# 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

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

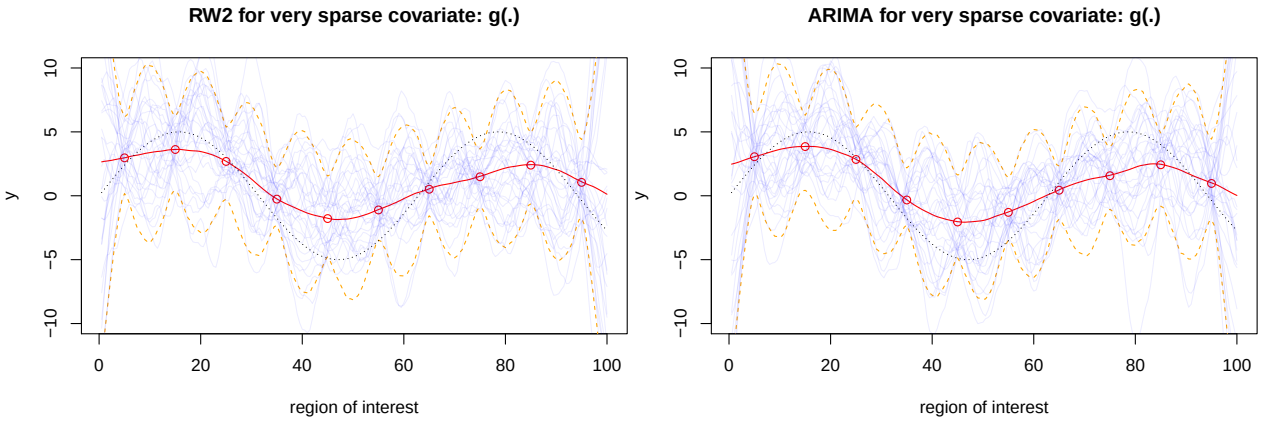
with $\epsilon_i \sim N(0, 3)$.

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

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

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

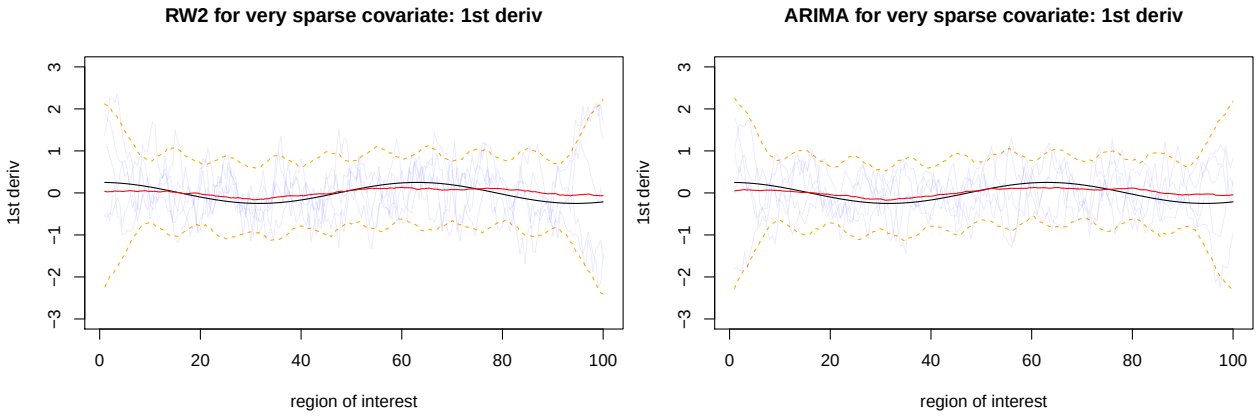
## Sample Size being 10




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 90 % 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 $g(\boldsymbol{x})$.

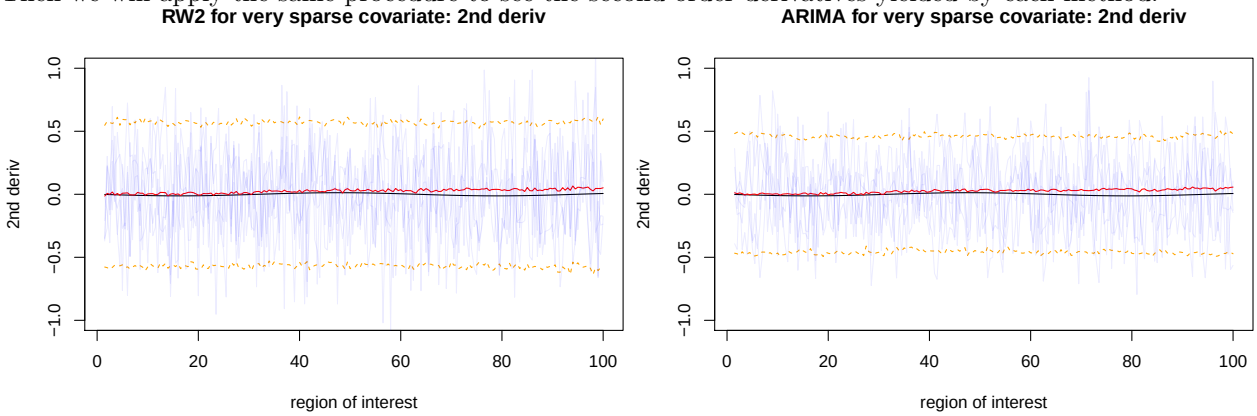
Just looking at the posterior mean and the posterior credible intervals, the two methods yield similar inference for $g(.)$. The corresponding RIAE and MCI are 1.284 and 10.427 for RW2 method; 1.238 and 10.401 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ 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.




0.05 For the first order derivative, there seems to be no big difference between the two approaches as well. The corresponding RIAE and MCI are 0.333 and 1.951 for RW2 method; 0.323 and 1.884 for ARIMA method.

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

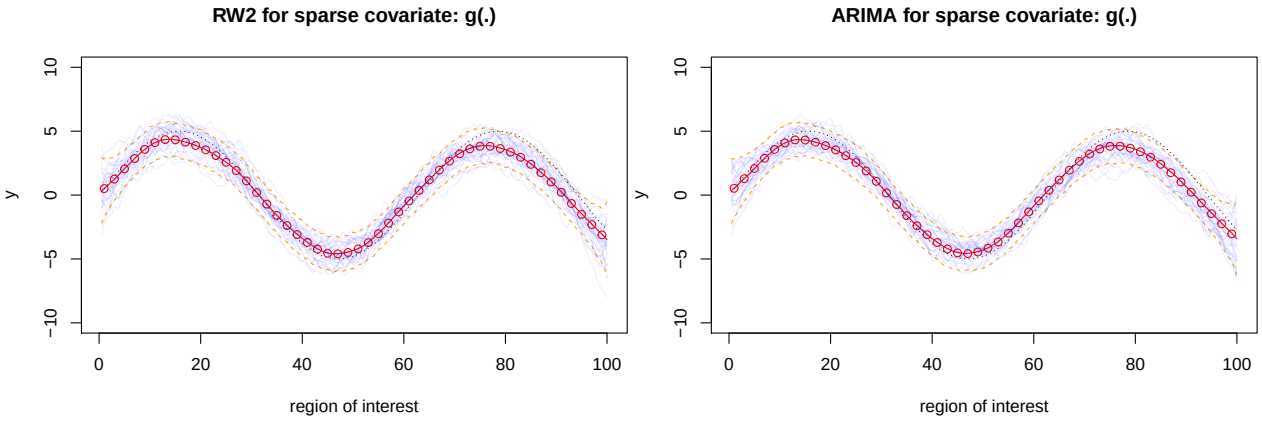



For the second order derivative, the corresponding RIAE and MCI are 0.165 and 1.14 for RW2 method; 0.164 and 0.921 for ARIMA method.

## Sample Size being 50

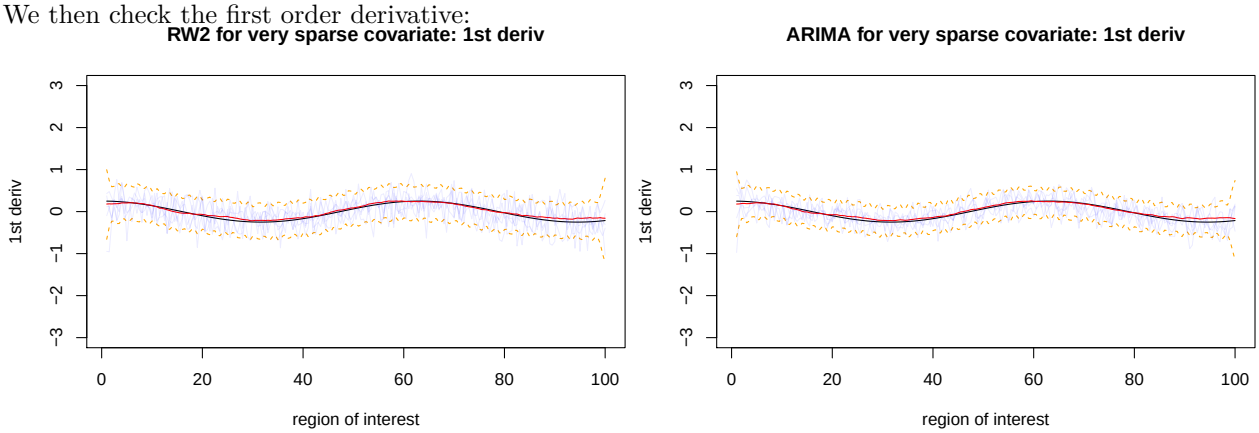
We can see that at sample size of 10, neither methods provide satisfactory inferences for $g(.)$. We then consider the same setting for the case when sample size is 50.

## Case 1: The locations are all unique



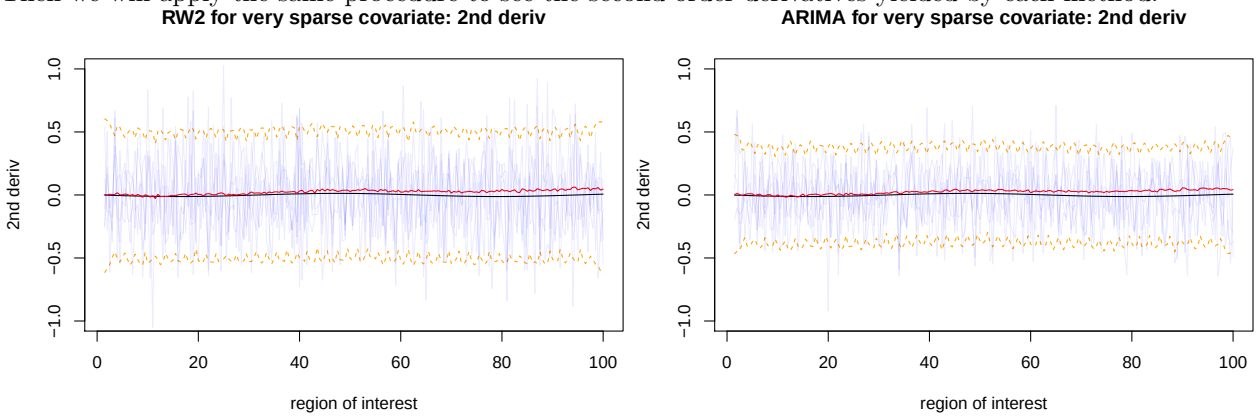

The corresponding RIAE and MCI are 0.78 and 2.822 for RW2 method; 0.781 and 2.768 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ that are observed. The conclusion is similar to what we observed for the case when $n = 10$, with the difference between MCI getting smaller.

We then check the first order derivative:



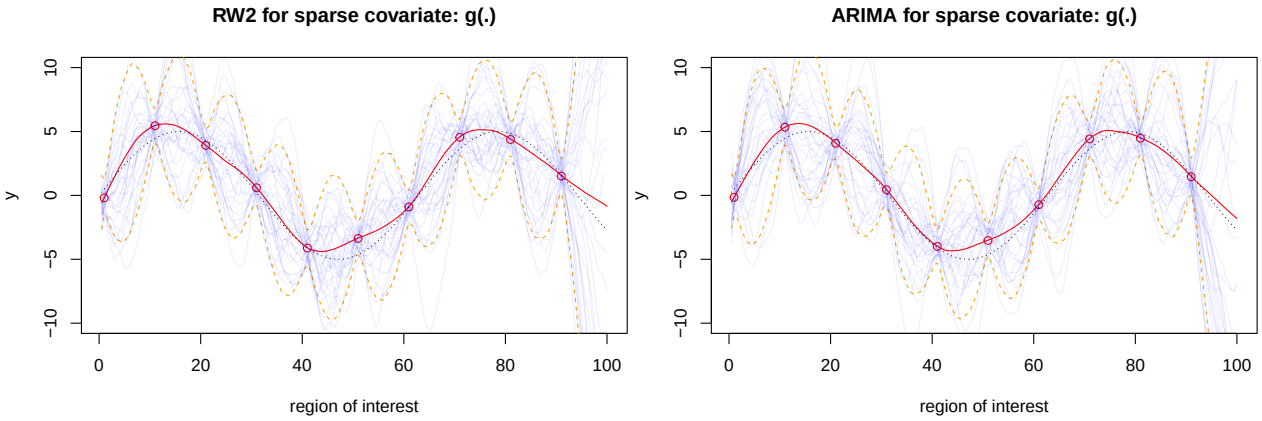

The corresponding RIAE and MCI are 0.186 and 0.821 for RW2 method; 0.186 and 0.693 for ARIMA method. We can see that the ARIMA approach yields on average a narrower credible interval

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



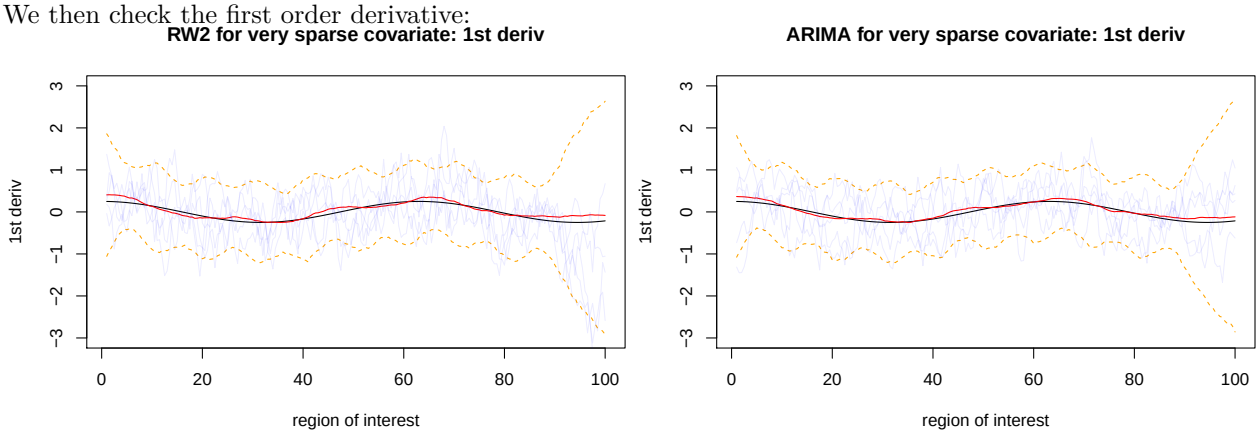
For the second order derivative, the corresponding RIAE and MCI are 0.162 and 1.016 for RW2 method; 0.161 and 0.768 for ARIMA method.

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



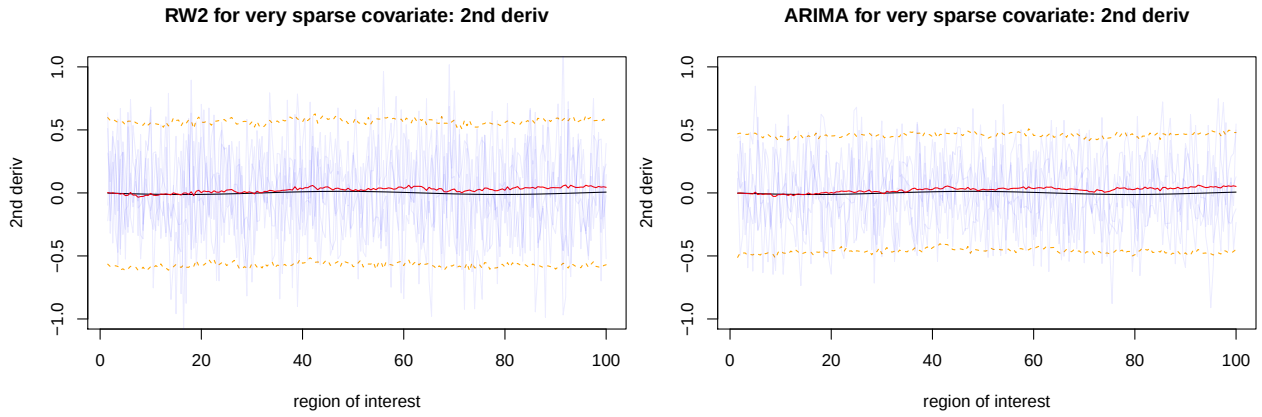

The corresponding RIAE and MCI are 0.732 and 10.396 for RW2 method; 0.653 and 10.299 for ARIMA method. The posterior credible intervals are shrunk at the locations of $\{z_i; i \in [M]\}$ that are observed. The conclusion is similar to what we observed for the case when $n = 10$, with the difference between MCI getting smaller.

We then check the first order derivative:




The corresponding RIAE and MCI are 0.259 and 1.984 for RW2 method; 0.228 and 1.915 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.

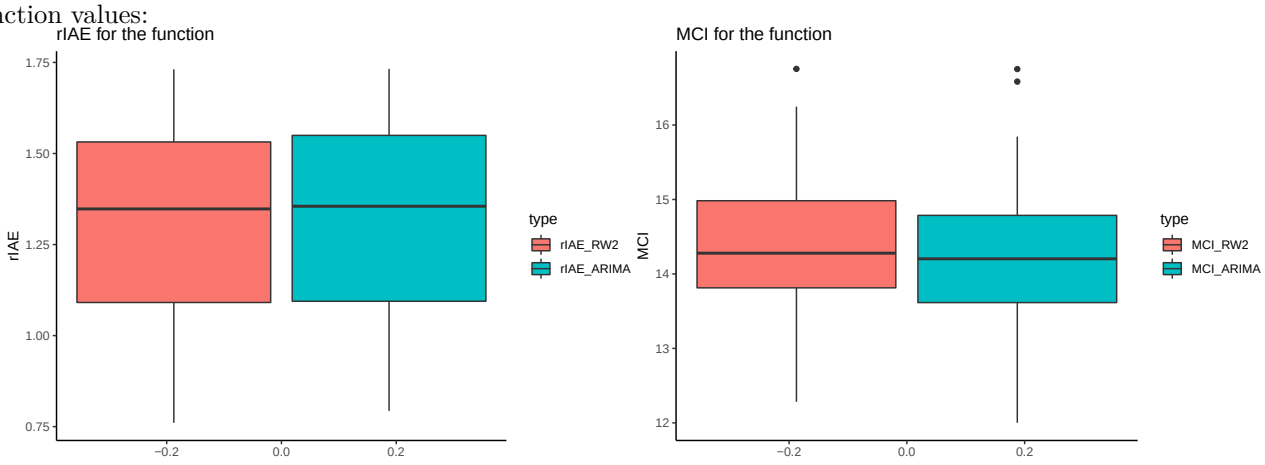



For the second order derivative, the corresponding RIAE and MCI are 0.164 and 1.141 for RW2 method; 0.161 and 0.922 for ARIMA method.

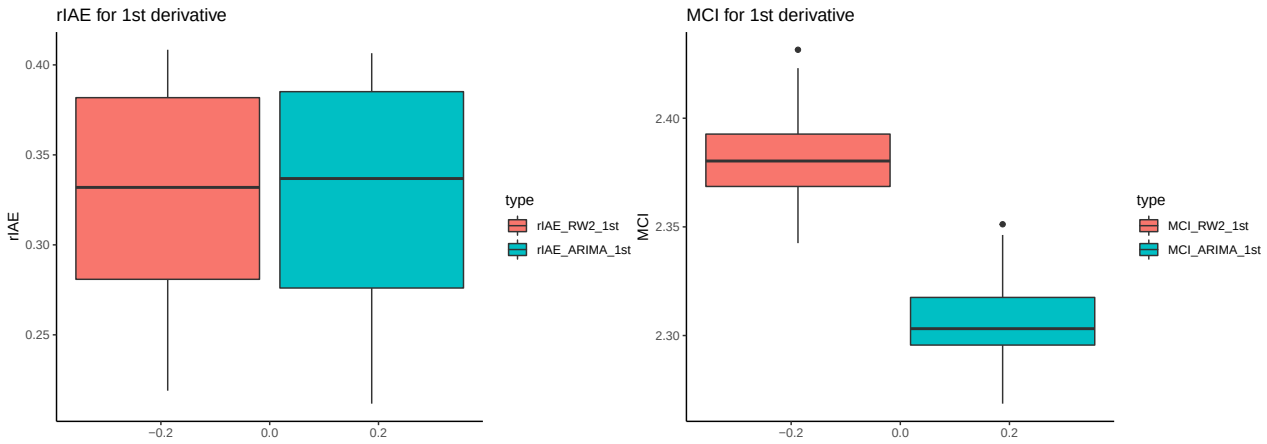
## Replication for 100 independent datasets:

To formally compare the performance of the two methods in these three different scenarios (n = 10, n = 50 with unique locations, n = 50 with repeated measurements), we replicate the inferences in each scenario with 100 independent datasets.

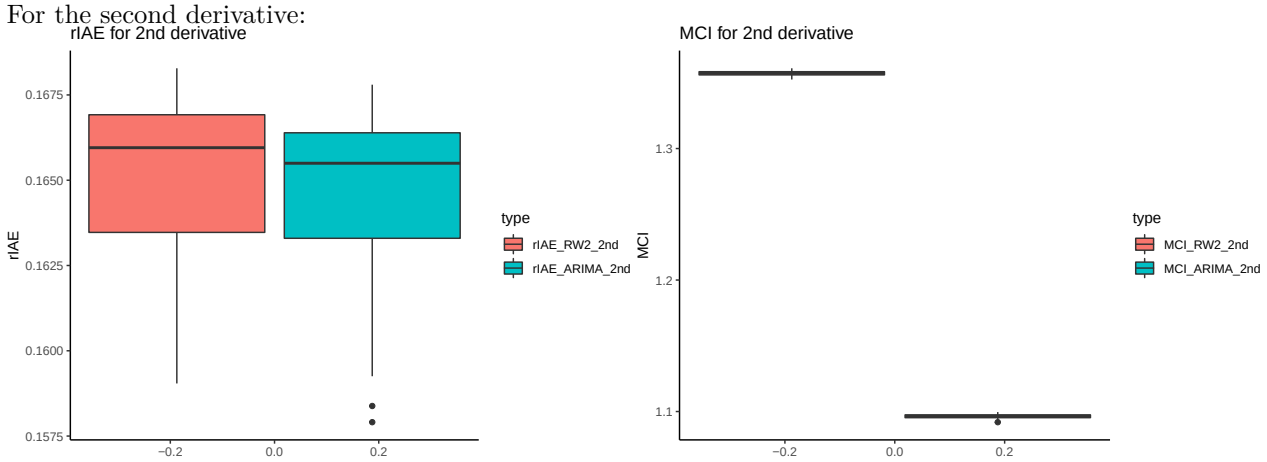
We first consider the first scenario with 10 unique locations and no repeated measurements. For the function values:



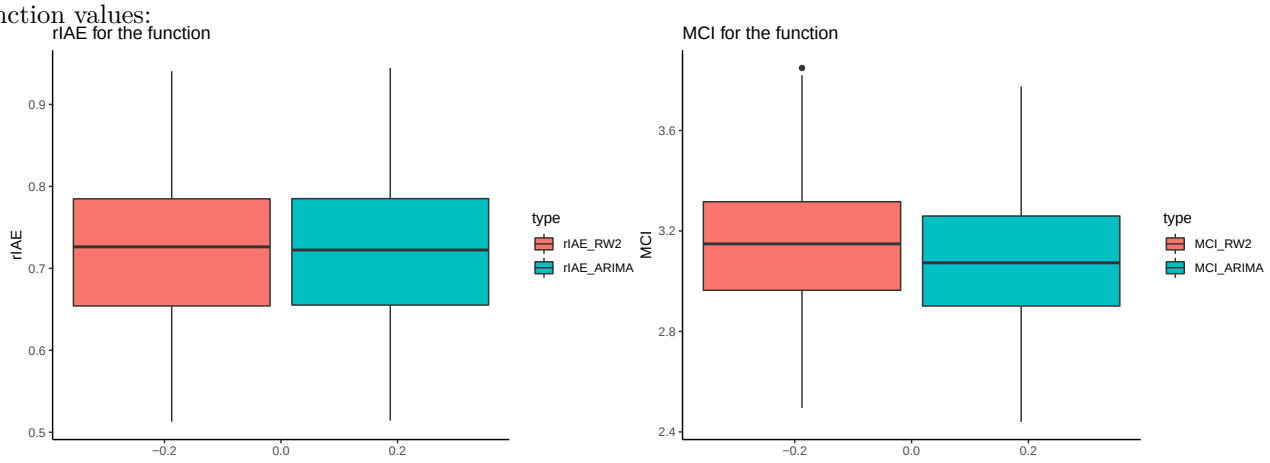

For the first derivative:



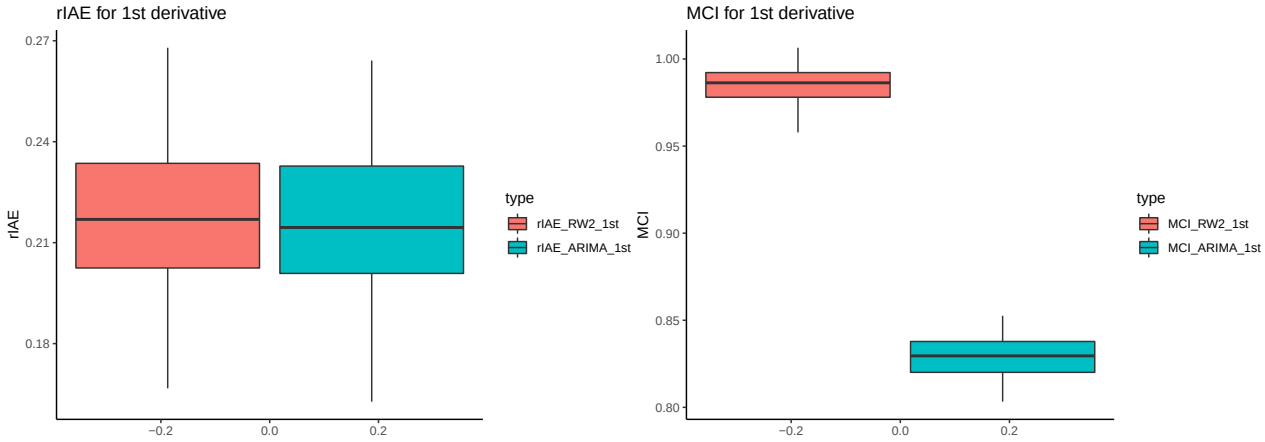

For the second derivative:



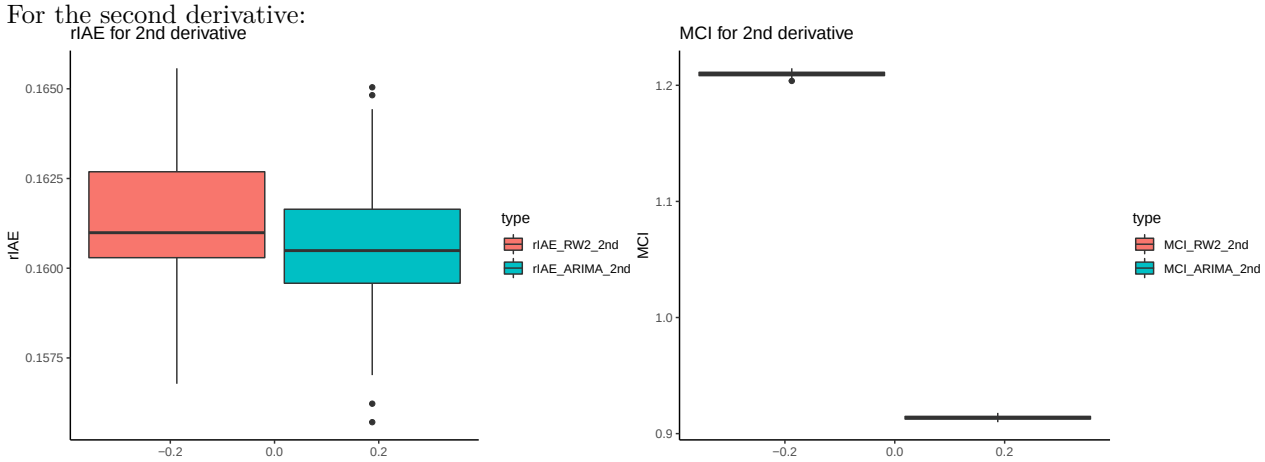

We then consider the second scenario with 50 unique locations and no repeated measurements. For the function values:



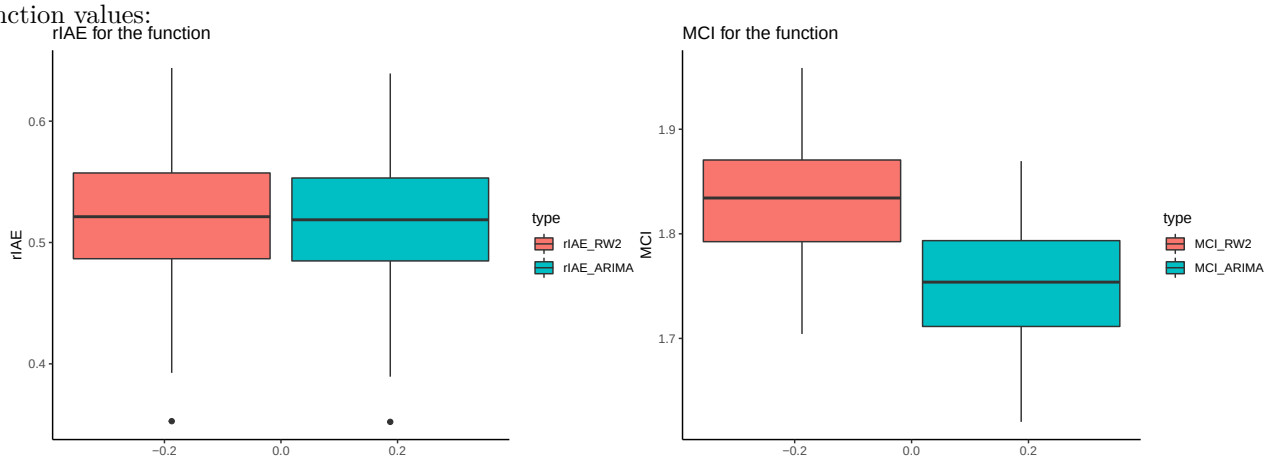

For the first derivative:



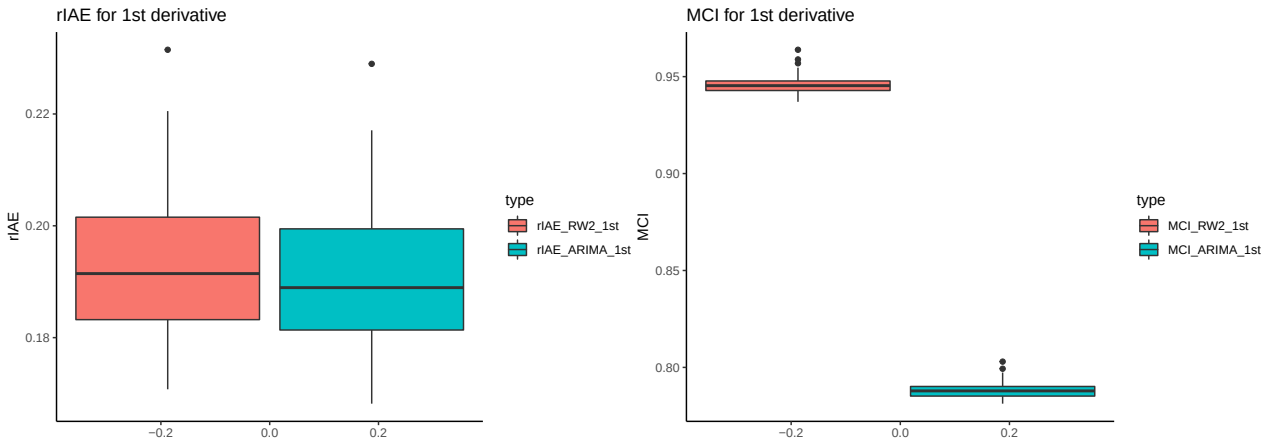

For the second derivative:



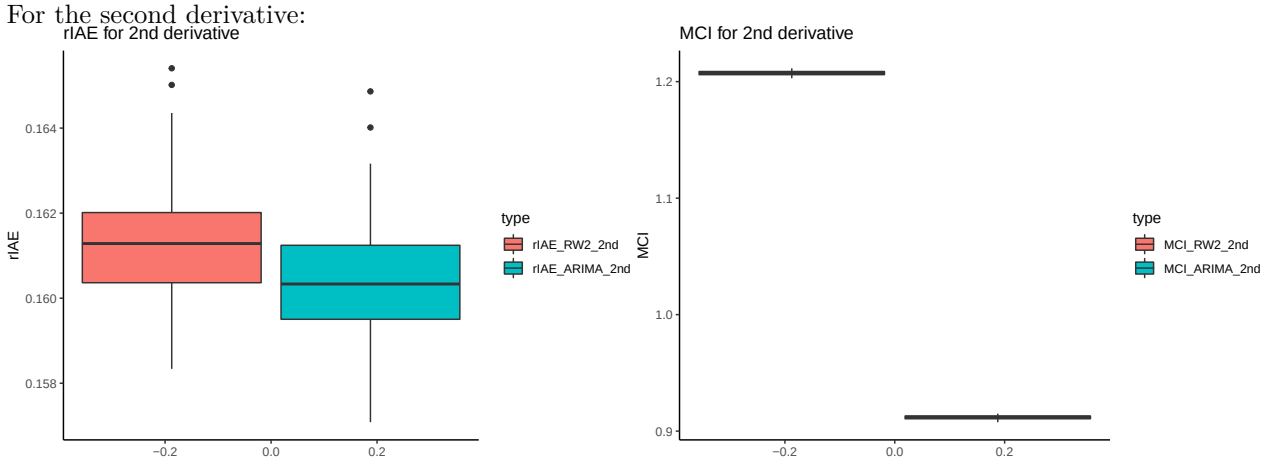

We then consider the third scenario with 50 unique locations and 5 repeated measurements. For the function values:




For the first derivative:




For the second derivative:




### Next step

(Summarize what you currently observed with these cases; what will be done for the next step: 1. Change the nominal rate to 90 percent. 2. Study the point-wise coverage probability (in each simulation: compute the posterior credible interval first, then compute the coverage rate by checking whether each true g(z_i) is contained) 3. Try the simulation with different types of true functions, and see if the performance will differ

## Conclusion:

Based on the results above, it seems like the two methods provide similar point estimate for $g(), g'()$ and $g''()$ in terms of posterior mean. However, the sample paths drawn from different methods have different behaviors, as shown by the difference between mean credible interval width (MCI). In particular, the MCI differs more for the inference of higher order derivatives.