1 Setup

Suppose we have high-dimensional data $Z \in \mathbb{R}^{n \times p}$ with MDS embedding $X \in \mathbb{R}^{n \times q}$ for $q << p$. We would like to test how various out-of-sample embedding techniques perform when the new point is noised. Let $w + \epsilon \in \mathbb{R}^p$ be a noised out-of-sample point with embedding $y + \epsilon' := \varphi(w + \epsilon) \in \mathbb{R}^q$. Here φ is the embedding function. We want to study the reconstruction error

$$\text{error}_{\text{rec}} = f(w, \varphi(w + \epsilon)),$$

where $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ is a cost function. Given a distribution on ϵ, we can then compute the risk

$$\text{risk} = \int f(w, \varphi(w + \epsilon))\, dp(\epsilon).$$

In practice, φ will not always have a closed form, so will compute empirical risk

$$\hat{\text{risk}} = \int f(w, \varphi(w + \epsilon))\, d\hat{p}(\epsilon)$$

by randomly sampling values of ϵ from the chosen error distribution. Given this framework, we must decide which cost functions f to use and what distribution p to impose on ϵ.

1.1 Cost Functions

Toward a Quantitative Survey of Dimension Reduction Techniques (Espadoto et. al) uses a variety of cost functions to compare and contrast different dimension reduction techniques. These cost functions often take the mean of reconstruction errors for each point, so they can be easily adapted to the out-of-sample problem.

1.1.1 Trustworthiness

$$f(w, \varphi(w + \epsilon)) = 1 - \frac{2}{6n - 6k - k(k-1)} \sum_{j \in U^{(k)}} [r(j) - k]$$

$$U^{(k)} = \{\text{points among } k \text{ nearest neighbors of } y + \epsilon', \text{ but not } w\}$$

$$r(j) = \text{rank of } z_j \text{ among } k \text{ nearest neighbors of } w$$

1.1.2 Continuity

$$f(w, \varphi(w + \epsilon)) = 1 - \frac{2}{6n - 6k - k(k-1)} \sum_{j \in V^{(k)}} [\hat{r}(j) - k]$$

$$V^{(k)} = \{\text{points among } k \text{ nearest neighbors of } w, \text{ but not } y + \epsilon'\}$$

$$r(j) = \text{rank of } z_j \text{ among } k \text{ nearest neighbors of } y + \epsilon'$$

1.1.3 Projected Precision Score

$$f(w, \varphi(w + \epsilon)) = \frac{1}{2} \left\| \frac{\mathbf{d}(w, Z_{n(w,k;Z)})}{\|\mathbf{d}(w, Z_{n(w,k;Z)})\|} - \frac{\mathbf{d}(y + \epsilon', X_{n(w,k;Z)})}{\|\mathbf{d}(y + \epsilon', X_{n(w,k;Z)})\|} \right\|$$

$$n(w, k; Z) = \{\text{indices of } k \text{ nearest neighbors of } w \text{ in } Z\}$$

$$Z_{n(w,k;Z)} = \{z_i : i \in n(w, k; Z)\}$$

$$X_{n(w,k;Z)} = \{x_i : i \in n(w, k; Z)\}$$

$$d(w, Z_{n(w,k;Z)}) = \{d(w,z) : z \in Z_{n(w,k;Z)}\}$$
$$d(w, X_{n(w,k;Z)}) = \{d(w,x) : x \in X_{n(w,k;Z)}\}$$

*Note: We've used set notation for brevity, but these "sets" should be treated as vectors. $n(w,k;Z)$ is ordered from nearest neighbor to furthest neighbor and this ordering is preserved during the construction of the other vectors.

1.1.4 Normalized Stress

$$f(w, \varphi(w+\epsilon)) = \frac{\sum_i [d(w,z_i) - d(y+\epsilon', x_i)]^2}{\sum_i d(w,z_i)^2}$$

1.1.5 Correlation of Distances

$$f(w, \varphi(w+\epsilon)) = \sigma_{\text{Spearman}}\left(\{d(w,z_i), d(y+\epsilon', x_i)\}_{i=1}^n\right)$$

1.1.6 Proportion of Triplets Preserved

$$f(w, \varphi(w+\epsilon)) = \frac{1}{|T|} \sum_{(i,j)\in T} t(y+\epsilon'; i,j)$$
$$T = \{(i,j) : i \neq j \text{ and } d(w,z_i) < d(w,z_j)\}$$
$$t(y+\epsilon'; i,j) = \begin{cases} 1 & d(y+\epsilon', x_i) < d(y+\epsilon', x_j) \\ 0 & \text{otherwise} \end{cases}$$

2 Assessing and Visualizing Out-of-Sample Performance

Given an out-of-sample point w, we can quantify the performance of an out-of-sample embedding technique at that particular point. The performance, however, is heavily dependent on the location of the new point and the structure of the original data. As such, we would like to study the performance as function of the location of w:

$$w \mapsto \int f(w, \varphi(w+\epsilon))\, dp(\epsilon),$$

where f is one of the aforementioned cost functions or the aggregate embedding score. Unfortunately, this is a q-dimensional functional, so it cannot be easily visualized. If $p = 2$, we can instead plot $\int f(w, \varphi(w+\epsilon))\, dp(\epsilon)$ at the point $\varphi(w) \in \mathbb{R}^2$. A bilinear interpolation can then be applied to give a smooth visual.

3 Interpreting ϵ as Uncertainty

Noise refers to meaningless data, which is ambiguous in the case of (unsupervised) dimension reduction. The underlying signal is dependent on the specific problem and nonexistent in some cases. For example, when reducing the data to two dimensions for the sake of visualization, what underlying signal are we after?

Instead, we can interpret ϵ as uncertainty, which is universal to all collected data. Typical dimension reduction methods all assume the high-dimensional data are exact, which is rarely the case in practice. Our goal is to analyze how the presences of uncertainty affects the performance of out-of-sample embeddings.

3.1 Research Questions

1. How does the presence of uncertainty affect out-of-sample embeddings?
2. How can we study/visualize these effects?
3. Which techniques are more robust to the presence of uncertainty?
4. In the presence of uncertainty, do dimension reduction techniques behave similarly to their out-of-sample versions?

3.2 Performance in the Presence of Uncertainty

The risk

$$\text{risk}(w) = \int f(w, \varphi(w + \epsilon)) \, dp(\epsilon)$$

quantifies the performance at a point $w \in \mathbb{R}^p$ in high dimension in the presence of uncertainty. Risk computes the expected loss. To complete the picture, we can also estimate the distribution and variance of loss.

To see how the addition of uncertainty affects performance, we can compare risk to performance in the absence of uncertainty:

$$\frac{f(w, \varphi(w)) - \text{risk}(w)}{f(w, \varphi(w))}.$$

We may then study the effect of uncertainty as a function of w.

3.3 Extension to Full Dimension Reduction Methods

Linear techniques such as PCA and Kernel PCA have natural out-of-sample extensions. Non-linear techniques such as t-SNE have out-of-sample extensions that adopt the same loss functions. The framework above can be used to analyze the behaviors of these out-of-sample extensions in the presence of uncertainty, but do the original techniques also exhibit the same behaviors when uncertainty is present in all of the data?

In this context, we could define

$$\text{risk} = \int \sum_{i=1}^{n} f(z_i, \varphi(z_i + \epsilon_i)) \, dp(\epsilon_1, \ldots, \epsilon_n),$$

but an adequate estimate of this value would requires us to thoroughly sample from the joint distribution $p(\epsilon_1, \ldots, \epsilon_n)$ and run the DR algorithm for each sample. This only seems feasible for very simple toy examples.