SMMED GP Variable Selection Idea

Kristyn Pantoja

20 February 2020

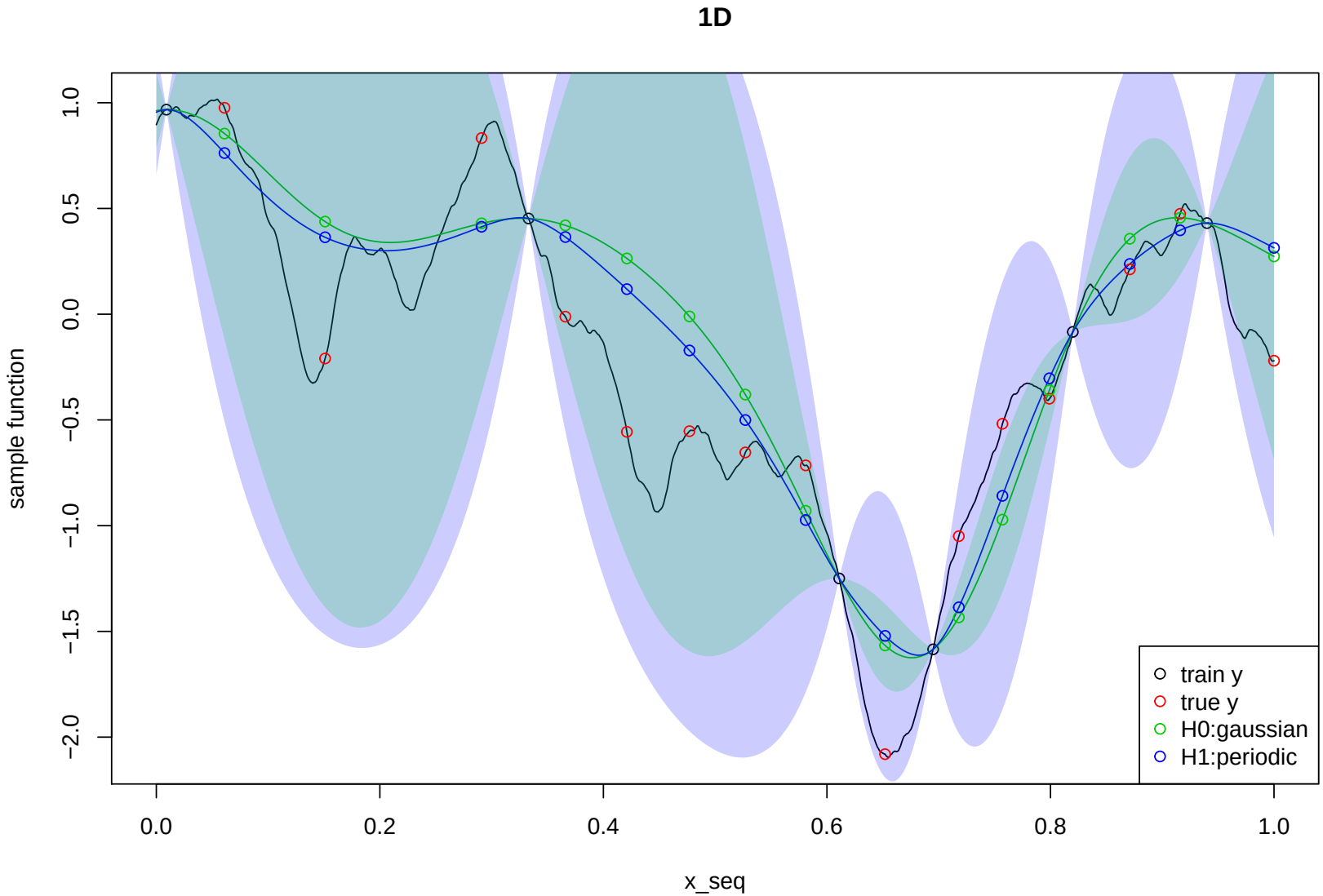
Recall 1D hypothesis tests

Hypothesis tests for variable selection

Next time

Recall 1D hypothesis tests

Comparing 2 kernels in 1D



Hypothesis tests for variable selection

Comparing 1 Variable and 2 variable

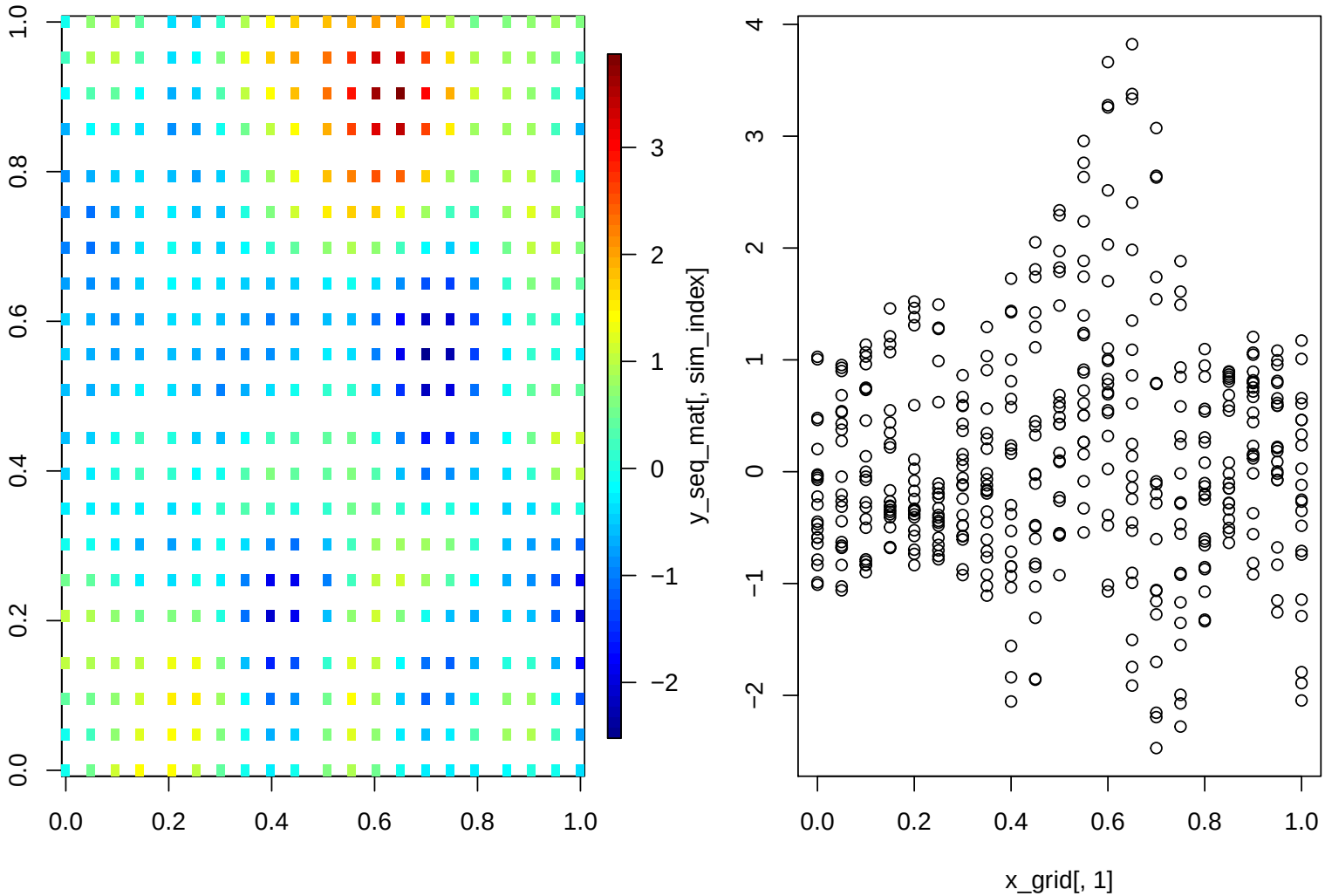
In order to use SMMED for variable selection, I consider the following set-up:

There are 2 variables being considered, X_1 and X_2, but the true function f was generated from a Gaussian kernel in X_1 only. Then the hypotheses are:

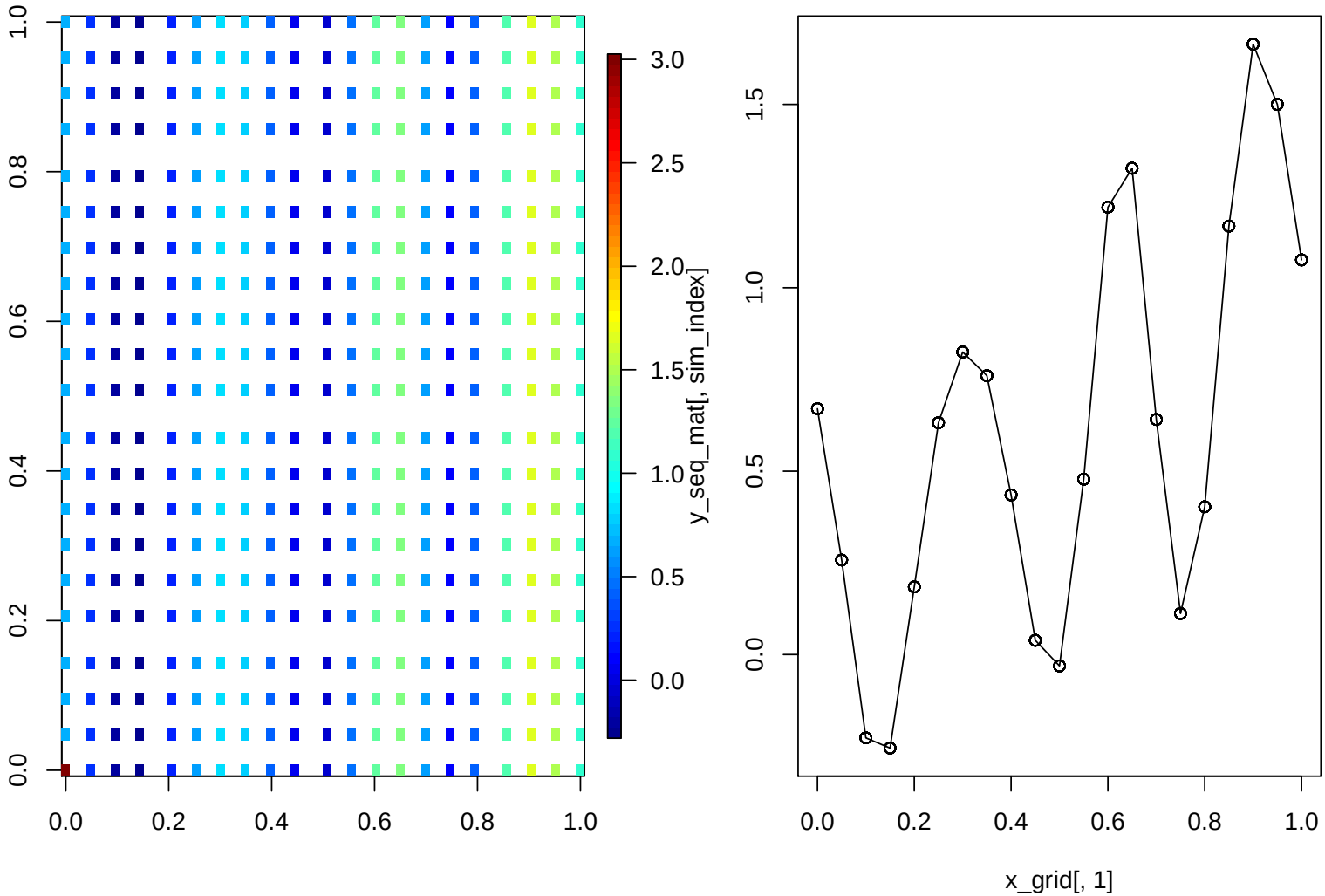
1. f is generated from a Gaussian kernel with 1-dimensional input, $x_i = (x_{i,1})$.
2. f is generated from a Gaussian kernel with 2-dimensional inputs, $x_i = (x_{i,1}, x_{i,2})^T$, where $x_{i,j}$ is a value of the jth variable, $j = 1, 2$.

To apply SMMED, I consider the Gaussian kernel with 2-dimensional inputs and I also consider the Gaussian kernel with the 1st variable only (by stripping the observed data's input x_i of its second variable) and apply the Wasserstein distance for each x like usual.

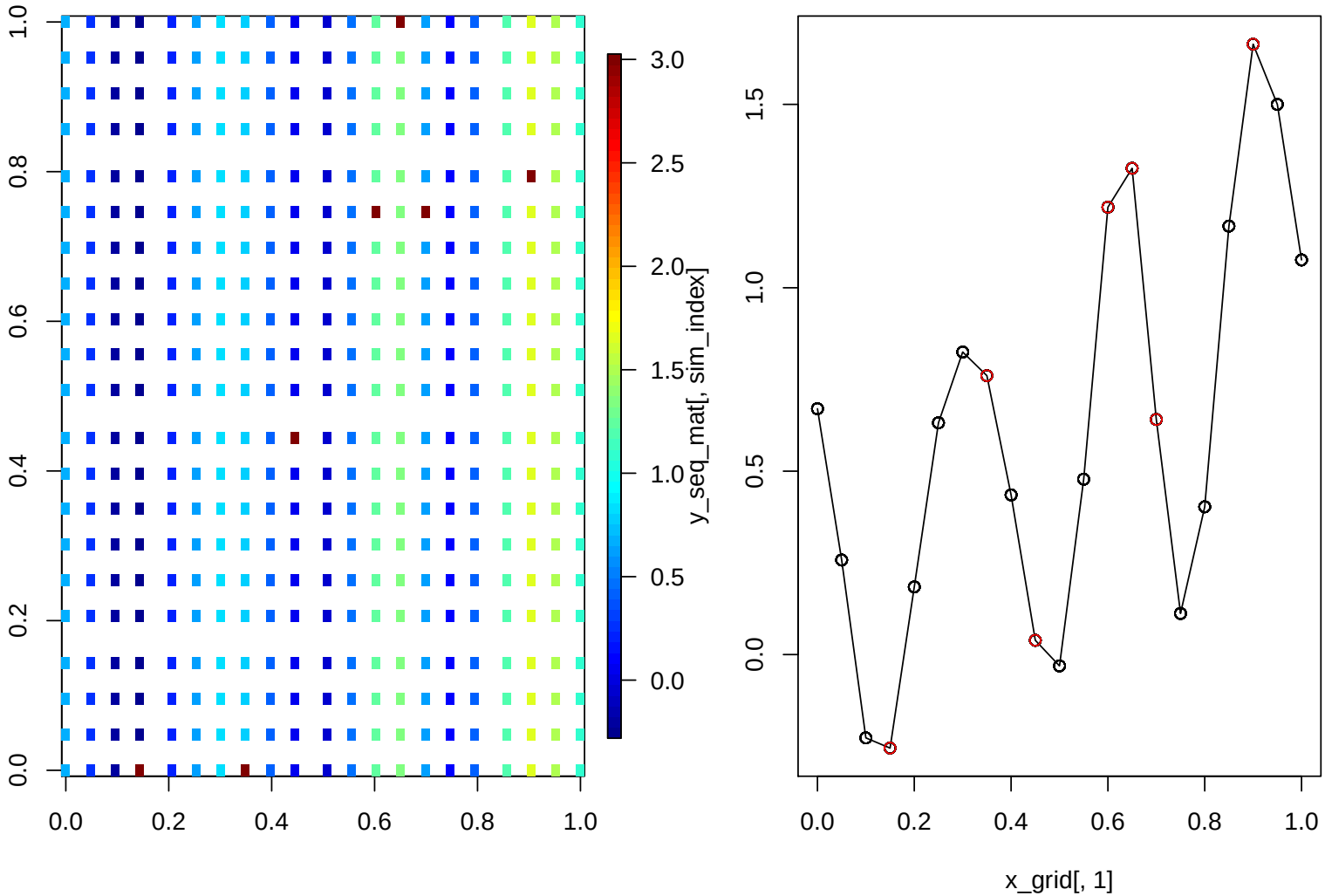
Generating from 2 dimensions



Generating from 1 dimension



MMED for 1 vs 2 dimensions (true f is 1D)



Input Points and SMMED

```
x_train
```

```
##     Var1 Var2
## 42  1.00 0.05
## 52  0.45 0.10
## 76  0.60 0.15
## 98  0.65 0.20
## 109 0.15 0.25
## 138 0.55 0.30
```

```
mmed_gp2d$add
```

```
##      [,1] [,2]
## [1,] 0.65 1.00
## [2,] 0.15 0.15
## [3,] 0.45 0.45
## [4,] 0.90 0.90
## [5,] 0.60 0.60
## [6,] 0.35 0.35
## [7,] 0.70 0.70
```

SMMED has determined that the best way to distinguish the two models (one variable vs. two variables) is to test points that have the same values first variable as some of the initial points, but different values in the second variable, in order to see if there is a change in y and to determine whether the second variable is useful or not.

Next time

Next Time

1. Add noise to be able to pick points that have the same input, but different output.
   - specifically, this is a problem when stripping variables from data where the leftover variables have the same values, and the output is the same, leading to a covariance that is not invertible)
2. More dimensions! Maybe, 3 dimensional function.
   - for more than 2 dimensions, try a bottom-up method where we start with the best variable (in terms of some metric - one that gives highest posterior probability to the true model? or is that pre-emptive to choosing the correct model?)
   - in this case, we would do pairwise comparisons, or we could simply test to see if at least one of the two other variables is useful.
   - if there is another way, try to think of it because this is not exactly ideal.
3. Design to help LASSO
   - compare LASSO model (for some fixed lambda) to saturated model, see if we can determine the points that help LASSO find the true, smaller model.