**\*\*Explain the purpose of t-SNE and how it is used, including an example to illustrate its use and interpretation. [1 page plus figure(s)]\*\*:**

In statistics, high-dimensional data problems often involve situations where there are more features in the data than there are observations (the $p>n$ problem), or even when there are enough observations, most of the variables contribute noise rather than signal to the process. Dimension-reduction techniques focus on identification of a “smaller fundamental set of independent variables… which determine the values of the original p variables”(Hotelling, 1933). This concept that a narrower set of $k<p$ features that may contain the majority of the information in the original p features drove the development of PCA, MDS, Sammon Mapping, and other multivariate mapping techniques. High-dimensional data visualization, on the other hand, seeks to find ways to visualize more than 2 or 3 dimensions of data without fundamentally changing the underlying scales of interest; a good example of this idea can be found in Chernoff faces (Chernoff, 1973).

T-distributed Stochastic Neighbor Embedding (t-SNE) seeks to solve both problems simultaneously by identifying a reduction of the data to 2 (or 3) dimensions and producing a 2 (or 3)-dimensional map of the original data, with meaningful groupings from the high-dimensional space preserved in the low dimension. In plain English, if observations are similar when using all the features, they should be mapped close together in the resulting t-SNE map. More technically, t-SNE minimizes the symmetricized Kullback-Leibler divergence between two joint probability distributions, P\_{ij} and Q\_{ij}. Without getting too technical here, P can be interpreted as a joint probability (using the full feature set) that a given observation (i) would choose another observation (j) as a neighbor, under the assumption that P is distributed normally and centered at i. The Q similarity operates in the same way, except that it uses a heavier-tailed t-distribution with 1 degree of freedom to estimate the joint probability using only a subset of the features. If the lower-dimensional mapping is faithful to the high-dimensional setting (the actual data), the KL-divergence between P and Q will be small (van der Maaten, 2008), so t-SNE seeks to minimize this for all the pairs of datapoints.

T-SNE is surprisingly easy to implement, with only the choice of a perplexity parameter necessary to run. Perplexity can be thought of as a rough estimate of the expected number of neighbors in each group. Van der Maaten (2008) notes that t-SNE is fairly robust to changes in perplexity values; however, I have not found this to be the case. Below, I implement t-SNE with the package Rtsne on a dataset of NBA Rookies drafted in 2017, with a perplexity of 10. I chose 10 because the dataset is relatively small and I expected to see only two groups based on previous research with PCA. It is useful to scale the data prior to running t-SNE (in part because the default setting in Rtsne is to run PCA in the background and then t-SNE the already dimension-reduced scores). The plot of the 2-D t-SNE map is shown below in Figure X. Care should be taken when interpreting these maps; groups of observations have meaning but positions of groups do not. Indeed, based on the random seed chosen, positions of groups may differ substantially across runs.

T-SNE does not offer much in the way of interpretation or creation of new latent scales to assess performance – the axes of variation (X and Y) do not have meaning in an interpretable sense. That being said, the groupings do have meaning. In this case, t-SNE grouped the players who contributed meaningful statistics to their teams in their rookie year in the bottom left corner (note Donovan Mitchell, Jayson Tatum, and Ben Simmons were all in the running for rookie of the year in 2018). Because of the difficulty of interpretation and black-boxed nature of t-SNE, I would not use this method as a replacement for PCA or a latent-variable model. I like t-SNE for exploratory analysis primarily. Additionally, I came across a use case where I had the ability to add additional features to a model but due to software constraints, I could only add a single feature to the model (I had 722 to choose from) – the goal in this case was to increase predictive ability of the model without necessarily creating interpretable features to be used. One use of t-SNE might have been to take the 722 feature set and create a 2-D mapping, and then apply model-based clustering to the map. Those clusters could then be added to the model as an additional single feature that contains as much information as possible from the original 722 feature set. I demonstrate this concept in Figure X.

References

Hotelling, H. (1933). Analysis of a complex statistical variables into principal components. J. Educ. Psychol., 24, 414-441.

[*Herman Chernoff*](https://en.wikipedia.org/wiki/Herman_Chernoff)*(1973).*[*"The Use of Faces to Represent Points in K-Dimensional Space Graphically"*](https://web.archive.org/web/20120415030406/http:/www.apprendre-en-ligne.net/mathematica/3.3/chernoff.pdf)*(PDF). Journal of the American Statistical Association. American Statistical Association.****68****(342): 361–368.*[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier)*:*[*10.2307/2284077*](https://doi.org/10.2307%2F2284077)*.*[*JSTOR*](https://en.wikipedia.org/wiki/JSTOR)[*2284077*](https://www.jstor.org/stable/2284077)*. Archived from*[*the original*](http://www.apprendre-en-ligne.net/mathematica/3.3/chernoff.pdf)*(PDF) on 2012-04-15.*