Dr. Janet Kelso and Dr. Alfonso Valencia
Editors-in-Chief
Bioinformatics

November 1, 2023

Dear Editors-in-Chief,

We wish to submit an original research article titled *Avoiding overfitting during the calibration of perplexity (t-SNE)* for consideration by *Bioinformatics*.

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere. All authors have approved the manuscript and agree with its submission to *Bioinformatics*.

In this paper, we discuss the overfitting problem in the context of dimension reduction and demonstrate how its ignorance leads to miscalibration of t-SNE hyperparameters. More specifically, we show low values of perplexity tend to overfit the data.

Overfitting is heavily studied within the context of other modeling problems, but entirely ignored when discussing dimension reduction. Previous works disregard high-dimensional data's composition of signal and noise, so they assess dimension reduction techniques on their ability to capture both the signal and the noise. We provide a framework that assesses dimension reduction techniques on their ability to capture just the signal.

The discovery of new technologies has given rise to an influx of high-dimensional data. In the word of microbiology and single-cell transcriptomics, for example, the data are notoriously complex and noisy. This has led to an increase in popularity of powerful nonlinear dimension reduction techniques like t-SNE and UMAP. When used incorrectly, however, these methods are known to produce unfaithful results. Our framework provides an effective way to calibrate these methods in the presence of noise.

We have no conflicts of interest to disclose. Thank you for your consideration of this manuscript.

Sincerely,

Justin Lin
PhD Candidate, Department of Mathematics
Indiana University
linjus@iu.edu

Julia Fukuyama,
Assistant Professor, Department of Statistics
Indiana University
jfukuyam@iu.edu