On NS dataset

Marina Berbel
(Dated: April 8, 2022)

I. ANALYSIS OF THE RAW DATA

Here we do some analysis of the data, without any treatment, just to explore its features.

A. Error in the recovery

Here we compute the relative error in each quantity as

$$error = \frac{(q_{inj} - q_{rec})}{q_{inj}}. \tag{1}$$

Using this there are division by zero in the spins, so we do not show them. In figure 1 we can see the results for the train and test dataset.

The error is bigger for $m_1 \leq 50$ and $m_2 \leq 20$. One of the meanings of this is that the BNS will be poorly recovered, making it harder to classify them.

COMMENTS ON EFFECTS OF THIS IN THE REGRESSION?

B. Distribution of the label

We can draw in the $m_1 - m_2$ plane the events in the data, colored by their label. This gives us the idea of the function used for labelling and if it will be easy or more difficult to classify the events. Also we can later compare the probability distribution given by the classifier with the actual distribution of the label.

In figure 2 we can see that the event is classified as *HasNS* if $m_2 \leq 3$. In figure 3 we can see that in the data we will use for classification, this is the recovered values of the event from the pipeline, the sharp cut no longer exists because the masses are distorted. We can still observe a dependency on m_2 stronger than on m_1.

The error in the recovery of m_2 will highly affect the classification of events.
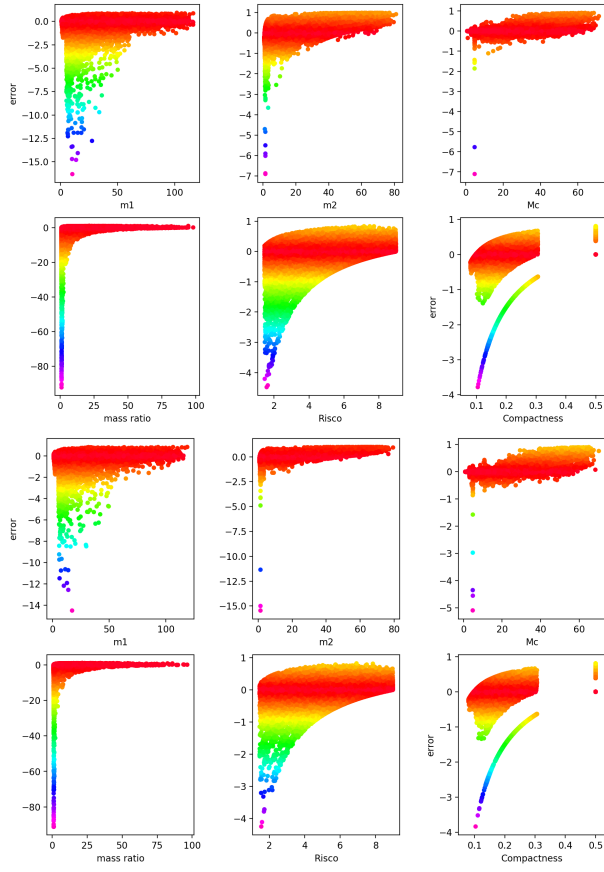
FIG. 1. Relative error in the recovery of each quantity, independent and derived. Top picture: in train dataset. Bottom picture: in test dataset.
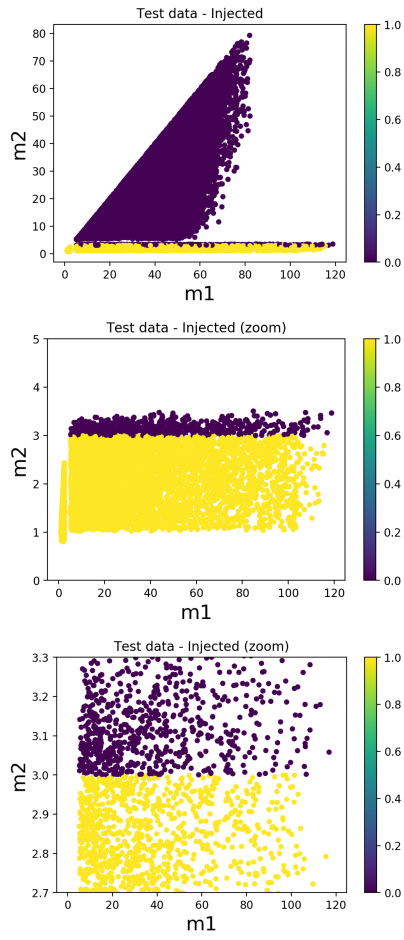
FIG. 2. From the test dataset, injected masses colored by their label. Second and third pictures are zooms of the first one. We can see that there is a sharp cut at $m_2 = 3$.
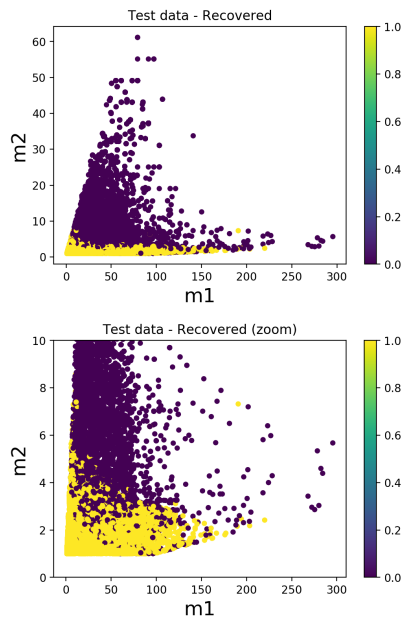
FIG. 3. From the test dataset, recovered masses colored by their label. Second picture is a zoom of the first one. The sharp cut is lost in the error of the recovery.