Data Generation

Jayjay, Tuna, Jason, Richard

2024-09-30

# Surrogate Modeling for Which System?

1. Simplified Geological Carbon Storage (Francis’ paper)
2. Incompressible Navier Stokes

# Twophase flow for the CO2 saturation

* We regenerate Francis’ dataset, and additionally compute Fisher Information Matrix as well.
* For the purpose of validation, we currently form full Fisher Infromation Matrix and then compute eigenvector.
* Our next step will be low rank approximation or trace estimation so that we don’t have to form the full matrix.

## Dataset

Our dataset consists of pairs of .

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) K0 | |  |

|  |  |
| --- | --- |
| |  | | --- | | (b) K1 | |

Figure 1: Example Permeability Model

|  |  |  |
| --- | --- | --- |
| |  | | --- | | (a) Time Series of Saturation of K0 | |  |

|  |  |
| --- | --- |
| |  | | --- | | (b) Time Series of Saturation of K1 | |

Figure 2: Example Saturation Time Series

## Fisher Information Matrix

* To find the optimal number of observations, , we visualize eigenvector and vector jacobian product.
* We observe that as increases, the clearer we see the boundary of the permeabiltiy, which will be more informative during training and inference. [[1]](#footnote-40)
* Given 1 pair of dataset, , we get a single FIM.

### Computing Fisher Information Matrix for each datapoint

We consider a realistic scenario when we only have access to samples, but not distribution. When is number of samples and , neural network model learns mapping from . For each pair of , we generate .

* : number of data points,
* : number of observation,

For a single data pair, we generate multiple observations.

As we assumed Gaussian, we define likelihood as following.

A FIM for a single data pair is:

### How does FIM change as number of observation increases?

FIM is expectation of covariance of derivative of log likelihood. As we expected, we see clearer definition in diagonal relationship as increases.

|  |  |  |
| --- | --- | --- |
| M = 1  M = 1 | M = 10  M = 10 | M = 100  M = 100 |

Figure 3: Change in FIM[:256, :256] of single data pair as number of observation, increases

### Making Sense of FIM obtained

Still, does our FIM make sense? How can we better understand what FIM is representing?

Let’s look at the first row of the FIM and reshape it to [64, 64].

|  |  |  |
| --- | --- | --- |
| FIM[0,:]  FIM[0,:] | FIM[1,:]  FIM[1,:] | FIM[2,:]  FIM[2,:] |

Figure 4: Fist, Second, and Third row in FIM

* Like we expected from the definition of FIM, we observe each plot is just different linear transformation of
* As we will see from below, each rows in FIM is noisy version of its eigenvector.

### How does eigenvectors of FIM look like as increases?

#### (Single Observation)

|  |  |  |
| --- | --- | --- |
| First Eigenvector  First Eigenvector | Second Eigenvector  Second Eigenvector | Third Eigenvector  Third Eigenvector |

Figure 5: First three largest eigenvector of FIM

* Even when FIM is computed with single observation, we see that the largest eigenvector has the most definition in the shape of permeability. Rest of eigenvector looks more like noise.

|  |  |  |
| --- | --- | --- |
| First Eigenvector  First Eigenvector | Second Eigenvector  Second Eigenvector | Third Eigenvector  Third Eigenvector |

Figure 6: First three largest eigenvector of FIM

|  |  |  |
| --- | --- | --- |
| First Eigenvector  First Eigenvector | Second Eigenvector  Second Eigenvector | Third Eigenvector  Third Eigenvector |

Figure 7: First three largest eigenvector of FIM

|  |  |  |
| --- | --- | --- |
| First Eigenvector  First Eigenvector | Second Eigenvector  Second Eigenvector | Third Eigenvector  Third Eigenvector |

Figure 8: First three largest eigenvector of FIM

* As increases, we observe flow through the channel clearer.
* We see the boundary of permeability gets clearer.
* In general, it gets less noisy.

### How does vector Jacobian product look like as increases?

|  |  |
| --- | --- |
| vjp (M=1)  vjp () | vjp (M=10)  vjp () |

|  |  |
| --- | --- |
| vjp (M=100)  vjp () | vjp (M=1000)  vjp () |

Figure 9: Normalized Vector Jacobian Product when vector is the largest eigenvector

* We observe that vector Jacobian product looks more like saturation rather than permeability.
* As increases, scale in color bar also increases.
* One possible conclusion:
  + vjp tells us the location in the spatial distribution (likelihood space) where there exists the largest variation, thus have the most information on parameter.
  + , when is the largest eigenvector of FIM, is projecting Jacobian onto direction of maximum sensitivity.

# Incompressible Navier Stokes

## Dataset

|  |  |
| --- | --- |
| Vorticity at t=0  Vorticity at | Vorticity at t=40  Vorticity at |

Figure 10: The first and the last vorticity in a single time series

Our dataset consists of 50 pairs of , where . Initial vorticities are a Gaussian Random Fields.

## Fisher Information Matrix

### How do we compute FIM?

* Just means that we are computing FIM with respect to the initial vorticity, .

### How does FIM looks like as changes?

|  |  |
| --- | --- |
| M=10 | M=100 |

Figure 11: FIM[:100, :100] of varying

### Making Sense of FIM obtained

Still, does our FIM make sense? How can we better understand what FIM is representing?

Let’s look at the first row of the Fisher Information Matrix and reshape it to [64,64].

|  |  |
| --- | --- |
| FIM[0, :]  FIM[0, :] | Input Vorticity  Input Vorticity |

Figure 12: Comparison of the input parameter with the first element of FIM

Also, let’s look at how the first row of the FIM changes as time evolves. When ,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| t=1 | t=5 | t=10 | |  | | --- | |  | | |  | | --- | |  | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | |  | | |  | | --- | |  | | |  | | --- | |  | | |  | | --- | |  | | |  | | --- | |  | |

Figure 13: The evolution of the first row of FIM

## Future Step

1. TODO: Debug NS eigenvector and vjp.
2. TODO: Want to generate the full dataset for Francis’ dataset (which might take 1 or 2 days).
3. TODO: Try it on Jason’s dataset (Now that we fixed the problem with FIM computation, we are optimistic about the experiment, so we want to try it again.)

## Question

1. What would be the optimal number for observations, when computing Fisher Information Matrix?

1. [Note on Learning Problem](https://www.overleaf.com/1149716711hxnvfbyfpzvb#a799ce). [↑](#footnote-ref-40)