University of California, Merced

Capstone Project

Using Principal Component Analysis and Dynamic Mode Decomposition to Analyze Spatio-Temporal Data

Maureen Long

A technical report submitted
in partial fulfillment of the requirements for the degree of

Master of Science in Applied Mathematics

July, 2011

University of California, Merced
Graduate Division

This is to certify that I have examined a copy of a technical report by

Maureen Long

and found it satisfactory in all respects, and that any and all revisions required by the examining committee have been made.

Research Advisor: _____
Arnold D. Kim

Reading Committee: _____
Harish Bhat

Applied Mathematics Graduate Studies Chair: _____
Boaz Ilan

_____
July 28th, 2011

ABSTRACT OF THE CAPSTONE PROJECT

Using Principal Component Analysis and Dynamic Mode Decomposition to Analyze Spatio-Temporal Data

by

Maureen Long
July 2011

University of California, Merced

Abstract

We study two methods to analyze spatio-temporal data. To describe data, we use principal component analysis. To predict data, we use dynamic mode decomposition. We compute numerical solutions of the complex Ginzburg-Landau equation and we use that numerical solution as data. Using principal component analysis we identify a low-dimensional subspace spanned by only 3 principal components. Using these 3 principal components we can reconstruct the original data matrix with approximately 2% error, and construct out-of-sample data with less than 3% error. Using dynamic mode decomposition we are able to predict the temporal evolution of out-of-sample data as far as 500 time steps into the future with less than 5% error. The combination of these two techniques provides robust and reliable methods to analyze complex data sets.

1 Introduction

Modern society has been generating ever increasing amounts of data, which has led to the demand for data mining and analysis techniques. These techniques span the entire breadth of the computational sciences, and are rapidly developing into their own interdisciplinary field.

The two main goals of data mining are description and prediction. Description focuses on finding features or patterns of a given set of data. Prediction involves using some variables to predict future values of other variables. The main techniques that are used to fulfill these goals are classification, regression, clustering, summarization, dependency modeling, and change and deviation detection [1].

In this paper we use two techniques to both describe spatial patterns of a data set, and offer predictions as to how it will temporally evolve. We create a dynamically rich data set by solving a model system, which enables us to assess the accuracy of our predictions. We find the dominant behaviors using a spatial analysis referred to as Principal Component Analysis. We construct the principal components of the system, and attempt to describe the original data set using only these components. To predict how the system evolves, we use the Dynamic Mode Decomposition of the system.

Using these two techniques enables us to provide both a description of the data set and make some predictions about how the system will continue to evolve.

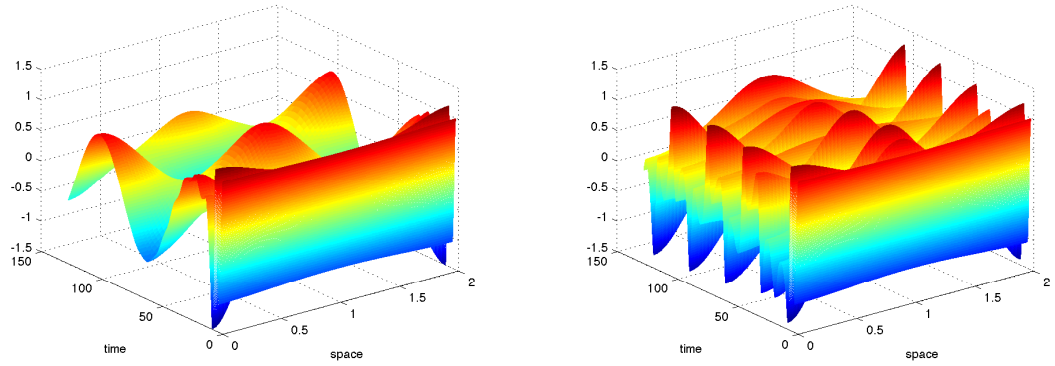
2 Model System

We study the complex Ginzburg-Landau equation

$$iu_t + q^2(1 - ic_0)u_{xx} - i\rho u + (1 + i\rho)|u|^2 u = 0, \tag{1}$$

as our model system. This equation exhibits diffusion, dispersion, linear growth and nonlinearity. Consequently, this model system possesses a rich set of dynamics. We use the complicated dynamics of this model system to generate data to analyze using principal component analysis and dynamic mode decomposition.

The parameter q is a bifurcation parameter. Sirovich and Rodriguez [2] have characterized the dynamics of the solution in terms of q. For example, they have found that for $0.6 \leq q \leq 0.7$, there exists a limit cycle (Figure 1(a)). In addition, for $0.7 \leq q \leq 0.827$, there are two frequencies (Figure 1(b)). To solve (1) numerically, we need a method that is accurate enough to handle the complicated nonlinear dynamics exhibited by this model system. For that reason, we use a Fourier pseudo-spectral method. In addition, we use the implicit Crank-Nicolson scheme for the linear terms, and the Adams-Bashforth scheme for the nonlinear term. We explain the details of the numerical method below.



(a) Limit cycle behavior ($q = 0.67$)

(b) Two frequency behavior ($q = 0.78$)

Figure 1: Solution of the Ginzburg-Landau equation with various bifurcation parameters.

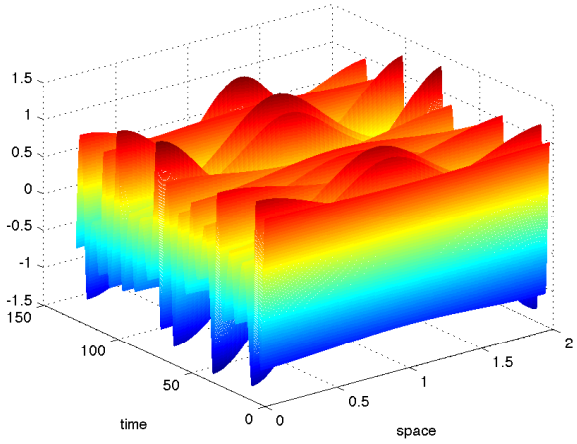


Figure 2: Solution of the Ginzburg-Landau equation exhibiting chaotic behavior ($q = 0.95$)

2.1 Numerical Method

We seek the numerical solution of (1) with periodic boundary conditions using a Fourier pseudo-spectral method on the collocation points $x_j = 2\pi j/N$ for $j = 0, 1, \cdots, N-1$.
We define the approximation

$$U_j(t) \approx u(x_j, t),$$

and the spectral approximation of the second derivative

$$D_N^2 U_j(t) \approx u_{xx}(x_j, t)$$

Our equation for (1) then becomes:

$$U_j'(t) = iq^2(1 - ic_0)D_N^2 U_j(t) - \rho U_j(t) + i(1 + i\rho)\left|U_j(t)\right|^2 U_j(t)$$

When we apply Crank-Nicolson/Adams-Bashforth ($U_j^n \approx u(x_j, t_n)$), we obtain

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{i}{2}q^2(1 - ic_0)D_N^2 U_j^{n+1} + \frac{i}{2}q^2(1 - ic_0)D_N^2 U_j^n \\ - \frac{1}{2}\rho U_j^{n+1} - \frac{1}{2}\rho U_j^n + \frac{i}{2}(1 + i\rho)\left[3\left|U_j^n\right|^2 U_j^n - \left|U_j^{n-1}\right|^2 U_j^{n-1}\right].$$

Rearranging terms, we arrive at the following update formula:

$$\left[1 - \frac{ik}{2}q^2(1 - ic_0)D_N^2 + \frac{k}{2}\rho\right] U_j^{n+1} \\ = \left[1 + \frac{ik}{2}q^2(1 - ic_0)D_N^2 - \frac{k}{2}\rho\right] U_j^n + \frac{k}{2}i(1 + i\rho)\left[3\left|U_j^n\right|^2 U_j^n - \left|U_j^{n-1}\right|^2 U_j^{n-1}\right] \quad (2)$$

Rather than solve (2) directly, we solve it in the transform space. Using

$$U_j^n = \sum_{\xi=-N/2}^{N/2-1} \tilde{a}_\xi(t_n) e^{i\xi x_j}$$
$$\tilde{a}_\xi(t_n) = \sum_{\xi=-N/2}^{N/2-1} \frac{1}{N} U_j^n e^{-i\xi x_j}$$
$$D_N^2 U_j^n = \sum_{\xi=-N/2}^{N/2-1} -\xi^2 \tilde{a}_\xi(t_n) e^{i\xi x_j},$$

our update formula then becomes:

$$\left[1 + \frac{ik}{2}q^2(1 - ic_0)\xi^2 + \frac{k}{2}\rho\right] \tilde{a}_\xi^{n+1} = \left[1 - \frac{ik}{2}q^2(1 - ic_0)\xi^2 - \frac{k}{2}\rho\right] \tilde{a}_\xi^n + \frac{i}{2}(1 + i\rho)\left[3\tilde{b}_\xi^n - \tilde{b}_\xi^{n-1}\right]$$

where $\tilde{b}_\xi = \text{FFT}\left[\left|U_j^n\right|^2 U_j^n\right]$. Rearranging terms, we obtain the update formula

$$\tilde{a}_\xi^{n+1} = \left[1 + \frac{ik}{2}q^2(1 - ic_0)\xi^2 + \frac{k}{2}\rho\right]^{-1} \times \left\{\left[1 - \frac{ik}{2}q^2(1 - ic_0)\xi^2 - \frac{k}{2}\rho\right] \tilde{a}_\xi^n + \frac{i}{2}(1 + i\rho)\left[3\tilde{b}_\xi^n - \tilde{b}_\xi^{n-1}\right]\right\}.$$

To obtain the solution in the physical domain, we compute

$$U_j^{n+1} = \sum_{\xi=-N/2}^{N/2-1} \tilde{a}_\xi(t_{n+1}) e^{i\xi x_j}.$$

Adams-Bashforth is a two-step numerical method. We initialize it using one step of forward Euler,

while still evaluating the nonlinear term in the transform space.

$$\tilde{a}_\xi^1 = \left[1 + \frac{ik}{2}q^2(1-ic_0)\xi^2 + \frac{k}{2}\rho\right]^{-1} \times \left\{\left[1 - \frac{ik}{2}q^2(1-ic_0)\xi^2 - \frac{k}{2}\rho\right]\tilde{a}_\xi^0 + \frac{i}{2}(1+i\rho)\left[\tilde{b}_\xi^0\right]\right\}.$$

For all computations, we used the initial condition

$$u(x,0) = 1 + 0.02\cos x$$

constants $\rho = c_0 = 1/4$ and bifurcation parameter, $q = 0.95$ (placing our dynamics in the chaotic regime). With $N = 64$ Fourier modes and $M = 128,000$ time steps each of size $\Delta t = 0.001$, we obtain the solution plotted in Figure 2.

We store the generated solution of (1) in a data matrix, D, which we describe in detail in the next section. We then use this generated data to study Principal Component Analysis for describing data and the Dynamic Mode Decomposition for predicting data.

3 Principal Component Analysis

Principal Component Analysis (PCA) is a technique to reduce the dimensionality of a data set by transforming to a new set of variables called principal components (PCs). The first few principal components may account for most of the variation present in the original system. If that is the case, we obtain a low-dimensional description of the data. We compute the Singular Value Decomposition (SVD) of a matrix containing the spatio-temporal data to compute the PCs. The SVD is defined as:

$$A = U\Sigma V^H,$$

where U is an orthogonal matrix containing the eigenvectors of AA^H, V is an orthogonal matrix containing the eigenvectors of A^HA, and Σ a diagonal matrix containing the nonnegative set of real square roots of the eigenvalues of A^HA and AA^H in descending order, which are the coefficients and standard deviations of the principal components, respectively [3]. To utilize this fact, we store the solution to (1) at each time step as,

$$d_n = \left[u(x_1,t_n), u(x_2,t_n), u(x_3,t_n), \ldots, u(x_N,t_n)\right]$$

with d_n denoting an individual snapshot. The matrix $D \in \mathbb{C}^{M\times N}$ (where M is the total number of time steps collected in D, and N is the number of collocation points, with $M \gg N$), is defined as

$$D = \{d_1, d_2, d_3, \ldots, d_M\},$$

where each d_M fills a row of the D matrix. We obtain our PCs by computing the SVD of this matrix, the PCs fill the columns of the V matrix.

Once we know the PCs, we first study how well they reconstruct our original data contained in D. We then look at using them to approximate other data not contained in D. The columns of V form an orthogonal basis which allows us to reconstruct the data as a linear combination of $\bar{v}_i$,

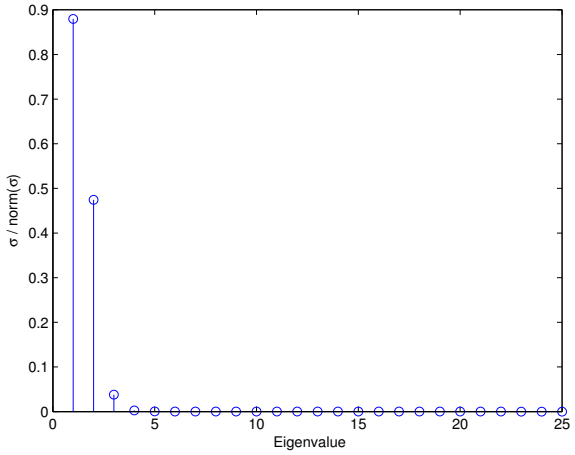


Figure 3: Stem plot of the singular values of the data matrix D, truncated for clarity.

where $\bar{v_i}$ are the columns of the eigenvector matrix V,

$$d_n = a_1\bar{v_1} + a_2\bar{v_2} + \ldots + a_N\bar{v}_N.$$

We can then construct a subspace of the solution using the most relevant principal components, and use this subspace to attempt to create an accurate low-dimensional approximation of the solution. Our approximation to the solution will then be:

$$d_n \approx a_1\bar{v_1} + a_2\bar{v_2} + \ldots + a_j\bar{v}_j$$

where there are $j < N$ relevant principal components.

From looking at Figure 3, it appears that only the first 3 principal components will be relevant. However, we can ensure this mathematically by first normalizing the values along the diagonal of Σ as

$$\sum_{j=1}^{M} \sigma_j^2 = 1,$$

and then finding some $M^* < M$ such that:

$$\sum_{j=1}^{M^*} \sigma_j^2 \geq 0.99.$$

In a least-squares sense, we can recover 99% of the information stored in D, by truncating this sum at M^*. Using our data, we find that $M^* = 3$. This result is expected since Sirovich and Rodriguez [2] have established that there exists a three-dimensional attracting subspace for the complex Ginzburg-Landau equation in this parameter regime.

Re-expressing what we defined before:

$$\begin{bmatrix} \bar{v_1} & \bar{v_2} & \bar{v_3} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \bar{d_i}$$

Since V is orthogonal, we find that:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \bar{v_1} & \bar{v_2} & \bar{v_3} \end{bmatrix}^T \bar{d_i},$$

To determine the accuracy of this reconstruction, we define our residual as:

$$r_i = \left\| \bar{d_i} - \sum_{n=1}^{M^*} a_n \bar{v_n} \right\|_2$$

which is plotted in Figure 4, where we can see that we are getting approximately 2% error from reconstructing D using only 3 vectors.

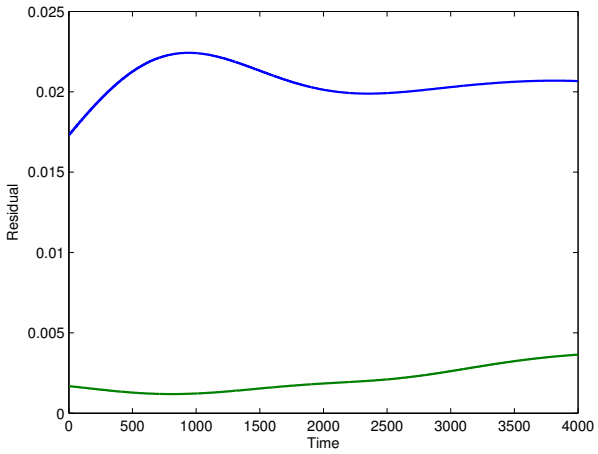


Figure 4: Plot of the residual from recreating D using $M^* = 3$ vectors and using $M^* + 1 = 4$ vectors.

We now test whether our results generalize to out-of-sample data, by checking to see if we can reconstruct data not contained in D. We utilize the same technique, applied to a different data set:

$$d_{2,i} = [u(x_1, t_{2n}), u(x_2, t_{2n}), u(x_3, t_{2n}), \ldots, u(x_N, t_{2n})]$$

and the data matrix:

$$D_2 = \{d_{2,1}, d_{2,2}, d_{2,3}, \ldots, d_{2,M}\},$$

and then using the same procedure as before:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{M^*} \end{bmatrix} = \begin{bmatrix} \bar{v}_1 & \bar{v}_2 & \dots & \bar{v}_{M^*} \end{bmatrix}^T \bar{d}_{2,i}$$

We now find the residual in reconstructing D_2, plotted in Figure 5. From Figure 5, we can see that

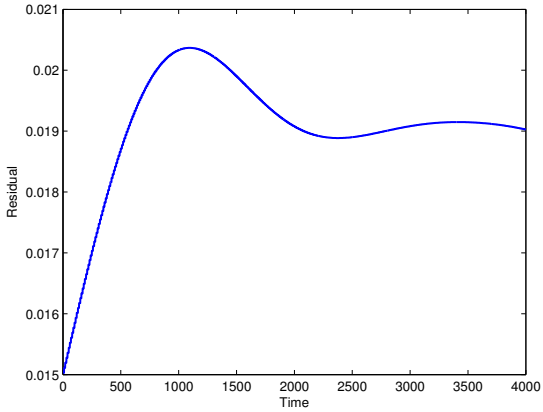


Figure 5: Plot of the residual from recreating D_2 using $M^* = 3$ vectors.

we are able to reconstruct out-of-sample data using data in D with only approximately 2% error.

The next step is to examine if we can reconstruct the solution of a slightly shifted problem. In this case we look at shifting from $q = 0.95$ to $q' = 0.97$. The same setup process applies. We simply apply our PCs from the original D_q to the $D_{q'}$ matrix, and evaluate the residual as before. The results are shown in Figure 6. Here, we actually see a good reconstruction with only 6% error.

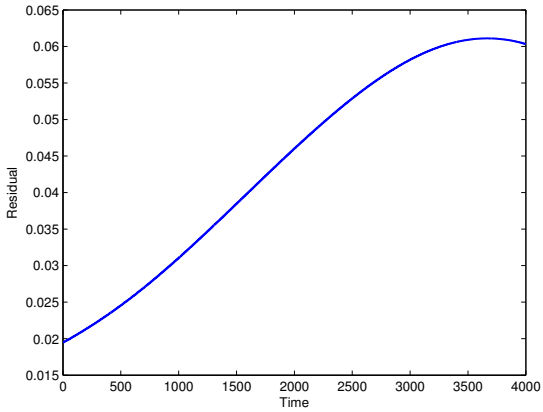


Figure 6: Plot of the residual from reconstructing $D_{q'}$ using $M^* = 3$ vectors.

For this data we have obtained a low-dimensional subspace that effectively reconstructs the data. The principal components of the system allow us to reduce our dimensionality from 64 to 3 (by 95%). A reconstruction of this data using only the spanning set of principal components is accurate to within 2.5% of the original data matrix. When we test our PCA models on out-of-sample data, we find that we can reconstruct data at a later time with only 2.1% error. We can also reconstruct data with a slightly modified bifurcation parameter with only 6% error. This technique is most useful when the original data set has a low-dimensional subspace.

4 Dynamic Mode Decomposition

The Dynamic Mode Decomposition, or DMD, allows us to predict how the solution will change in time using temporal evolution of the data. This technique requires us to make the assumption that the future data is a linear transformation of the past data. Therefore, we assume a model of the form:

$$D_{i+1} = AD_i \tag{3}$$

where we call A the system matrix. To compute this matrix A, we compute a SVD of the D_i matrix:

$$D_i = U\Sigma V^H.$$

Substituting that result into (3), we obtain

$$D_{i+1} = A \cdot U\Sigma V^H$$

If we define the pseudoinverse of Σ as Σ^+, we find that:

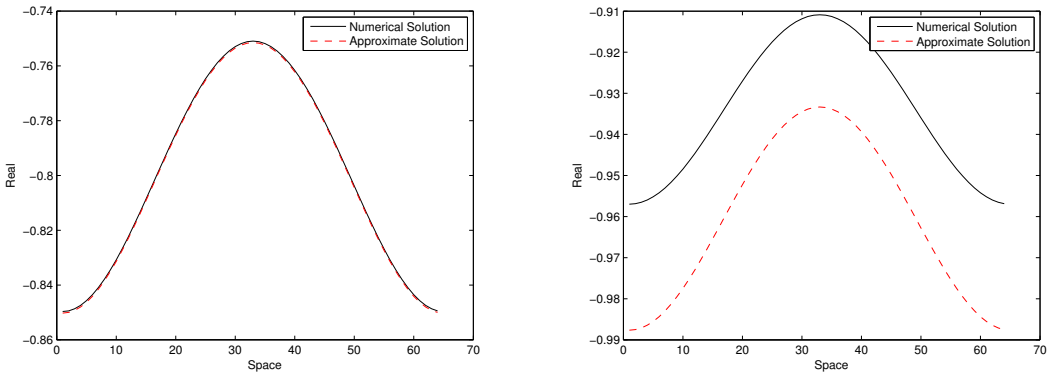
$$A \cong D_{i+1}V\Sigma^+U^H$$

We use A derived from a data set contained in D_1, as defined above, to predict what will happen in future time steps.

In Figure 7 we compare the numerical solution of (1) at $t_n = 5$ to the approximation given by Ad_{n-1}. Figure 7(a) is a much more accurate approximation than Figure 7(b) because we are computing an approximation that is only 5 time steps away from our system matrix's source data, instead of 1000 time steps away. Figure 8 plots the error in the approximation of predicting farther into the future without recalculating the system matrix.

To analyze the system, we look at the approximation

$$d_n \approx Ad_{n-1},$$



(a) $t_n = 5$ (b) $t_n = 1000$

Figure 7: Plots of the numerical solution to (1), compared to plots of the DMD prediction at each corresponding time-step

which leads to

$$d_n = A^n d_0$$

If we write diagonalize A, we obtain

$$u_n = S\Lambda^n S^{-1} u_0$$

However, Λ is just a diagonal matrix of the form:

$$\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

There are three cases to consider. The eigenvectors corresponding to $|\lambda| < 1$ will decay as n increases, those corresponding to $|\lambda| > 1$ will grow as n increases, and those corresponding to $|\lambda| = 1$ will be neutrally stable.

Plotting the eigenvalues of A and the unit circle, we can see that we have 6 eigenvalues that are close to the unit circle. There are also two eigenvalues situated outside the unit circle, which lead to the growth of error in the approximation.

For this data set, we obtained a predictive model that allowed us to accurately predict how the data evolves through time. We obtain well under 5% error for up to 500 time steps ($t = 0.5$) after the last snapshot in D_i.

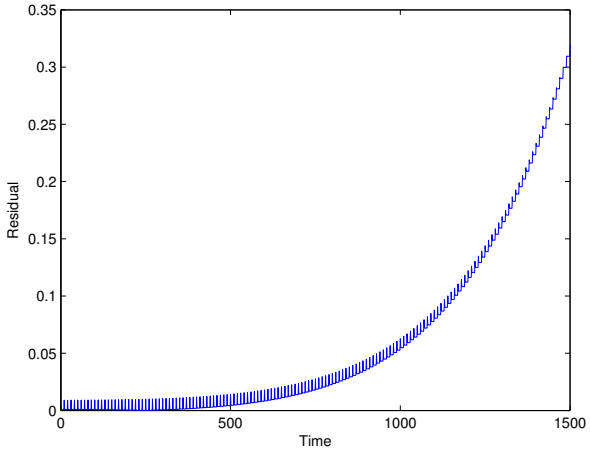
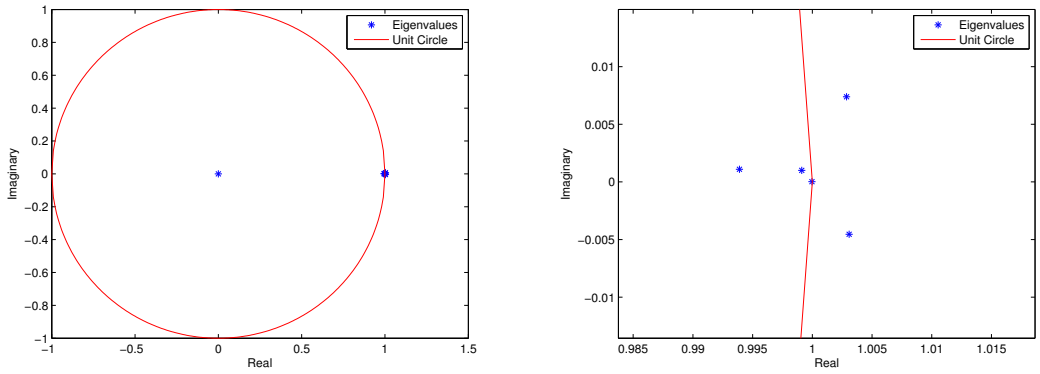


Figure 8: Plot of the residual from the numerical solution and the DMD prediction



(a) Zoomed out

(b) Zoomed In

Figure 9: Plots of the eigenvalues of the system matrix A, with the unit circle overlayed

5 Conclusions

We have presented a set of techniques for describing and predicting complex spatio-temporal data. Using a spatial analysis, we described our data set by constructing a set of principal components which nearly span the space of the solution. We can then use this subspace to approximately reconstruct data at other time steps and for other values of the bifurcation parameter. Using a temporal analysis, we found the dynamic modes of our system, which enables us to predict how the system will change at future time steps.

Using the principal component technique, we were able to reconstruct the original data using only 3 vectors, with under 3% error. The construction of data at a later time step, and of a shifted bifurcation parameter, also resulted in error under 6%. When the data exists on a low-dimensional

subspace, this technique will be very useful.

The dynamic mode decomposition allowed us to predict the temporal evolution of the out-of-sample data. In fact, the technique allows us to make predictions greater than 500 time steps into the future time with error under 5%. However, it is a linear approximation. We have found that the error for this data set rises exponentially.

Data description and prediction is a computationally expensive process. However, both the PCA and DMD techniques are computationally robust, reliable and effective methods for analyzing complex data sets.

References

[1] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine.* 17, 3 (1996): 37-54.

[2] Sirovich, L., and J.D. Rodriguez. "Coherent Structures and Chaos: A Model Problem." *Physics Letters A.* 120, 5 (1987): 211-214.

[3] Jolliffe, I.T. *Principal Component Analysis.* 2nd ed. Springer. 2002.

[4] Schmid, Peter J. "Dynamic mode decomposition of numerical and experimental data." *Journal of Fluid Mechanics.* 656 (2010): 5-28.