There are 6 files:

files “SGP\_x\_20180418\_\*” and “SGP\_y\_20180418\_\*”

\* can be “mean”, “random” and “gaussian”.

The ‘x’ and ‘y’ files with the same suffix are corresponding to each other. So totally there are 3 data sets to use.

Different suffix means different data generating process, and I think for algorithm test purpose, you can ignore the detail of this.

All x files have shape (3, 750000), so it has three variables at each time step.

All y files have shape (750000,) the same as each variable in file x.

It’s the label. For example, if the first 10 elements of y are “1”, and the next 10 elements are “3”, it means the first 10 time steps of x (3 variables for each of the 10 time steps) are from distribution “1” and the next 10 time steps of x are from distribution “3”, so there is a distribution change at time step 11.

You can divide it to test set, validation set and training set according to your idea. One possible way is [0, 500000] for training, [500000, 600000] for validation and [600000, 750000] for test.