Independent Study, Chaotic Modeling

Luke Beebe

2024-06-09

Introduction

This report covers the usage of **delay mapping** and **multi-view embedding** for future prediction of time-series data in Fresno, CA. This method is not dissimilar to **Empirical Dynamic Modeling** which has grown in popularity to approach questions in non-linear dynamical systems (population dynamics, ecosystem service, medicine, neuroscience, finance, geophysics…etc.), i.e. chaotic systems, where complete parameterization is nearly impossible. I will compare this method to simple historical distributions of temperature and explain the tools I've added onto the code before making suggestions as to next steps.

This independent study took place under Dr, Michael LuValle, alongside my classmate Kavi Chikkappa, who explored **time-series vs multi-view embedding** during the semester.

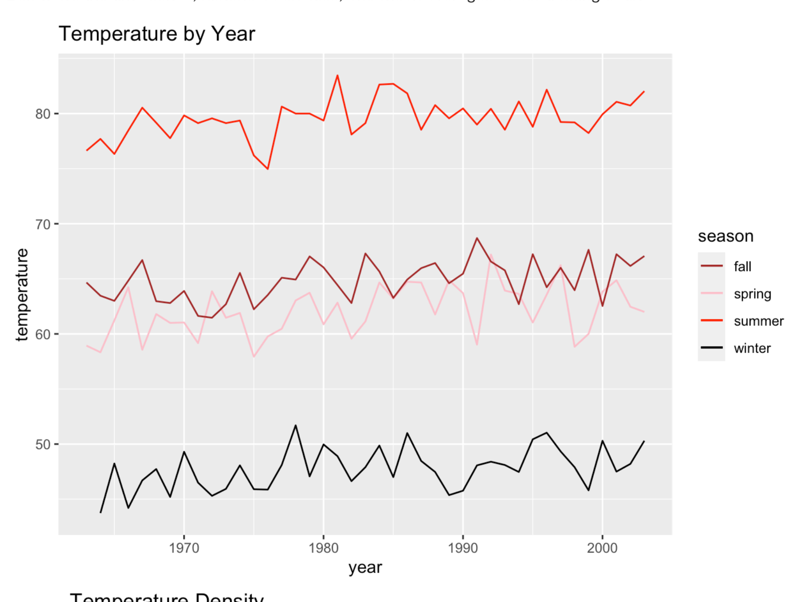
Data

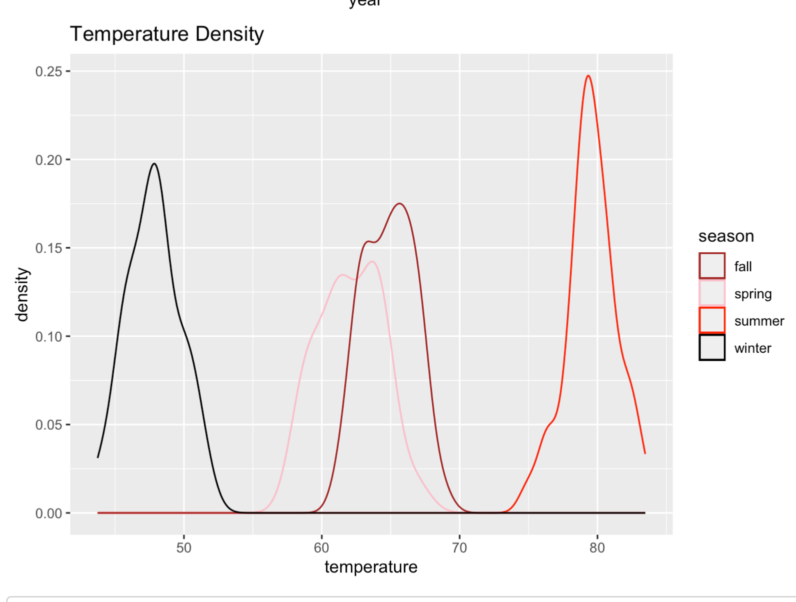
The original data can be found here which contains the variables listed below.

```
##   season year        MEI        PDO        IOD        AO1       PDO1
## 1      1 1963 -0.6410000 -0.5333333  0.1333333  0.3823333  0.8248396
## 2      2 1963  0.3006667 -0.9700000  0.4136667 -0.5043333 -1.0002820
## 3      3 1963  0.8306667 -0.7166667  0.6106667  0.2443333 -0.7216494
## 4      4 1964  0.7080000 -0.4266667 -0.2060000 -0.4560000  0.7074671
## 5      1 1964 -0.6890000 -1.2700000 -0.1610000  0.4263333  0.7835343
## 6      2 1964 -1.3250000 -0.6200000 -0.3413333 -0.1103333 -1.1025918
##   my.sunspot.vec     F_PRCP   F_TAVG
## 1      29.800000 2.05333333 58.93333
## 2      29.566667 0.01333333 76.63333
## 3      32.500000 1.21333333 64.66667
## 4      15.966667 0.31333333 43.73333
## 5      11.533333 0.70666667 58.33333
## 6       7.166667 0.10333333 77.70000
```

The variables are a collection of atmospheric indicators, temperatures, sunspot cycle metrics, and precipitation in Fresno, California from Spring 1963 to Fall 2012 which was given to us by Professor LuValle as the the data he has worked with.

Below are the time series, historical densities, and mean averages of our training data.






```
## [1] "Spring Temp Avg: 62.0756097560976"
```

```
## [1] "Summer Temp Avg: 79.5723577235772"
```

```
## [1] "Fall Temp Avg: 64.8585365853659"
```

```
## [1] "Winter Temp Avg: 47.7675"
```

Technique

1) Take variables, generate *n lags* (*n=4* for this example) for each variable so that each row holds a limited history.

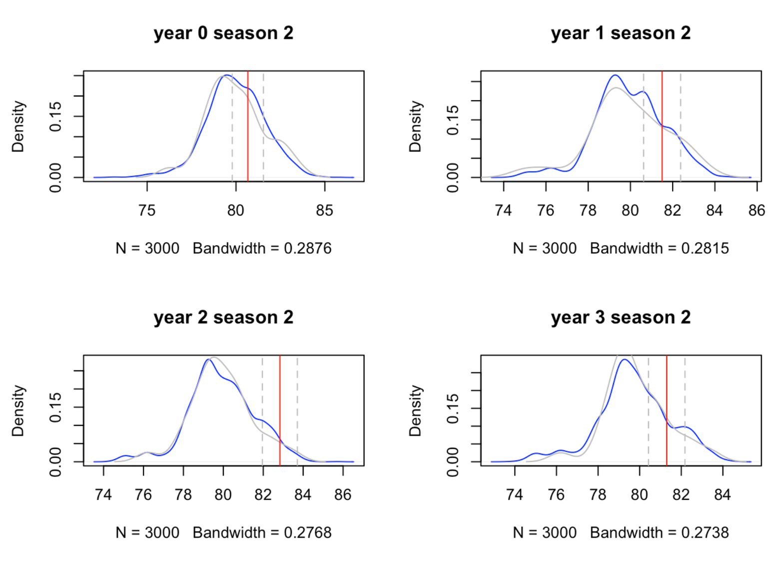
```
##            MEI        PDO        IOD         AO1       PDO1 my.sunspot.vec
## 5  -0.68900000 -1.2700000 -0.1610000  0.42633333  0.7835343      11.533333
## 6  -1.32500000 -0.6200000 -0.3413333 -0.11033333 -1.1025918       7.166667
## 7  -1.23033333 -0.6166667 -0.4153333 -0.07633333 -1.0310862       6.066667
## 8  -0.59066667 -1.3066667 -0.1186667 -1.12533333  0.4275651      15.633333
## 9   0.09266667  0.0000000 -0.1423333 -0.16333333  0.7713328      14.200000
## 10  1.23766667 -0.3566667 -0.1280000 -0.24233333 -0.8463471      12.233333
##         F_PRCP   F_TAVG     MEI t-1     PDO t-1    IOD t-1     AO1 t-1
## 5  0.706666667 58.33333  0.70800000 -0.4266667 -0.2060000 -0.45600000
## 6  0.103333333 77.70000 -0.68900000 -1.2700000 -0.1610000  0.42633333
## 7  0.906666667 63.46667 -1.32500000 -0.6200000 -0.3413333 -0.11033333
## 8  1.370000000 48.23333 -1.23033333 -0.6166667 -0.4153333 -0.07633333
## 9  1.380000000 61.23333 -0.59066667 -1.3066667 -0.1186667 -1.12533333
## 10 0.006666667 76.33333  0.09266667  0.0000000 -0.1423333 -0.16333333
##      PDO1 t-1 my.sunspot.vec t-1 F_PRCP t-1 F_TAVG t-1    MEI t-2    PDO t-2
## 5   0.7074671          15.966667  0.3133333   43.73333  0.8306667 -0.7166667
## 6   0.7835343          11.533333  0.7066667   58.33333  0.7080000 -0.4266667
## 7  -1.1025918           7.166667  0.1033333   77.70000 -0.6890000 -1.2700000
## 8  -1.0310862           6.066667  0.9066667   63.46667 -1.3250000 -0.6200000
## 9   0.4275651          15.633333  1.3700000   48.23333 -1.2303333 -0.6166667
## 10  0.7713328          14.200000  1.3800000   61.23333 -0.5906667 -1.3066667
##       IOD t-2     AO1 t-2   PDO1 t-2 my.sunspot.vec t-2 F_PRCP t-2 F_TAVG t-2
## 5   0.6106667  0.24433333 -0.7216494          32.500000  1.2133333   64.66667
## 6  -0.2060000 -0.45600000  0.7074671          15.966667  0.3133333   43.73333
## 7  -0.1610000  0.42633333  0.7835343          11.533333  0.7066667   58.33333
## 8  -0.3413333 -0.11033333 -1.1025918           7.166667  0.1033333   77.70000
## 9  -0.4153333 -0.07633333 -1.0310862           6.066667  0.9066667   63.46667
## 10 -0.1186667 -1.12533333  0.4275651          15.633333  1.3700000   48.23333
##       MEI t-3    PDO t-3    IOD t-3     AO1 t-3   PDO1 t-3 my.sunspot.vec t-3
## 5   0.3006667 -0.9700000  0.4136667 -0.50433333 -1.0002820          29.566667
## 6   0.8306667 -0.7166667  0.6106667  0.24433333 -0.7216494          32.500000
## 7   0.7080000 -0.4266667 -0.2060000 -0.45600000  0.7074671          15.966667
## 8  -0.6890000 -1.2700000 -0.1610000  0.42633333  0.7835343          11.533333
## 9  -1.3250000 -0.6200000 -0.3413333 -0.11033333 -1.1025918           7.166667
## 10 -1.2303333 -0.6166667 -0.4153333 -0.07633333 -1.0310862           6.066667
##    F_PRCP t-3 F_TAVG t-3    MEI t-4    PDO t-4    IOD t-4    AO1 t-4   PDO1 t-4
## 5  0.01333333   76.63333 -0.6410000 -0.5333333  0.1333333  0.3823333  0.8248396
## 6  1.21333333   64.66667  0.3006667 -0.9700000  0.4136667 -0.5043333 -1.0002820
## 7  0.31333333   43.73333  0.8306667 -0.7166667  0.6106667  0.2443333 -0.7216494
## 8  0.70666667   58.33333  0.7080000 -0.4266667 -0.2060000 -0.4560000  0.7074671
## 9  0.10333333   77.70000 -0.6890000 -1.2700000 -0.1610000  0.4263333  0.7835343
## 10 0.90666667   63.46667 -1.3250000 -0.6200000 -0.3413333 -0.1103333 -1.1025918
##    my.sunspot.vec t-4 F_PRCP t-4 F_TAVG t-4
## 5           29.800000 2.05333333   58.93333
## 6           29.566667 0.01333333   76.63333
## 7           32.500000 1.21333333   64.66667
## 8           15.966667 0.31333333   43.73333
## 9           11.533333 0.70666667   58.33333
## 10           7.166667 0.10333333   77.70000
```

2) Create a *delay mapping* which is random subsets of *m* variables, where *m* equals the dimension of the attractor (*m=4* for this example). The **attractor** is a geometric representation of a system's long term behavior (a multivariate mapping of points through time).

```
## [[1]]
## [1] 40 32  7 38
##
## [[2]]
## [1] 35 24 27 12
##
```

[illegible]

```
##  [7] "MEI t-14, #113"     "IOD t-15, #123"     "F_TAVG t-15, #128"
## [10] "IOD t-14, #115"
```



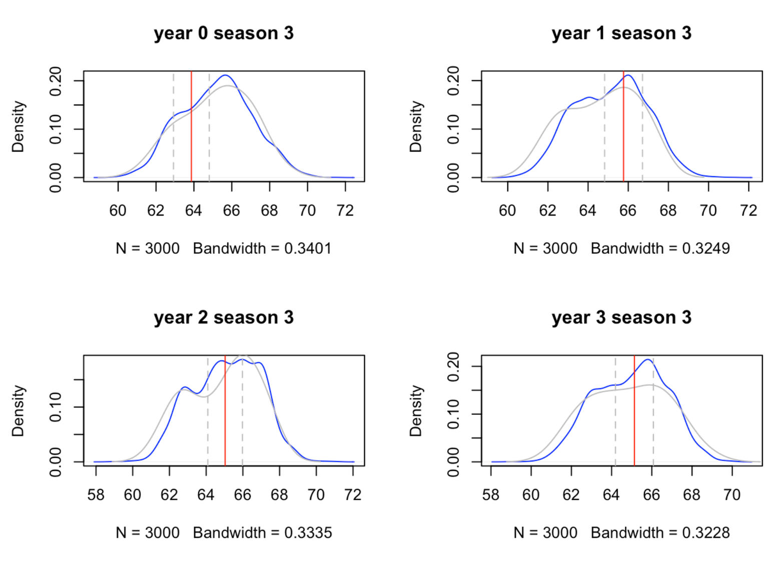

```
weather_prediction(as.matrix(weather_season), x=c(3:9), y=10, split=36,
                   seas=c(3), year=c(0,1,2,3),
                   lag=6, nn=30, dim=5, ntrial=1200,
                   sd=0.5, hist=F)
```

```
## [1] "Significant Variable, Index"
##  [1] "PDO1 t-2, #21"      "PDO1 t-1, #13"      "F_PRCP, #7"
##  [4] "PDO t-1, #10"       "PDO, #2"            "my.sunspot.vec, #6"
##  [7] "PDO t-2, #18"       "PDO1 t-5, #45"      "PDO1 t-3, #29"
## [10] "IOD t-2, #19"
```

```
## [1] "Significant Variable, Index"
##  [1] "AO1 t-6, #52"    "MEI t-6, #49"    "F_PRCP t-6, #55" "PDO1 t-5, #45"
##  [5] "PDO1 t-6, #53"   "PDO1 t-7, #61"   "IOD t-8, #67"    "MEI t-5, #41"
##  [9] "AO1 t-5, #44"    "MEI t-7, #57"
```

```
## [1] "Significant Variable, Index"
##  [1] "MEI t-11, #89"            "MEI t-12, #97"
##  [3] "AO1 t-11, #92"            "IOD t-12, #99"
##  [5] "F_TAVG t-10, #88"         "IOD t-8, #67"
##  [7] "F_PRCP t-11, #95"         "MEI t-8, #65"
##  [9] "PDO t-10, #82"            "my.sunspot.vec t-8, #70"
```

```
## [1] "Significant Variable, Index"
##  [1] "F_TAVG t-16, #136"         "PDO t-17, #138"
##  [3] "PDO t-14, #114"            "PDO t-15, #122"
##  [5] "IOD t-12, #99"             "MEI t-12, #97"
##  [7] "my.sunspot.vec t-15, #126" "PDO t-16, #130"
##  [9] "AO1 t-15, #124"            "PDO1 t-17, #141"
```



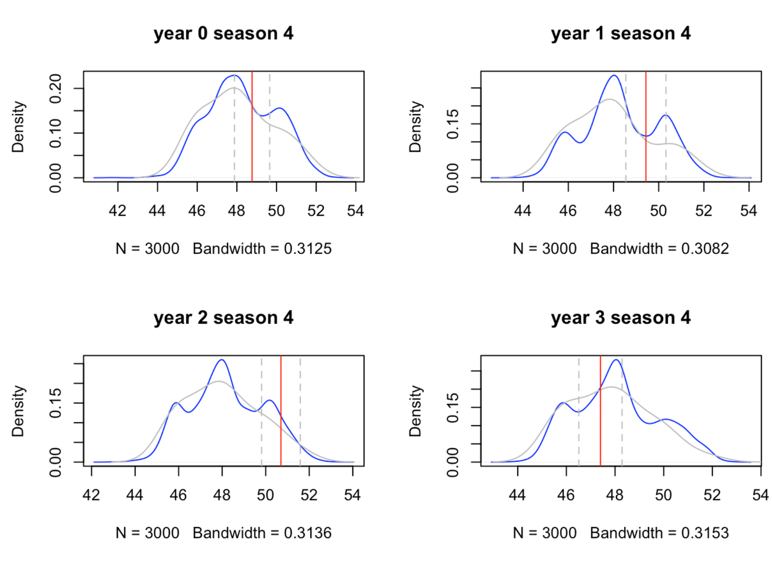

```
weather_prediction(as.matrix(weather_season), x=c(3:9), y=10, split=36,
                   seas=c(4), year=c(0,1,2,3),
                   lag=6, nn=30, dim=5, ntrial=1200,
                   sd=0.5, hist=F)
```

```
## [1] "Significant Variable, Index"
##  [1] "PDO1 t-2, #21"   "PDO, #2"         "AO1, #4"         "F_PRCP, #7"
##  [5] "F_TAVG t-1, #16" "PDO1 t-3, #29"   "PDO1, #5"        "IOD t-4, #35"
##  [9] "PDO t-1, #10"    "PDO1 t-1, #13"
```

```
## [1] "Significant Variable, Index"
##  [1] "F_PRCP t-6, #55" "PDO1 t-9, #77"   "IOD t-4, #35"    "PDO1 t-4, #37"
##  [5] "AO1 t-4, #36"    "MEI t-5, #41"    "MEI t-9, #73"    "AO1 t-6, #52"
##  [9] "F_TAVG t-9, #80" "F_PRCP t-5, #47"
```

```
## [1] "Significant Variable, Index"
##  [1] "PDO1 t-9, #77"      "IOD t-12, #99"     "AO1 t-12, #100"
##  [4] "F_TAVG t-13, #112" "PDO t-10, #82"     "PDO1 t-10, #85"
##  [7] "F_TAVG t-9, #80"    "IOD t-10, #83"     "PDO1 t-8, #69"
## [10] "MEI t-9, #73"
```

```
## [1] "Significant Variable, Index"
##  [1] "IOD t-12, #99"      "AO1 t-12, #100"    "F_TAVG t-13, #112"
##  [4] "F_TAVG t-14, #120" "F_PRCP t-17, #143" "AO1 t-16, #132"
##  [7] "MEI t-14, #113"     "PDO1 t-15, #125"   "AO1 t-17, #140"
## [10] "PDO t-17, #138"
```



From these densities produced it is not apparent whether we are gaining more predictive knowledge or not with our method compared to using their respective historical densities. The shape of the densities are different, but the change in density doesn't always increase in correlation with its predictive accuracy. There are a few things to consider.

The first is that its original ~41 years (50 (total) - 9 (testing)) is not enough to parse nuance from. Our methodology trains each model on the most similar 30 years and produces a prediction from running the numbers from the most similar 5 years. If we do this for each random set of variables *n=1200* times we will likely use all of our training years to build different regression formulas. The nearest neighbors will prevail as the most influential years on our predictions; however, most, if not all years will have influence on our predictions. The hope would be that if our data set was from 1900 and forward, we'd be able to produce densities with slimmer margins than the historical data. This historical density from 1900 would have wider margins than the historical densities we see above due to global temperature trends. It may be prudent to try this method to predict something seasonal in which has a longer recorded history.

Second, the knowledge gained might not be predictive, yet. If you can recreate the historical density, or at least have it similar, then you can understand the relationship better between the variables for each respective season and year. This model gives equal weight to all variables sampled until the regression line is fit with their coefficients optimized. It is important to know which combinations of variables yield good results, which do not, and to what degree they're used in yielding their results. If there is any learning to be done, it is to be with this information.

Third, we can predict many years ahead, but it doesn't mean we should. The densities are similar for each year ahead prediction because the model is trained by omitting one year and predicting the year afterwards. Because of this, the ~41 years become ~40 years, and the lags remain the same although offset by a year, resulting in similar densities. The one way to shake this up would be to add a bayesian approach to update the prediction by feeding it another year to re-weight the variables once that measurement has been made.

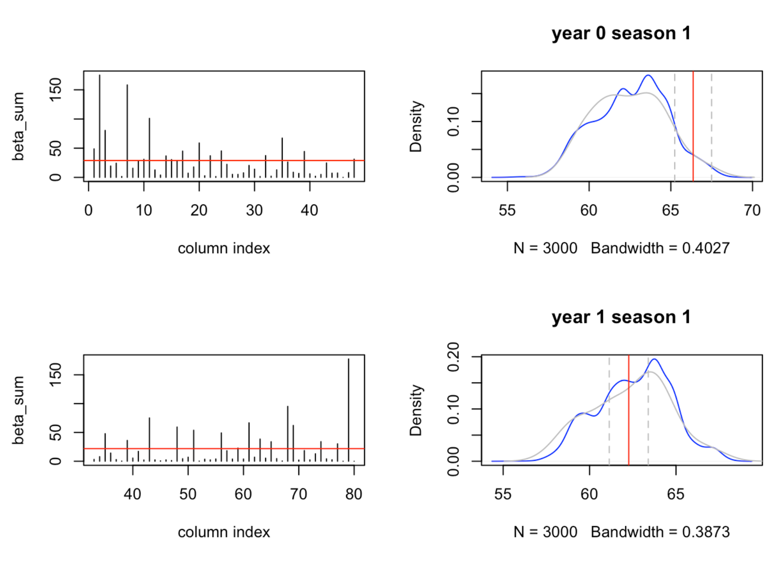
Most Used Predictive Variables

To solve the second and third considerations I have built a method to print out the most significant variables alongside their index number. You can also print a histogram of the index numbers with their total coefficient size. For example, years 1 and 2 of our Spring predictions share similarities in the variables that were 'most predictive' (see below). Below you will see that *"F_PRCP t-9, #79"* and *"AO1 t-8, #68"* have the largest coefficient sums across all random combinations in both models, meaning that if a re-weighting between the variables were to take place by "learning" from 1 year out to predict for year 2, then the density would bias those terms to be sampled more, and theoretically produce more accurate results.

```
weather_prediction(as.matrix(weather_season), x=c(3:9), y=10, split=36,
                   seas=c(1), year=c(0,1,2,3),
                   lag=6, nn=30, dim=5, ntrial=1200,
                   sd=0.5, hist=T)
```

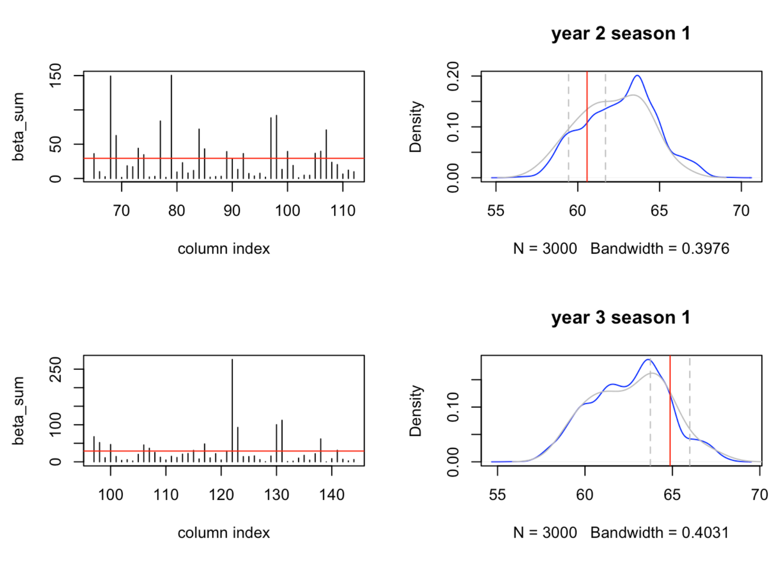
```
## [1] "Significant Variable, Index"
##  [1] "PDO, #2"         "F_PRCP, #7"      "IOD t-1, #11"    "IOD, #3"
##  [5] "IOD t-4, #35"    "AO1 t-2, #20"    "MEI, #1"         "F_TAVG t-2, #24"
##  [9] "MEI t-2, #17"    "F_PRCP t-4, #39"
```

```
## [1] "Significant Variable, Index"
##  [1] "F_PRCP t-9, #79" "AO1 t-8, #68"    "IOD t-5, #43"    "PDO1 t-7, #61"
##  [5] "PDO1 t-8, #69"   "F_TAVG t-5, #48" "IOD t-6, #51"    "F_TAVG t-6, #56"
##  [9] "IOD t-4, #35"    "F_PRCP t-7, #63"
```




```
## [1] "Significant Variable, Index"
##  [1] "F_PRCP t-9, #79" "AO1 t-8, #68"    "PDO t-12, #98"   "MEI t-12, #97"
##  [5] "PDO1 t-9, #77"   "AO1 t-10, #84"   "IOD t-13, #107"  "PDO1 t-8, #69"
##  [9] "MEI t-9, #73"    "PDO1 t-10, #85"
```

```
## [1] "Significant Variable, Index"
##  [1] "PDO t-15, #122"  "IOD t-16, #131"  "PDO t-16, #130"  "IOD t-15, #123"
##  [5] "MEI t-12, #97"   "PDO t-17, #138"  "PDO t-12, #98"   "PDO1 t-14, #117"
##  [9] "AO1 t-12, #100"  "PDO t-13, #106"
```



While I have not implemented a 'learning' function yet, Professor LuValle has, and before I do I would like to make sure I am not creating something wholly similar to his approach unless it is proven to work. But, the gist would be to use the "Most Predictive Variables" to learn further trends and aid future predictions.

There are other insights to gain from these print outs, mainly for Spring above, Precipitation and Temperature are predictive in years closer to present time, whereas in year 4, and even year 3, there are an overwhelming amount of ocean temperatures used to more accurately predict future events. Of course, this is with data local to Fresno and more research will have to be done in different locations to see the range of this effect. As a whole, I believe this method should be tested across many locations using different local variables to test its effects. I also believe if accuracy is to increase in its predictive densities, it will be within variable selection.

The predictive densities represent the possible paths the attractor can take, a shotgun approach, whereas after an observation we can restrict the possible paths to those only near our measurement, a sniper approach. The one thing this doesn't account for is whether the attractor is polynomial or not. Because of this, the predictions must be local, meaning that they should only be made, with confidence, one year ahead.

Conclusion

In this paper I have detailed a report on **delay mapping** and **multi-view embedding** in my independent study. I have detailed using it for seasonal weather predictions and explained the **Most Used Predictive Variables** function I built for further research of weather and chaotic modeling. Weather and other chaotic systems are notoriously difficult to predict, but with these tools we hope to further research in modeling these patterns.