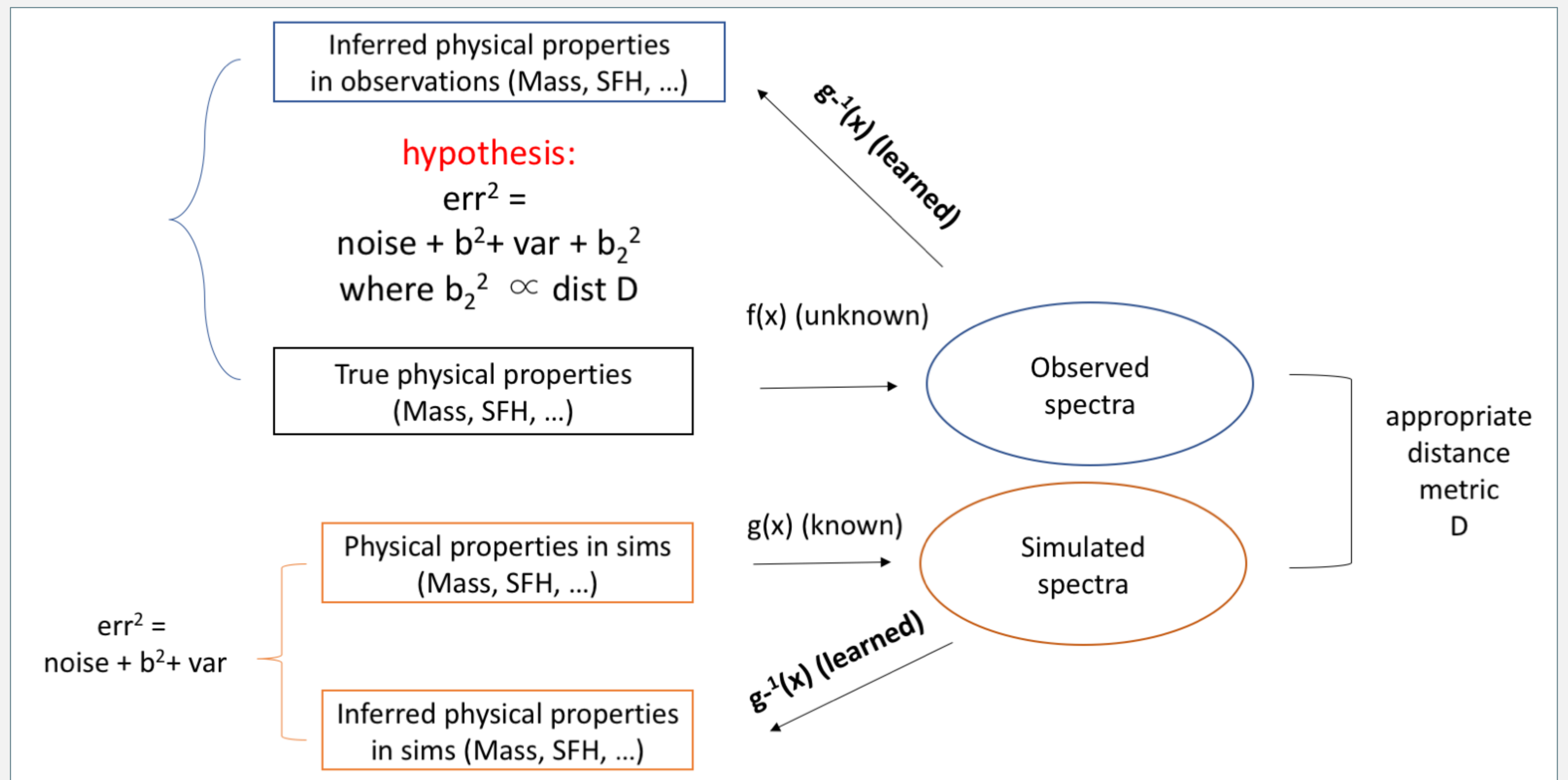
Debunking Generalization Error or: How I Learned to Stop Worrying and Love My Training Set

Viviana Acquaviva (CUNY NYC College of Technology), Christopher C. Lovell (University of Hertfordshire), Emille E. O. Ishida (Université Clermont Auvergne)

We aim to determine some physical properties of distant galaxies (for example, stellar mass, star formation history, or chemical enrichment history) from their observed spectra. Unfortunately, identifying a training set for this problem is very hard, because labels are not readily available - we have no way of knowing the true history of how galaxies have formed. One possible approach to this problem is to train machine learning models on state-of-the-art cosmological simulations; however, it is unclear how models will perform once applied to real data. In this paper, we attempt to model the generalization error as a function of an appropriate measure of distance between the source domain and the application domain. Our goal is to obtain a reliable estimate of how a model trained on simulations might behave on data.

vacquaviva@citytech.cuny.edu vacquaviva @AstroVivi

Framework



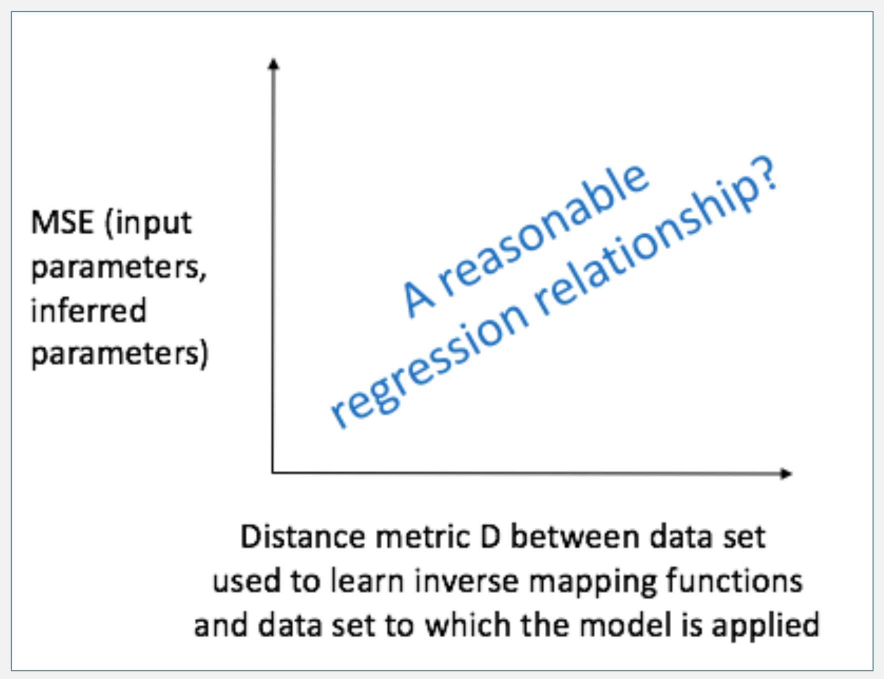
The mapping of the physical properties of galaxies (stellar mass, star formation history…) to observed quantities (spectra) is described in nature by a function f(x). **We are interested in learning its inverse f^{-1}**, which allows us to estimate the properties of galaxies from data.
In simulations, the direct function is g(x). We can verify that g(x) is a good approximation of f(x) by using the distance in spectral space as a proxy for the similarity between f(x) and g(x).

Now let us consider the other direction. The function g−1 can be learned by, for example, training a machine learning model. What happens if we apply the learned function g−1 *to the observed spectra* (in other words, when we use it a proxy for the function we want, f^{-1})?

There will be an additional term in the square error, which comes from the fact that we learned the "wrong" function, g^{-1} instead of f^{-1}. Our hypothesis is that the additional error will depend in a ***predictable*** way on an appropriate distance metric describing the similarity between the observed spectra and the simulated spectra. If we can show that this is true, **we have a way of predicting the generalization error on data.**

Methodology

- We generate 20 sets of simulations, changing the modeling assumptions, and find a suitable representation feature space for all the simulations sets
- We find an appropriate measure of distance between data sets (Di,j where i, j ∈ [1, 20]);
- We train 20 models, one per simulated set of spectra, excluding the objects who participated to the feature selection process, to learn as many inverse modeling functions (g^{-1}_i);
- We obtain 20 scatter plots of the distance metric versus the generalization error obtained by applying the 20 functions $g^{-1}_1, g^{-1}_2, \dots g^{-1}_{20} \dots$ to each simulation set i;
- We use these 20 examples to infer a robust regression between the distance metric Di,j mentioned above and the generalization error incurred, and populate this plot:

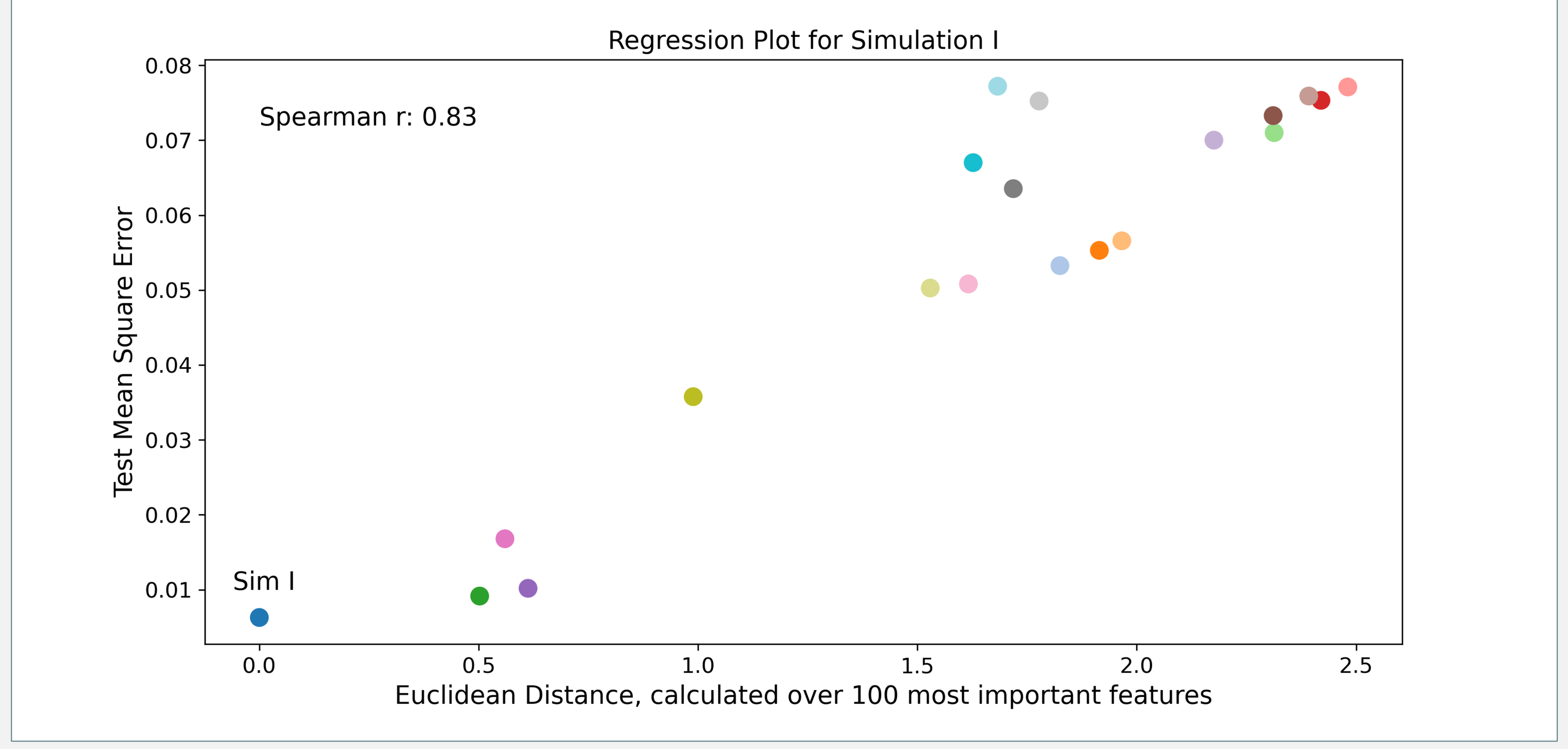


If the regression relationship is sufficiently tight, **we can use it to predict the generalization error on data, based on the distance between data and simulations.**

References

1. This paper is available at https://arxiv.org/abs/2012.00066
2. Our "star formation histories with CNNs paper (Lovell et al, MNRAS Vol. 490, 2019) is at https://academic.oup.com/mnras/article/490/4/5503/5586582

Preliminary Results / Next Steps



Our chosen metric is the mean Euclidean distance in the space of 100 most important features, ranked by running a robust Random Forest model on a superset that combines examples from all the 20 simulations. Objects in this superset are excluded from further processing.

We show one example plot where the "target" set of spectra is simulation 1, and we show the MSE that we obtain when we apply the 20 learned inverse functions to recover the stellar mass. There is a clear trend that suggest the possibility of fitting the regression successfully. The trends seen here are similar to what we observe in the other 19 plots. Next steps include:

- Refining our feature selection technique, e.g. by clustering highly correlated features and selecting one per cluster;
- Understanding "failing" cases, such as outliers in our distance/generalization error regressions;
- Using Convolutional Neural Networks, which have better generalization properties than tree-based methods, to derive the MSE used in these plots.
- Extending the framework to other tasks, such as inferring dust properties or star formation histories.