DATA GENERATION

A PREPRINT

Jayjay, Tuna, Jason, Richard

2024-09-30

1 Surrogate Modeling for Which System?

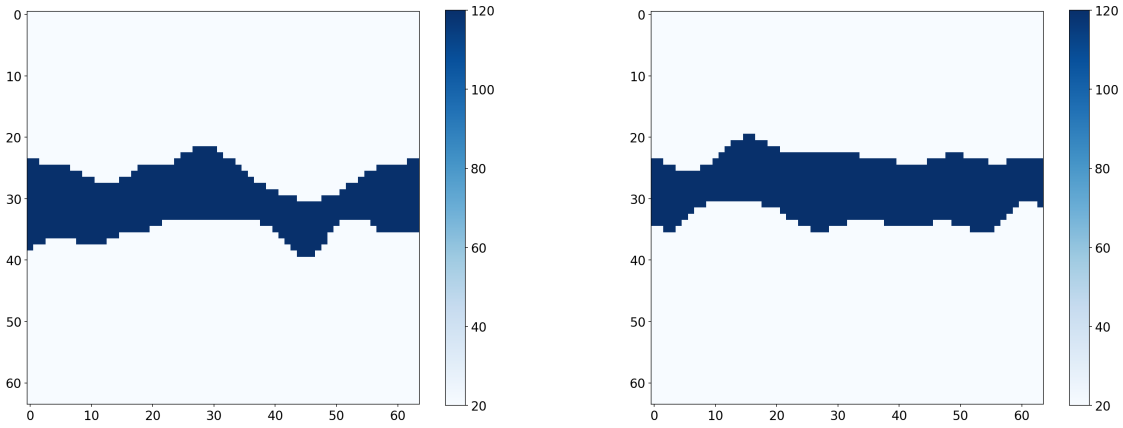
1. Simplified Geological Carbon Storage (Francis' paper)
2. Incompressible Navier Stokes

2 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.

2.1 Dataset

Our dataset consists of 2000 pairs of $\{K, S^t(K)\}_{t=1}^8$.



(a) K0 (b) K1

Figure 1: Example Permeability Model

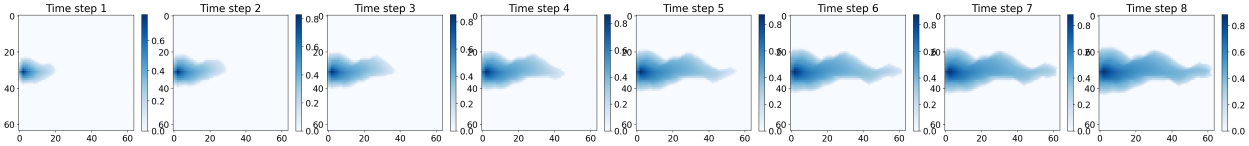
2.2 Fisher Information Matrix

- To find the optimal number of observations, M, we visualize eigenvector and vector jacobian product.
- We observe that as M increases, the clearer we see the boundary of the permeabiltiy, which will be more informative during training and inference. [1]
- Given 1 pair of dataset, $\{K, S^t(K)\}_{t=1}^8$, we get a single FIM.

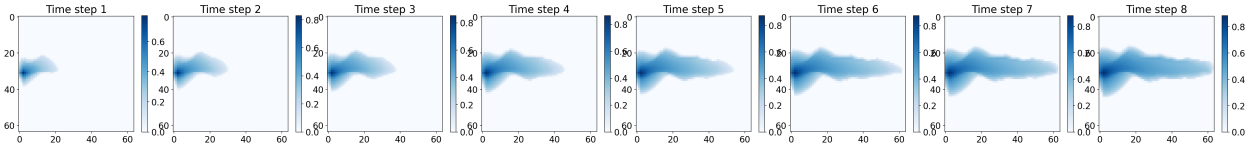
[1] Note on Learning Problem.

2.2.1 Computing Fisher Information Matrix for each datapoint

We consider a realistic scenario when we only have access to samples, but not distribution. When N is number of samples and $X \in \mathbb{R}^{d \times d}$, neural network model F_{nn} learns mapping from $X_i \to Y_i$. For each pair of $\{X_i, Y_i\}_{i=1}^N$, we generate $\{FIM_i\}_{i=1}^N$.



(a) Time Series of Saturation of K0



(b) Time Series of Saturation of K1

Figure 2: Example Saturation Time Series

- N : number of data points, $\{X_i, Y_i\}$
- M : number of observation, Y

$$\{X_i\}_{i=1}^{N} \sim p_X(X),\ \epsilon \sim \mathcal{N}(0, \Sigma),\ \Sigma = I$$

For a single data pair, we generate multiple observations.

$$Y_{i,J} = F(X_i) + \epsilon_{i,J}, \quad where\ \{\epsilon_{i,J}\}_{i,J=1,1}^{N,M}$$

As we assumed Gaussian, we define likelihood as following.

$$p(Y_{i,J}|X_i) = e^{-\frac{1}{2}\|Y_{i,J} - F(X_i)\|_2^2}$$

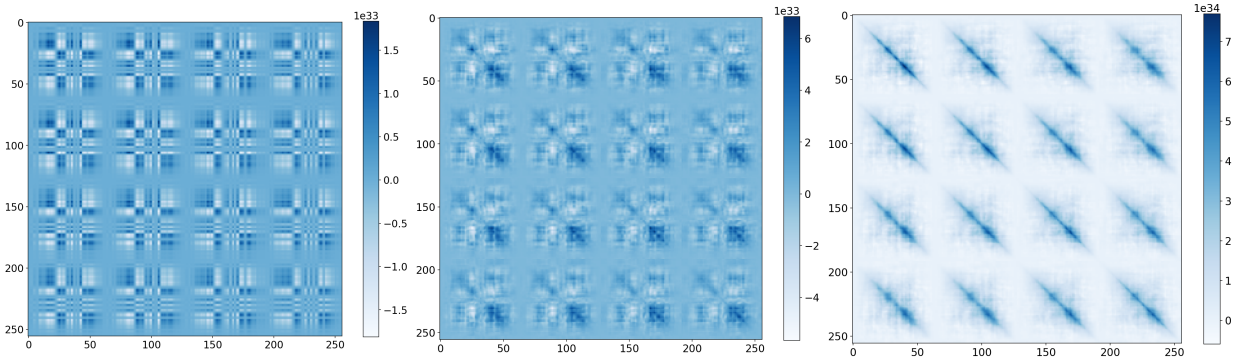
$$log\, p(Y_{i,J}|X_i) \approx \frac{1}{\Sigma}\|Y_{i,J} - F(X_i)\|_2^2$$

A FIM for a single data pair i is:

$$FIM_i = \mathbb{E}_{Y_{i,\{J\}_{i=1}^{m}} \sim p(Y_{i,J}|X_i)} \left[\left(\nabla log\, p(Y_{i,J}|X_i)\right) \left(\nabla log\, p(Y_{i,J}|X_i)\right)^T \right]$$

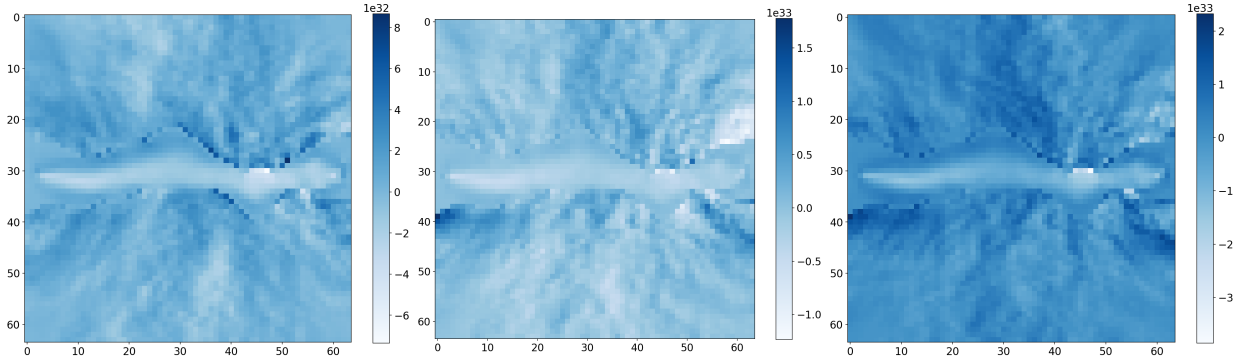
2.2.2 How does FIM change as number of observation increases?

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



(a) M = 1 (b) M = 10 (c) M = 100

Figure 3: Change in FIM[:256, :256] of single data pair $\{K, S^t(K)\}_{t=1}^{8}$ as number of observation, M increases



(a) FIM[0,:] (b) FIM[1,:] (c) FIM[2,:]

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

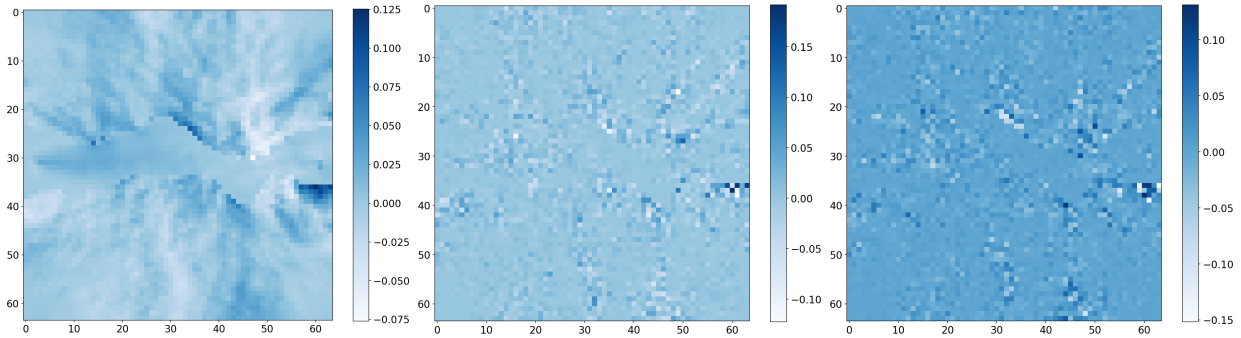
2.2.3 Making Sense of FIM obtained

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].

- Like we expected from the definition of FIM, we observe each plot is just different linear transformation of $\nabla log p(\{S^t\}_{t=1}^8|K)$
- As we will see from below, each rows in FIM is noisy version of its eigenvector.

2.2.4 How does eigenvectors of FIM look like as M increases?



(a) First Eigenvector (b) Second Eigenvector (c) Third Eigenvector

Figure 5: First three largest eigenvector of FIM

2.2.4.1 $M = 1$ (Single Observation)

- 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.

2.2.4.2 $M = 10$

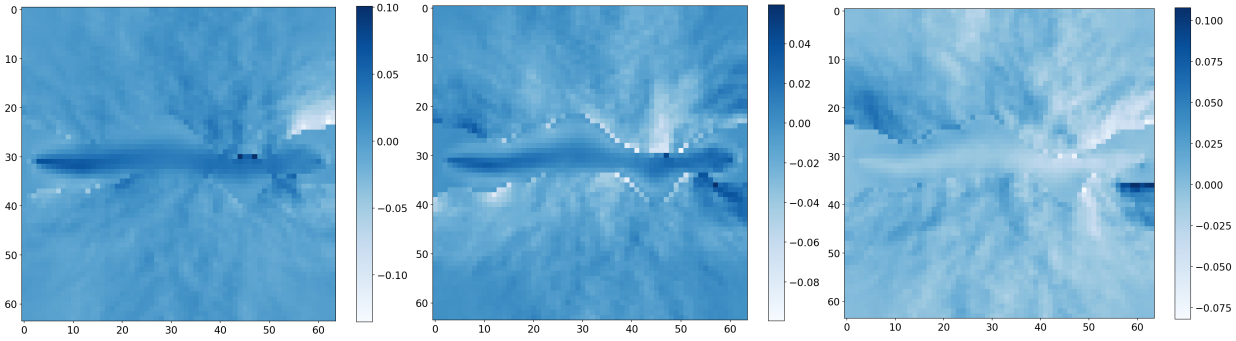
2.2.4.3 $M = 100$

2.2.4.4 $M = 1000$

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

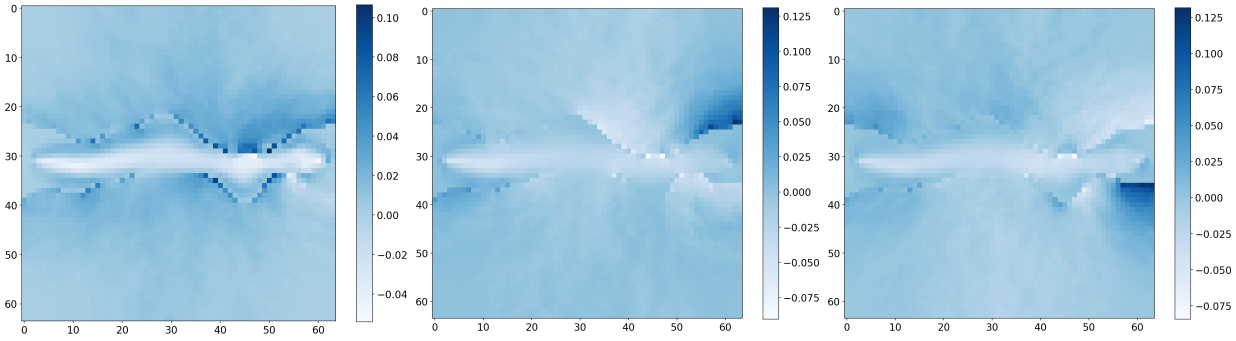
2.2.5 How does vector Jacobian product look like as M increases?

- We observe that vector Jacobian product looks more like saturation rather than permeability.



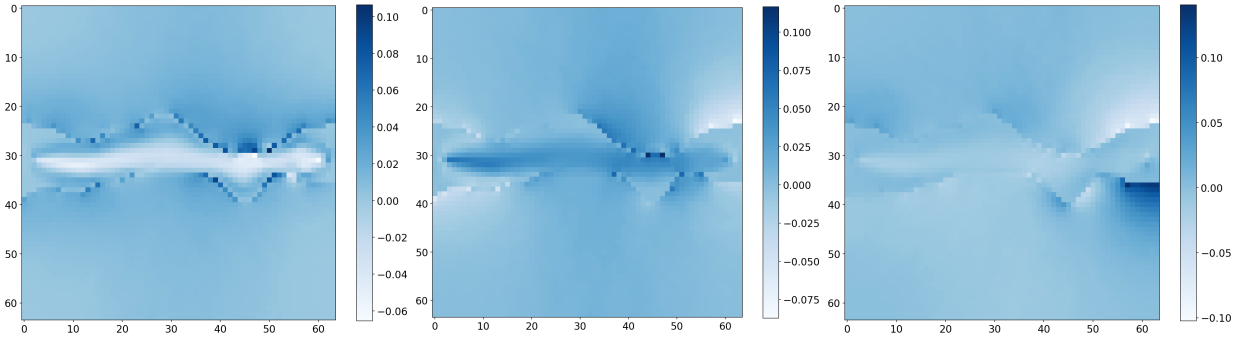
(a) First Eigenvector (b) Second Eigenvector (c) Third Eigenvector

Figure 6: First three largest eigenvector of FIM



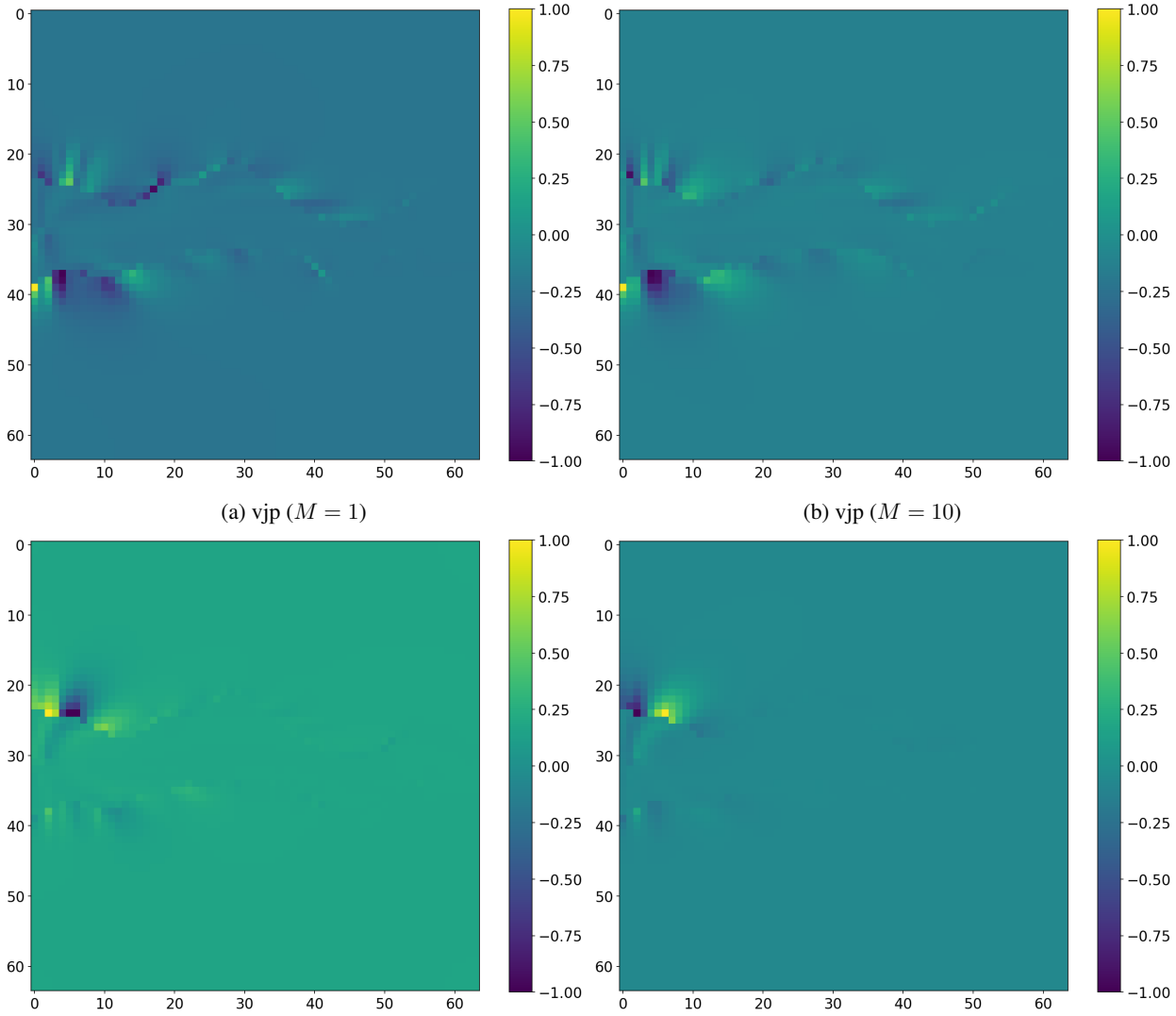
(a) First Eigenvector (b) Second Eigenvector (c) Third Eigenvector

Figure 7: First three largest eigenvector of FIM



(a) First Eigenvector (b) Second Eigenvector (c) Third Eigenvector

Figure 8: First three largest eigenvector of FIM



(a) vjp ($M = 1$)

(b) vjp ($M = 10$)

(c) vjp ($M = 100$)

(d) vjp ($M = 1000$)

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

- As M 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.
  - $J^T v$, when v is the largest eigenvector of FIM, is projecting Jacobian onto direction of maximum sensitivity.

3 Incompressible Navier Stokes

3.1 Dataset

Our dataset consists of 50 pairs of $\{\varphi^{t-1}(x_0), \varphi^t(x_0)\}_{t=1}^T$, where $T = 44$. Initial vorticities are a Gaussian Random Fields.

3.2 Fisher Information Matrix

3.2.1 How do we compute FIM?

$FIM = \left(\nabla logp(\varphi^t(x_0)|\varphi^0(x_0))\right)\left(\nabla logp(\varphi^t(x_0)|\varphi^0(x_0))\right)^T$