ST433/533 – Mid-term Exam 2 – Due 10/14

Geostatistical methods based on the multivariate normal distribution are computationally intensive for large datasets. Over the last 5-10 years several new methods have been proposed to overcome this limitation. The paper by Heaton et al (JABES; 2019)

`https://link.springer.com/article/10.1007/s13253-018-00348-w`

reviews these methods and the associated github page

`https://github.com/finnlindgren/heatoncomparison/tree/master/Code`

provides code to implement each method. In this exam, teams of three (table below) will review and illustrate these methods for the class.

We will use the LandSAT data to illustrate the methods (the data are summarized here). The three relevant variables (long, lat, and the response) are available for download at

`https://www4.stat.ncsu.edu/~bjreich/st533/E2_533_2020.csv`

These data have $n = 937,208$ observations but we will use only a subset. Use the following code to create the subset

```
> set.seed(919)
> n       <- 937208
> pool    <- sample(1:n,21000,replace=FALSE)
> group   <- rep(1:21,1000)
> g       <- rep(0,n)
> g[pool] <- group
> table(g)
```

```
0      1      2      3      4      5      6      7      8      9     10
916208 1000   1000   1000   1000   1000   1000   1000   1000   1000    1000
11     12     13     14     15     16     17     18     19     20     21
1000   1000   1000   1000   1000   1000   1000   1000   1000   1000    1000
```

The 916,208 observations with $g = 0$ are completely discarded from the analysis. The 1,000 observations with $g = 21$ are the test set for all Kriging predictions. The 20,000 observations with $g \in \{1, 2, ..., 20\}$ are used to construct training datasets of increasing size. You will first fit the model using observations with $g = 1$ and make predictions for the observations with $g = 21$, then fit the model using observations with $g \in \{1, 2\}$ and make predictions for the observations with $g = 21$ and so on until the final fit that uses the full training data of size 20,000 to make predictions for the observations with $g = 21$. You will perform these computations with the method you are assigned (table below) and standard MLE/Kriging. You can kill any jobs that run for more than an hour.

The final product is a 10-15 minute presentation. Your presentation MUST be broken into the following sections of 3-5 slides each:

1. **Description of the method**: Describe the approximation method. You do not need to review standard MLE/Kriging methods. Your review must include at least one of each of the following slide types: text slide to give the motivation/intuition behind the method; formula slide to pin down the details; a graphic to illustrate the method (not a plot of data, but the method, e.g., a plot of the partitions, knots, neighborhood structure, etc). Remember, your goal is to teach the other students in the class how to use the method, so pick an appropriate level of technical detail (not too high, not too low).

2. **Implementation details**: Describe how you implemented the methods (e.g., which R package you used) and any important details needed to apply the method. For example, all of the methods have tuning parameters (size of partitions, number of MCMC iterations, number of neighbors/knots) and you should discuss their role in model fit versus speed and how to select them for a given analysis.

3. **Results**: Compare your method with MLE in terms of both computing time and prediction performance across various sizes of the training set. You may consider multiple versions of your method defined by different values of the tuning parameters.

You will present your results to the class over zoom during lab on Wednesday 10/14. Email your slides to the instructor (bjreich@ncsu.edu) by midnight on the 13th. Teams must also submit commented code used to implement their analysis. Each member of the team must present one of the three sections above. Please put the presenter's name on each slide. You are responsible for the material on your slides, and your individual grade will be weighted by the material and presentation of your slides.

ID	Team	Method (section number)
1	Fidan, Foraker, H Jiang	Predictive processes (2.1.2)
2	Z Jiang, Schappe, Zheng	Predictive processes (2.1.2)
3	Wiecha, Yang, Yao	Spatial partitioning (2.2.1)
4	Carbajal Carrasco, Goodman, Hutchens	Spatial partitioning (2.2.1)
5	Bai, Fan, Pammer	LatticeKrige (2.3.1)
6	Colonnese, Freedman, Tabor	LatticeKrige (2.3.1)
7	Hasnat, Rangwala, Watson	Stochastic PDEs (2.3.3)
8	Adams, Peng, Wang	Stochastic PDEs (2.3.3)
9	Dixit, Harris, Pena	Nearest neighbor Gaussian process (2.3.4)
10	Schulte, Turner, Xu	Nearest neighbor Gaussian process (2.3.4)