# Progress Report

## Synopsis of the Last Meeting

During our last meeting, we discussed two probable extensions to our work as mentioned below

1. Project the trained map of an individual player in to a lower dimensional space for visualization. Have the complete set of maps projected to the lower dimensional space to check for overlapping and player similarity.
2. Automate the path of spread and generate an FSM according to the clustering observed in the POS.

## Work Thus Far

I have started working on the first possible extension, which is the mapping of the GSOM to a lower dimensional space using Sammon’s Projection.

* Our high dimensional vectors are the weight vectors of the winner nodes in the trained and tested GSOM. The map I chose for the initial analysis has 13 – dimensions and has 24 winners. (The dataset is attached as a separate attachment)
* As per the Sammon’s paper, it clearly says that in practice the initial configuration for the high dimensional vectors is found by projecting the high dimensional data orthogonally into a lower dimensional space spanned by the lower dimensional coordinates with a large variance.
* By taking the directive given in the Sammon’s paper, I used orthogonal projection of the data first in the three dimensional space. Following are my observations:

Let’s consider one weight vector of our winner nodes (W3 in the attached data set)

When I project the above vector in to three-dimensional space with basis vectors x = [1 0 0], y=[0 1 0] and z = [0 0 1] I get the following projection

The projection effectives zeros out all the dimensions greater than 3 and keeps everything upto it. I find this as an unfair estimation of our initial vectors.

To make it a fair projection I tried to reduce the dimensionality of the data set by using Principle Component Analysis or PCA.

After conducting PCA, I found 10 principle components. This technically means that I only could do away with three dimensions. The cumulative variance of the principle components are mentioned below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Col | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| % | 0.1710 | 0.3419 | 0.5029 | 0.5998 | 0.6718 | 0.7439 | 0.8127 | 0.8733 | 0.9243 | 0.9622 | 1 | 1 | 1 |

The ***“Col”*** represents the dimension of the weight vector and the ***“%”*** represents the variance percentage captured by each dimension.

By looking at the table above, it is clear that PCA would not work.

I searched around and found another technique called kernel PCA which technically transforms the data into a different space and performs PCA on that. However I did not have the time to check this out.

I also think the number of points I used for the PCA is not enough so I was thinking about using the complete GSOM not only the winners to do the projection. Is this a viable projection?

## What I need to be discussed during this meeting

* Would it be fair if I simply project orthogonally and minimize the Sammon’s error?
* Should I think of principle components and its extensions such as kernel PCA to create the initial estimates for the Sammon’s Projection
* Why should we map only the winner nodes? Why not map all the nodes of the finalized GSOM? (resulting in a larger data pool)
* Apart from that my raw data set contains positions of the ghosts and other information such as male/female and the hand orientation should I use them at this time or leave it for later analysis after PhD?