Principal Curves and Chaos

Sandeep Rajput[1] and Duane D. Bruns[2]

Department of Chemical Engineering, The University of Tennessee, Knoxville
419 Dougherty Engineering Building, Knoxville, TN 3799

Abstract. A *Principal Curve* is a hypercurve that passes through the center of the data cloud. We adapt and expand the principal curve algorithm to develop a non-parametric approach called Cluster-linked Principal Curves (CLPC) that locally approximates the structure and scatter in a distribution of data points. The iterative algorithm is based on Expectation-Maximization (E-M) principle. The projections of data points on the principal curve or arc lengths are capable of characterizing the data in fewer dimensions and with greater accuracy than PCA. The distribution of arc lengths is used for gauging stationarity and reversibility, and monitoring. For illustration we use the embeddings formed from chaotic gas pressure time series measurements collected before an electrified capillary nozzle that injects bubbles into a liquid-filled column.

INTRODUCTION

Delay embedding is the standard procedure for analyzing almost all nonlinear and chaotic time series. The upper limit for the embedding dimension required to faithfully reproduce the attractor geometry is established in [1-2], but it is an upper bound and a smaller embedding dimension may be sufficient [3]. It is very profitable to reduce the dimensionality of embedding vector representation while keeping the necessary information. The tasks of gauging stationarity and reversibility, process monitoring and control become much more amenable if the effective dimension of the data is smaller.

In this paper we introduce the concept of Cluster-linked Principal Curves (CLPC) that approximate the distribution of data points by a locally linear hypercurve that is obtained by iterative Expectation-Maximization Principle. The approach is an extension of the Principal Curves [4]. The projections on the hypercurve or *arc lengths* are examined for their potential in reducing the dimension, testing for stationarity and reversibility, and for monitoring and control. The experimental data used in this paper were collected on a liquid-filled column with an electrified capillary through which gas is bubbled. The column is referred to as *bubble column* in this document. The bubble column is a low-dimensional system that exhibits period-doubling route to chaos, and its return maps are quite similar to that of the dripping faucet experiment [5]. Experimental setup and relevant references can be found in [6-8].

Email: rajput@engr.utk.edu[1] , DBruns@utk.edu[2]

2. CLUSTER-LINKED PRINCIPAL CURVES

A *Principal Curve*, defined by Hastie and Stuetzle (henceforth referred to as HSPC) is a hyper-curve that locally approximates the data density [4]. The curve is a non-parametric polygonal line. Projecting the point on the polygonal line reduces a point $\mathbf{x}$ to an *arc length* $\lambda(\mathbf{x})$ along it. The arc length is the line integral along the polygonal line from its origin to the point where $\mathbf{x}$ projects on it. The curve is self-consistent, i.e., any point on the curve is the expected value of the distribution at that point. If $\lambda(\mathbf{x})$ is the arc length of the point $\mathbf{x}$, then $\lambda(\mathbf{x})=E(\lambda(\mathbf{x_j})|\lambda(\mathbf{x_j})=\lambda(\mathbf{x}))$.The HSPC algorithm is based on Expectation-Maximization (E-M) principle which successively refines the curve. The iterations are stopped when the fractional change in the sum of squared distances from the curve falls below a predefined threshold.

When dealing with a finite data set, $\lambda(\mathbf{x})$ cannot be found for every possible vector $\mathbf{x}$, because there may not be any other points having the same arc length. In that case, smoothing or kernel estimation techniques can be used to estimate $\lambda(\mathbf{x})$ for a given $\mathbf{x}$. To contend with very small conditional probabilities, HSPC algorithm involves a scatterplot smoothing step, thus leading one to designate one variable as dependent and another as independent. Faced with multivariate data, that choice is not simple to make and is rather restrictive –thus the algorithm is not readily applicable to data with dimension greater than two.

Cluster-linked Principal Curve (CLPC) is an adaptation and expansion of HSPC, and is also a non-intersecting curve through the data space so that it minimizes the orthogonal distances of the data points from it [9]. It treats all dimensions symmetrically and together, and performs smoothing by clustering instead of kernel smoothing. CLPC algorithm forms clusters of data based on their arc lengths. The curve is then redefined by connecting the cluster means with straight lines. The iterations are stopped when the curve stabilizes [9]. Every cluster formed in the CLPC algorithm has an equal number of points. This helps to improve the approximation where the data density is high or where we have more information. Figure 1 shows how CLPC approximates a bubble column return map and a noisy circle. Fifteen clusters were used for both approximations. Note that the return map (figure 1b) is characteristic of many chaotic systems.

CLPC is a continuous polygonal line and can be defined so that it closes on itself, i.e. forms a closed loop. This particular instance can be seen in figure 1a. Our algorithm has only one *hyper-parameter* –viz. the number of cluster centers employed to approximate the structure of data. The optimum number of cluster centers depends on the data density and the complexity of the distribution. Our trials indicate that using a cluster for 10 to 20 points usually yields satisfactory results. Cross-validation methods can be used to check the 'goodness of fit' and to prevent overfitting.

Standard statistical techniques can be used for residual analysis. The residuals are desired to have zero mean, small variance compared to the original data and be independent. The generalized variance of the distribution of points $\{\mathbf{x}\}$ is defined as the trace of $\mathrm{Cov}(\mathbf{x})$ or the sum of variances for each dimension. The ratio of the generalized variance of the residuals to that of the original data provides a good estimate of the remaining variability. In the examples shown in fig. 1, the generalized variance of the residuals was less that 0.1% of that of the data sets.

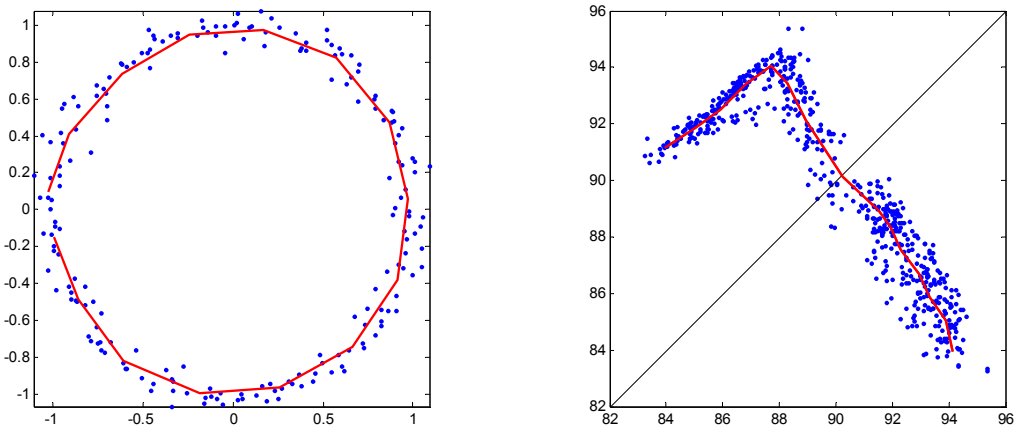
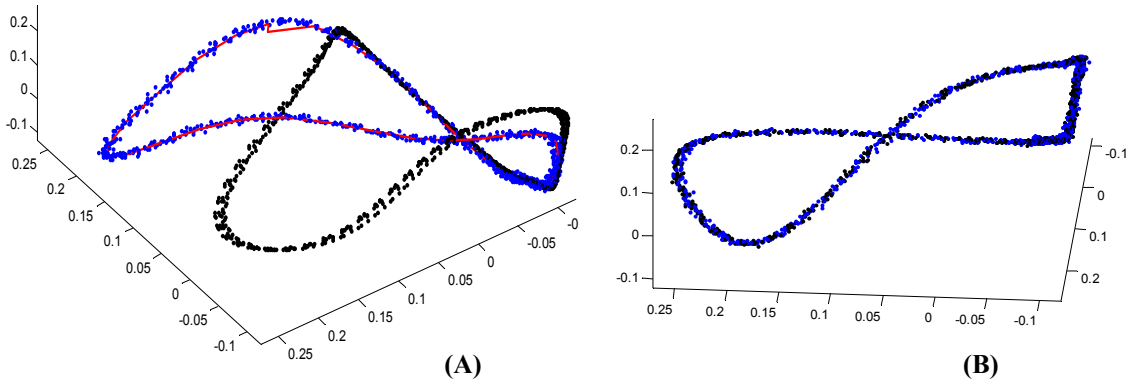


FIGURE 1. Principal Curve fit to a noisy circle (left) and a return map (right)

Beyond these simple examples, CLPC algorithm can be used to approximate the distribution in the embedding space. Figure 2 shows how the CLPC can approximate the attractor of a bubble column, with an embedding dimension of 3. The embedding delay was so chosen to unfold the geometry of the attractor. Figures 2a and 2b show overlaid attractors reconstructed from the first and last one-thirds of a series for Cases A and B respectively. The solid line shows the principal curve. All time series were chaotic. Note that for case A, the electrostatic potential was slightly changed halfway into the experiment and the experiment was allowed to stabilize before resuming measurements. For Case B, the operating conditions were not changed.



(A) **(B)**

FIGURE 2. Principal curves for bubble column embeddings. See text for detail.

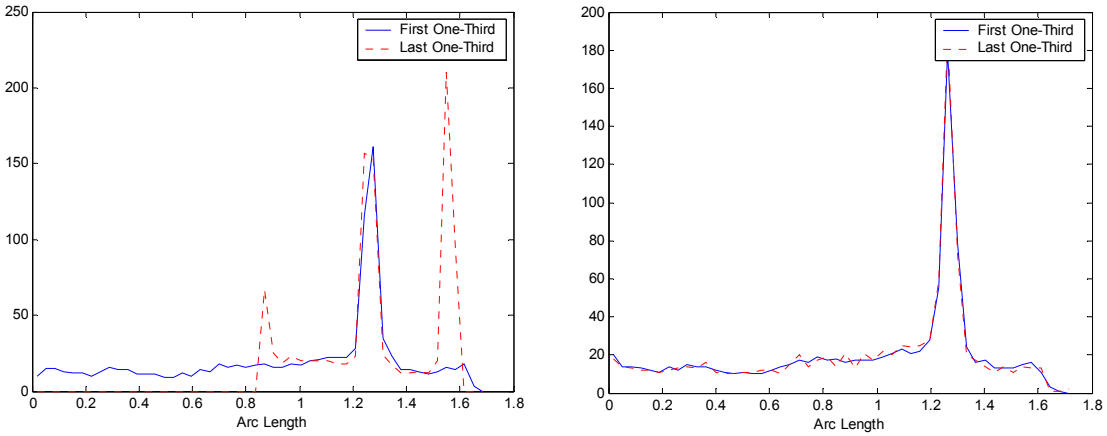
It can be seen that the curve approximates the geometry excellently. Some regions of the attractor are much denser than the rest but the curve approximates the trajectories throughout the embedding space quite well. Note that we do not expect the principal curve to capture the intricate fractal nature of the attractor since that aspect cannot be captured by any curve. However, principal curves reduce the effective dimension of the data from three to one. For lower dimensional systems, our investigations indicate that usually one principal curve is adequate and the arc lengths

along it characterize the system reasonably well. For the attractors shown in figure 2, the generalized variance of the residuals was less than 0.02% of that of the data. Adding random noise to the embedding data does not alter the principal curve appreciably. We note that the magnitude of residuals can be used as an indicator of the noise level.

3. APPLICATIONS OF PRINCIPAL CURVES: TESTING FOR STATIONARITY AND REVERSIBILITY

Non-stationarity implies the inconstancy of the probability density over time. To explore stationarity, one can compare the first and last one-thirds of the data set. The middle one-third is not used in order to temporally isolate the two segments. Discarding the middle one-third of the data helps detect even slow dynamical changes. A principal curve can be fitted to the first one-third of the data and the arc length distribution **R** can be obtained for the corresponding data points. Data points from the last one-third of the time series can be projected on the principal curve to yield another arc length distribution **S**. The χ^2 statistic provides the easiest way to compare these distributions. The chi-square statistic is defined as $\sum[(R_i-S_i)^2/(R_i+S_i)]$ where R_i and S_i denote the elements in the i^{th} bin of **R** and **S**. The corresponding degrees of freedom are N_B-1 where N_B is the number of bins.

If there is no reason to suspect pockets containing very few points in the data space, and some bins are empty, no correction should be made to the degrees of freedom. Otherwise, we suggest reducing the degrees of freedom to account for the empty bins. Note that reducing degrees of freedom increases the Type I error (erroneous rejection of null hypothesis). Figure 3 below shows the overlaid PDFs of the first and last one-thirds of the time series for Cases A and B.



(A) **(B)**

FIGURE 3. Overlaid PDFs for the first and last one-thirds of the time series. Solid and dashed lines represent the first and last one-thirds respectively. See text for detail.

From figure 3a it is obvious that the distributions are very different and the time series was not stationary. On the other hand, from figure 3b it is clear that the distributions are not different. A chi-square test did not reject the null hypothesis of stationarity for case B but did so for case A. Thus the procedure does not reject arbitrarily but captures the shift quite well. Note that one does not have to apply the chi-square test to compare the distributions. The Monte Carlo methods can be used on the surrogate data. Surrogate data [10] can be prepared from the raw time series, and then embedded to form *surrogate embedding vectors*. The arc length distributions of the surrogate embedding vectors then provide confidence limits for the distribution of arc lengths.

This framework can be extended to cover monitoring in a straightforward fashion. The principal curve obtained on the nominal data set can be used to find the arc length distributions for a running window and nominal data. These distributions can then be compared online using the χ^2 statistic. The confidence limits can be found from a probability table. When the χ^2 statistic for the difference between the nominal and the current data exceeds a limit, it indicates a significant change. With à priori information about the classes or régimes and their representative embeddings, a library of arc length distributions can be constructed that can be compared online with the current arc length distribution. Detection of a shift in régimes or states allows better control as well. In addition to the χ^2 statistic, a L^1, L^2 or other measure of distance can also be defined on the distributions and then used for monitoring.

Knowledge of the temporal reversibility of the time series is important since it rules out a random mechanism or its static transformations [11], and limits the modeling approaches suitable for the time series. To gauge for reversibility, one has to explore the differences between the distribution of a time series and that of its time-reversed counterpart. Note that only stationary time series need be tested for irreversibility since non-stationary time series are irreversible.

The procedure detailed above can be utilized to test for reversibility. Time-forward embedding vectors can be formed as $\{x(t)\ x(t-\tau)\ x(t-2\tau)\}$ and time-reversed embedding vectors as $\{x(t-2\tau)\ x(t-\tau)\ x(t)\}$. The distributions of these vectors can then be reduced to that of the arc lengths along the principal curve, and the latter can then be compared under the null hypothesis of reversibility. We use the time series (Case B), which was found to be stationary (see figures 2b and 3b). Figure 4 shows the overlaid PDFs for the time-forward and time-reversed data sets. The distributions are widely different, and a χ^2 test rejected the null hypothesis of reversibility very strongly. The data was chaotic and consequently irreversible. The procedure upheld our belief and thus proved useful. Once again, Monte Carlo simulations can be used for more rigorous hypothesis testing. The definition of principal curves implies reduction of noise, which can be implemented through kernel smoothing applied to a neighborhood of a point **x**.

4. CONCLUSIONS

In this document we showed how the CLPC algorithm can be used to reduce the dimensionality of an embedding, to test for stationarity and reversibility, and extended

it to cover process monitoring. The procedures outlined in this document can be applied to the return maps as well. Further research is underway on this subject, especially about using the CLPC framework for prediction.

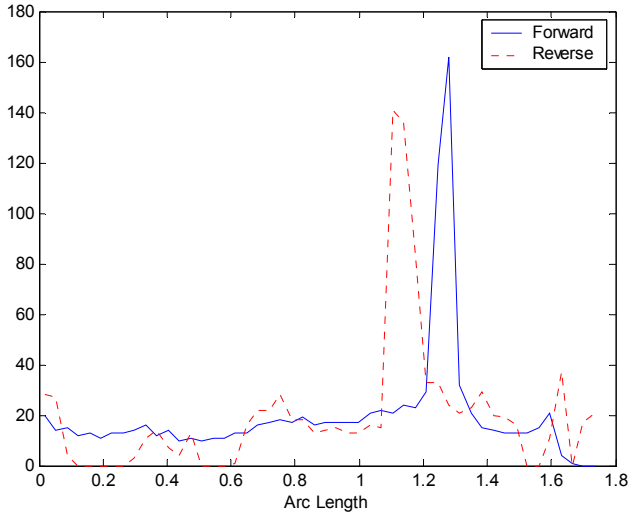


FIGURE 4. Overlaid PDFs for the time-forward (solid line) and time-reversed (dashed line) series. See text for detail.

ACKNOWLEDGMENTS

This work was supported by a grant from the Measurement and Control Engineering Center at The University of Tennessee which is an NSF Industry/University Cooperative Research Center (award number 9908040). The authors thank Dr. C. Stuart Daw for helpful discussions.

REFERENCES

1. Takens, F., in "Dynamic Systems and Turbulence" in *Lecture Notes in Mathematics 898,* edited by D. Rand and L.- S. Young, Berlin: Springer, 1981, p. 366.
2. Sauer, T., Yorke, J. A. and Casdagli, M., *J. Stat. Phys.* **65**, 579-616 (1991).
3. Kennel, M., Brown, R., and Abarbanel, H. D. I., *Phys. Rev. A* **45**, 3043 (1992).
4. Hastie, T. and Stuetzle, W., *J. Am. Stat. Soc.* **84**(406), 502-516 (1989).
5. Bruns, D. D., DePaoli, D. W., Rajput, S. and Menako, R., AIChE Annual Meeting 2002, Indianapolis, IN
6. Menako, C. R., *M. S. Thesis* (2001), The University of Tennessee, Knoxville.
7. Kang, Y., Cho, Y. J., Woo, K. J., Kim, K. I., and Kim, S. D., *Chem. Eng. Sci.*, **55**, 411-419 (2000).
8. Bruns, D. D., Cheng, M., Nguyen, K., Finney, C.E.A., Daw, C. S. and Kennel, M., *J. Chem. Eng.*, **64**(1), 191-197 (1996).
9. Rajput, S., *Ph.D. Dissertation* (2002), The University of Tennessee, Knoxville.
10. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J. D., *Physica D* **58**, 77 (1992).
11. Tong, H., *Non-linear Time Series: A Dynamical System Approach*, Oxford: Clarendon Press, 1990, pp. 193-198.