Good morning, I'm João and am here to propose you the theme for my master thesis on maximum variance unfolding applied to disjoint manifolds.

So, what is it about. Maximum variance unfolding is a dimensionality reduction method, which means that it looks for a way to take a dataset with some given dimensions and reduce the number of variables used, without losing much of the structure.

Not only MVU, but most of the methods that we are going to consider, do that by assuming local linearity. This, considers that in a local patch of a manifold, its curvature is essentially null, its locally flat. Which allows us to use the Euclidean distance as it was the true distance between points, or compute the local tangent of the manifold as the pain containing the points in each local neighbourhood. Many methods that we are going to talk about are based on neighbourhood graphs, precisely defining areas where this local linearity can be considered.

In general, most of the considered methods, make part of the subset of spectral methods. These aim to find a set of independent components to represent the given dataset in order of the orthogonal basis found, generally through an eigenvalue decomposition.

With this said, we are going to analyse the main linear method, PCA, then proceed to MDS, that can take linear or non-linear approaches, suck as classical MDS that leads to the exact same solution as PCA or, consider the geodesic distance between points to reach the non-linear solution found by Isomap.

From Isomap onwards, most of our analysed methods involve non-linear dimensionality reduction based on neighbourhood graphs. These, capture better the local properties throughout the dataset’s manifold, over its global geometry, leading to more accurate results when analysing related datapoints.

The other related methods are:

* LE and LLE aimed to build the reduced position based on the weighted influence from each point’s neighbours,
* H-LLE and LTSA, based on flattening the geometry of the dataset based on the manifold’s local tangent space at each point,
* t-SNE focused on computing the points distribution in the neighbourhood of each and maintain it in the low dimensional space. (Kullback-Leiber Divergence)
* MVU mimics the human intuition of stretching the manifold, by distancing the points away from each other while limiting the distances between neighbours.

Multiple extensions to these methods have been studied till date, being some of them:

**Landmark:**

* Synthetise the geometry of the data composed of all the data points by a set of landmark data points which will then be used to compute the rest of the points.
  + This simplification already showed some performance improvements when applied to LLE and Isomap.

The process of rebuilding the full-sized dataset from the landmarks can be a computationally heavy task, but when taken by comparison methods that scale heavily with the number of datapoints, it ends up being beneficial. One of these methods that scales heavily with the number of datapoints is MVU, which scales cubically with *n*.

Comparably to the landmark extensions, our solution proposes to reduce a given dataset to a smaller dataset of points representing the same D-dimensional volume of points. A value that, by the other landmark extensions was left as an adjustable hyperparameter, we believe that it can be approximated to a heuristic.

The focus of our solution, however, is to evaluate a reliable approach for MVU to perform over datasets that are not fully connected. The landmark extensions do not approach this setback, but the out-of-sample extensions does.

In practice, it is possible to connect the manifold by enlarging the neighbourhood at each point, however, since MVU and many other methods also scale heavily on k, there is the need for better solutions. A possible solution is:

**Out-of-sample:**

* Preform MVU concurrently on each disconnected component
* Use the Nystrom approximation to project each reduced disjoint component into the embedded space of a chosen central component.

Even though, this extension already solves the problem we present on MVU, over large distances, it can get inaccurate on the inter-component relations. So, although it also simplifies the computational costs by parallelising the MVU computation over the disjoint components, the heuristic to locate each component could be improved.

Objectively, the solution proposed in this thesis takes an approach relatable to the out-of-sample extension and the landmark’s, simultaneously:

It evolves, from each connected component to its closest neighbour components, the creation of inter-component connections between the points from each pair of components that are closest together. Additionally, synthetising each component’s defining volume with *d+1* data points, creating inner-component neighbourhood connections between all the representing points from each component.

Applying the normal MVU over these landmark points, linearises the global disjoint dataset, but doesn’t find a good reduced representation of each component. To solve this problem, we can compute MVU separately for each disjoint component, and since the main constraint on MVU’s optimisation problem is to keep the distance between neighbouring points exactly the same, the overlap between landmark point from the global computation of MVU and the separate component’s will have marginal error.

HOW TO CONNECT SUBMANIFOLDS?

HOW TO COMPUTE REPRESENTATIVE SUBSET

* RANK(Ci) = DIM(Yi)
* APPROX. MAXIMISE COVERAGE

GLOBAL MVU FORMULATION

COMPUTE AFFINE TRANSFORMATION FOR EACH SUBMANIFOLDS

CHECK IF ISOMAP HAS MANIFOLD TRACE