $n = 10$, with the difference between MCI getting smaller.

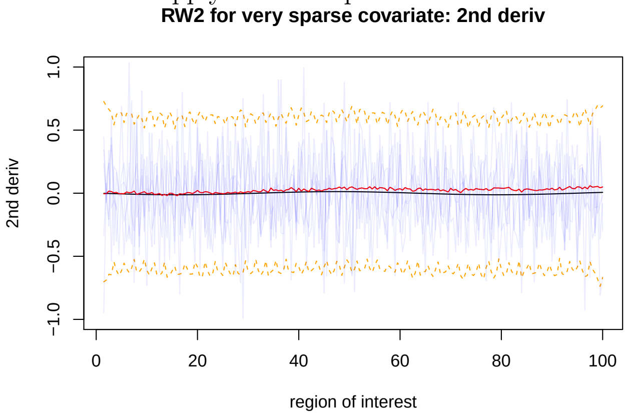
We then check the first order derivative:

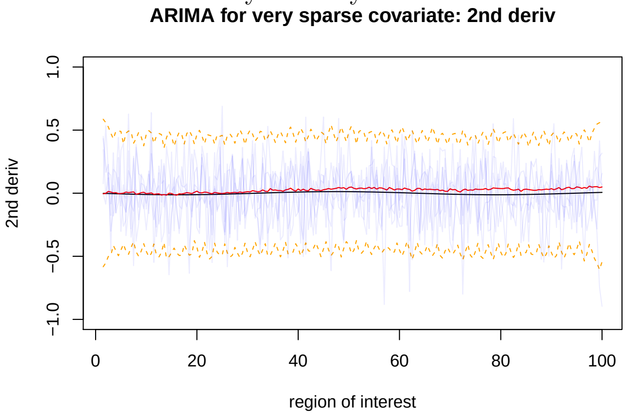




The corresponding RIAE and MCI are 0.195 and 0.979 for RW2 method; 0.193 and 0.823 for ARIMA method. We can see that the ARIMA approach yields on average a narrower credible interval

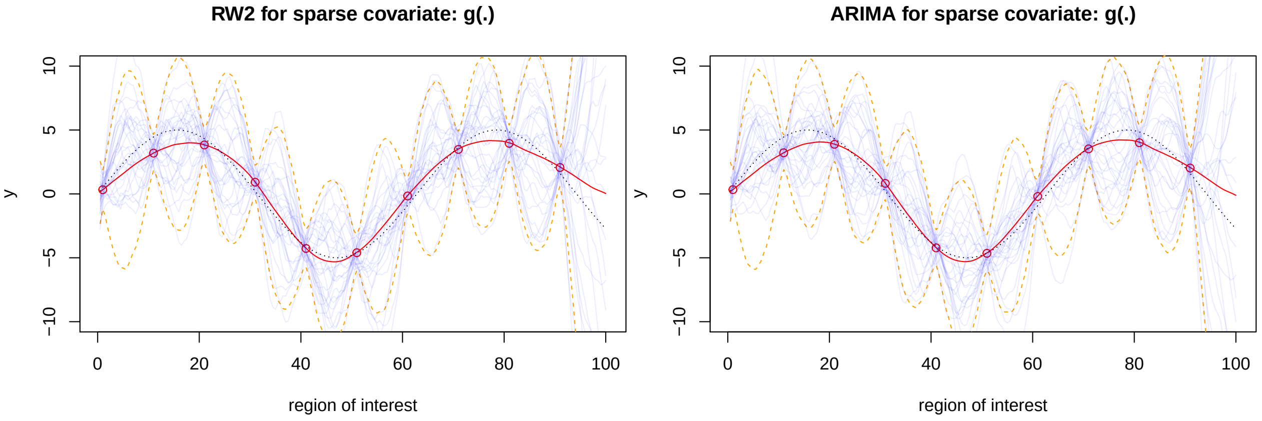
Then we will apply the same procedure to see the second order derivatives yielded by each method.





For the second order derivative, the corresponding RIAE and MCI are 0.159 and 1.211 for RW2 method; 0.158 and 0.915 for ARIMA method.

Case 2: There are 5 repeated measurements at 10 unique locations

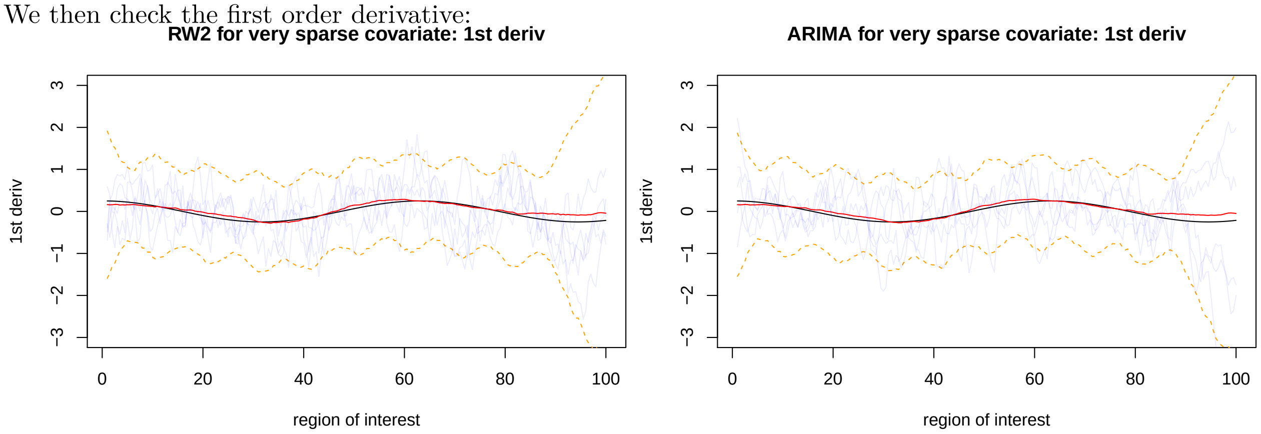


RW2 for sparse covariate: g(.)

ARIMA for sparse covariate: g(.)

The corresponding RIAE and MCI are 0.801 and 12.264 for RW2 method; 0.777 and 12.158 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ that are observed. The conclusion is similar to what we observed for the case when $n = 10$, with the difference between MCI getting smaller.

We then check the first order derivative:

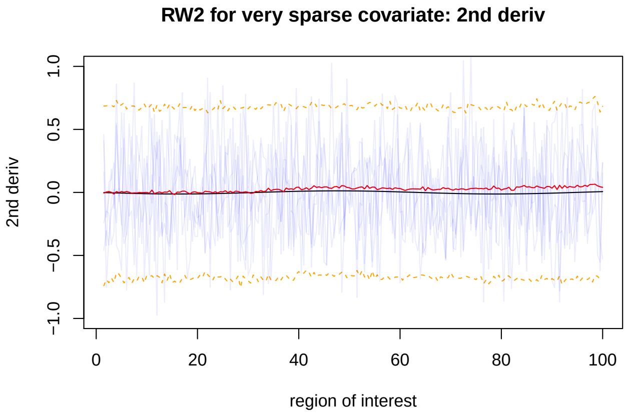


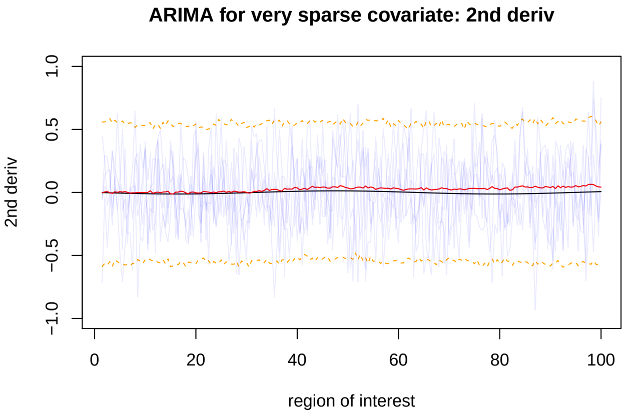
RW2 for very sparse covariate: 1st deriv

ARIMA for very sparse covariate: 1st deriv

The corresponding RIAE and MCI are 0.248 and 2.368 for RW2 method; 0.241 and 2.295 for ARIMA method. We can see that the ARIMA approach yields on average a narrower credible interval

Then we will apply the same procedure to see the second order derivatives yielded by each method.





For the second order derivative, the corresponding RIAE and MCI are 0.163 and 1.358 for RW2 method; 0.163 and 1.097 for ARIMA method.

Summary

(Summarize what you currently observed with these cases; what will be done for the next step: 1. Replication study using independent samples. 2. Study the relationship between (aggregated) difference in MCI, with the nominal coverage rate. 3. Study the more complex functional/operator defined for the function $g()$.)

Conclusion:

From the first three set of simulation studies (where n is changing from small to large), we can make the observation that, the two Bayesian smoothing methods perform similar in terms of MSE, but the ARIMA method gives smaller values of MCW. This difference gets larger as the sample size n declines.

The same conclusion is also supported by the second sets of simulation studies (where the region of interest is changing). Unless the sample size is large enough and the spacing between locations is small, ARIMA method yields more favorable inferential result than the RW2 method.