

Fourier Harmonic Approach for Visualizing Temporal Patterns of Gene Expression Data

Li Zhang and Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
{lizhang,azhang}@cse.buffalo.edu

Murali Ramanathan
Department of Pharmaceutical Sciences
State University of New York at Buffalo
Buffalo, NY 14260
murali@acsu.buffalo.edu

Abstract

DNA microarray technology provides a broad snapshot of the state of the cell by measuring the expression levels of thousands of genes simultaneously. Visualization techniques can enable the exploration and detection of patterns and relationships in a complex dataset by presenting the data in a graphical format in which the key characteristics become more apparent. The purpose of this study is to present an interactive visualization technique conveying the temporal patterns of gene expression data in a form intuitive for non-specialized end-users.

The first Fourier harmonic projection (FFHP) was introduced to translate the multi-dimensional time series data into a two dimensional scatter plot. The spatial relationship of the points reflect the structure of the original dataset and relationships among clusters become two dimensional. The proposed method was tested using two published, array-derived gene expression datasets. Our results demonstrate the effectiveness of the approach.

Keywords: *visualization, gene expression, time series, Fourier harmonic projection*

1 INTRODUCTION

Knowledge of the spectrum of genes expressed at a certain time or under given conditions proves instrumental to understand the working of a living cell. DNA microarray technology allows measurements of expression levels for thousands of genes simultaneously [9]. Extensive research has been conducted on the study of temporal patterns of gene expressions [10, 16]. Clustering methods which group genes or samples with similar patterns have become mainstream analysis tool [14]. Visualization can facilitate the discovery of structures, features, patterns, and relationships in data and may provide more insightful information than

traditional numerical methods. By visualization, we hope to gain some intuition regarding the data, but more importantly, we would like to understand the relationships between data points and detect the intrinsic structure, or possible cluster tendencies. Visualization is especially important in the early stages of data analysis in which qualitative analysis is primary to quantitative. Early success will enhance the users' performance in the remaining stages of analysis. Array-derived gene expression datasets present analysis and visualization challenges because of their dimensionality, noisy environment, and pattern varieties.

The parallel coordinates approach [18] is perhaps the simplest method to display patterns of gene expression profiles. Here the data in each dimension are plotted along a separate axis. Holter et al. [8] have used parallel coordinates to visualize the temporal progression in the yeast cell cycle data. Self-organizing maps (SOM) [12] is another approach. More recently, Hautaniemi et al. [6] have presented a heat map-based strategy for visualizing the U-matrix from SOM. The most prominent visualization-enhanced analysis tool for gene expression data is TreeView [5], which provides a user-friendly computational and graphical environment for assessing the results from hierarchical clustering. The graphical presentations from TreeView include a dendrogram to reflect the distance relationships between clusters and a heat plot to visually convey gene expression changes between samples. The heat plot can be viewed as variation of the parallel coordinates plot in which color is used to convey dimension values.

Here, we present an alternative mapping for multi-dimensional data that is based on the first harmonic of the discrete Fourier transform. The mapping has interesting properties and preserves certain key characteristics of a variety of data sets, especially time series data. Unlike parallel coordinates and heat plot which display all individual dimensional information, our approach uses a two dimensional point to represent each gene over the time at the com-

putational cost of $O(N \log N)$. It focuses on one very important aspect of the visualization: revealing the structure of the entire dataset. Tested using two published, array-derived gene expression time series datasets, our results indicated that temporal patterns were well reflected in the visualization: cluster relationship became two dimensional, clusters were apparent, and outliers were clear. An interactive visualization tool, VizStruct, was implemented to perform the visualization.

The remainder of this paper is organized as follows. Section 2 presents the model of visualization. In section 3, we show our analysis results. The final section discusses other issues in this approach. Proofs for all mapping propositions are included as appendix.

2 METHODS AND SYSTEM

The Mapping

Mapping converts multi-dimensional data to two-dimensions for visualization. An array-derived profile for M genes with N measurements results in $M \times N$ -dimensional data point containing real valued numerical data. Time series data in its simplest form is merely a set of data $\{y_t, t = 0, \dots, N - 1\}$ where the subscript t indicates the time at which the datum y_t was observed [4]. On the other hand, a discrete-time real signal on N evenly distributed time points is represented as an indexed sequence of N real numbers $0, \dots, N - 1$ denoted by $\mathbf{x}[n]$ [1]. Each term of $\mathbf{x}[n]$ is denoted by $x[n]$. The denotation similarity between time series and digital signal suggests that we can view each data point in a time series as a discrete-time real signal (It is not necessary for the signal's time index to comply with the actual time points). In this scenario, the problem of a two dimensional visualization of the time series is transformed into the problem of finding a two-dimensional point estimation for signals (data points).

The frequency domain representation of discrete-time signals is through discrete-time Fourier transform, or DFT [13]. The DFT of a N -point signal $\mathbf{x}[n]$ is a frequency sequence with N complex values: $\mathcal{F}(\mathbf{x}[n]) = [\mathcal{F}_k(\mathbf{x}[n])]$, where each

$$\mathcal{F}_k(\mathbf{x}[n]) = \sum_{n=0}^{N-1} x[n] \mathbf{W}_N^{nk}, \quad k = 0, \dots, N-1. \quad (1)$$

$\mathbf{W}_N = e^{-i2\pi/N}$ is called twiddle factor.

Each harmonic \mathcal{F}_k in the DFT is a measure of the k th sinusoidal frequency component in the signal. For example, the zero harmonic, \mathcal{F}_0 , is the mean value, the first harmonic \mathcal{F}_1 measures the base frequency component, the second harmonic \mathcal{F}_2 measures the component in the signal that is twice the base frequency, and so forth. Because Fourier

harmonics, in general, are complex numbers, they provide the two-dimensional point estimate for mapping a multi-dimensional signal. For this reason, we refer to the mapping as the Fourier harmonic projections. In particular, we are interested in the first Fourier harmonic projection (FFHP):

$$\mathcal{F}_1(\mathbf{x}[n]) = \sum_{n=0}^{N-1} x[n] \mathbf{W}_N^n = \sum_{n=0}^{N-1} x[n] e^{-i2\pi n/N}. \quad (2)$$

The relationship between the DFT and the mapping allows the fast Fourier transform algorithm (FFT), originally discovered by Cooley and Tukey [13], to be used for computation. The FFT is a computationally efficient algorithm and has a complexity of $O(N \log N)$.

The complex number of $\mathcal{F}_1(\mathbf{x}[n])$ in Equation (2) can be expressed in terms of magnitude r and phase θ to provide a useful geometric interpretation of the mapping. The data set was normalized so that the range of values of each dimension across the dataset was 0 to 1. For a data point with N dimensions, the complex exponential divides a unit circle centered at the origin of the complex plane into N equally spaced angles. The value of the first dimension is projected on the radial line corresponding to $\theta = 0$ and, similarly, the value of the k th dimension is projected on to the radial line corresponding to the $\theta = 2\pi(1 - k)/N$ radians. The overall two-dimensional FFHP mapping is the complex sum of all N projections from a data point. Figure 1 illustrates the geometric interpretation for a point containing 6 dimensions.

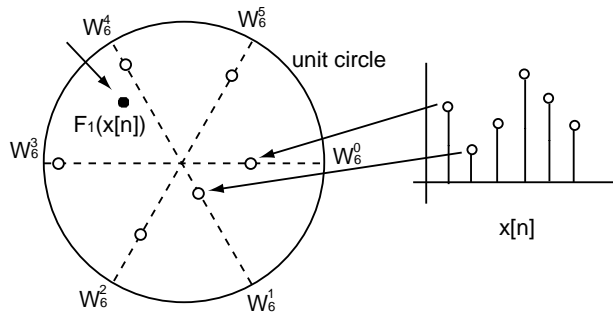


Figure 1. A geometric interpretation of the first Fourier harmonic projection. A normalized 6-dimensional data point is shown on the right by the stem plot. The powers of twiddle factor W_6 divide the unit circle centered at the origin into 6 equal angles and each dimension of the data point is projected onto a different radial angle (open circle). The projections are taken complex number sum to give a 2-dimensional image (indicated by a filled circle).

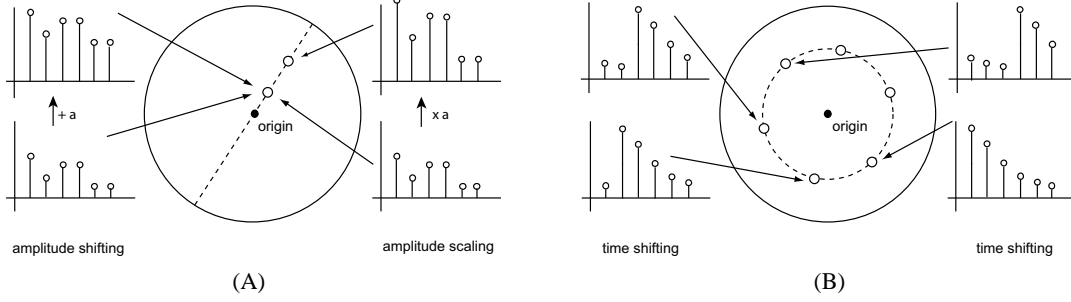


Figure 2. Illustration of the effect of (A) amplitude shifting and multiplying, and (B) time shifting of the first Fourier harmonic projection.

Properties of FFHP

The first Fourier harmonic projection has useful properties that preserve the correlation between dimensions in the multi-dimensional data point. We summarize them as propositions below. Detailed proofs for propositions are provided as the appendix.

1. Data points with equal values for all the dimensions are mapped to the origin. If $\mathbf{x}[n] = [a, \dots, a]$, then $\mathcal{F}_1(\mathbf{x}[n]) = \mathbf{0}$.
2. Two data points whose dimension values differ due to the amplitude shifting of a constant are mapped to the same point. If $\mathbf{y}[n] = \mathbf{x}[n] + a$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])$.
3. Two data points whose dimension values differ due to the amplitude multiplying a constant are mapped to the two points on a line through the origin. If $\mathbf{y}[n] = a\mathbf{x}[n]$, then $\mathcal{F}_1(\mathbf{y}[n]) = a\mathcal{F}_1(\mathbf{x}[n])$. See Figure 2A.
4. Two data points whose dimension values are transposing each other, i.e. symmetric regarding the middle time point, are mapped to the points symmetric to the real axis. If $\mathbf{y}[n] = \mathbf{x}[N - n - 1]$, then $\mathcal{F}_1(\mathbf{y}[n]) = \overline{\mathcal{F}_1(\mathbf{x}[n])}$.
5. Data points that differ only because they are “time-shifted” by d dimensions relative to each other are mapped to the circumference of the circle that is concentric with the unit circle and the angle between the points in the visualization is $\phi = 2\pi d/N$. If $\mathbf{y}[n] = \mathbf{x}[n - d]$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])W_N^d$. This property is illustrated in Figure 2B.
6. Let $\mathbf{w}[n] = \mathbf{x}[n] - \mathbf{y}[n]$ be the difference between the two N -dimensional points, $\mathbf{x}[n]$ and $\mathbf{y}[n]$. The distance between these two points in the visualization is:

$$\|\mathcal{F}_1(\mathbf{w}[n])\|^2 = g_0 N \left(1 + 2 \sum_{k=1}^{N-1} r_k \cos(2\pi k/N) \right) \quad (3)$$

Theory in Practice

We will demonstrate in the next section that the relative locations of temporal profiles’ mapping images can be predicted by Propositions 1–5. For example, genes with relatively low levels of expression throughout the time line are mapped close to the origin. Genes that steadily increase over the time and genes that steadily decrease over the time line are likely mapped symmetric to the real axis.

In Equation (3), the g_0 is the variance of $\mathbf{w}[n]$, r_k is the k th-sample autocorrelation coefficient of $\mathbf{w}[n]$. It can be shown [4] that $-1 \leq r_k \leq 1$ and for mutually independent random sequences or white noise, $r_k = 0$. Equation (3) provides insight into the cluster delineation capabilities of the first Fourier harmonic projection. The variance and k th-sample autocorrelation coefficients of the difference between two points within a given cluster are likely to be small because they will share similarities across many dimensions. Thus, points within cluster are likely to map close to each other in the visualization.

The first Fourier harmonic projection is well suited for visualizing temporal patterns. FFHP has a better class separation capability when the first harmonic is more dominant among all harmonics. Since the first harmonic is a measure of the base sinusoidal frequency component in the signal, a lower frequency signal tends to have a larger first harmonic. This is the case for a large number of time series data where most trend patterns show low frequency variations (not having multiple cycles). The first Fourier harmonic projection is sensitive to dimension order. For time series data, the time points (dimensions) are naturally ordered.

VizStruct System

The first Fourier harmonic projection approach is implemented as a visualization tool called VizStruct, which is written in Java and is available on request from the first author (lizhang@cse.buffalo.edu). The name VizStruct em-

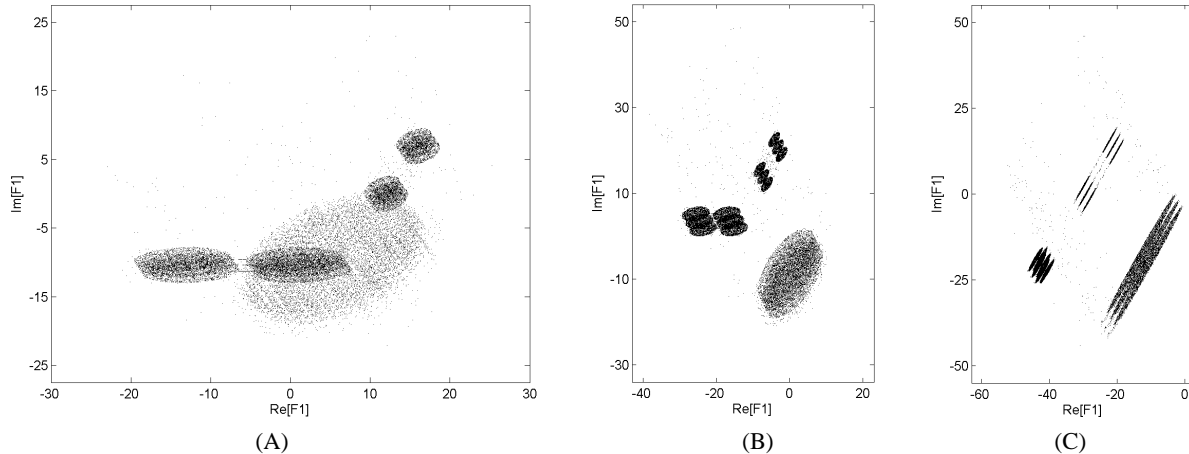


Figure 3. Three snapshots from a dimension tour through a synthetic three-dimensional data set containing 25,000 points. The parameter settings for (A)–(C) were $(0.5, 0.5, 0.5)$, $(0.3, 0.5, 1)$, and $(0.05, 1, 1)$, respectively.

phasizes its capability of visualizing the structure of the dataset.

Dimension Tour

The dimension tour is a feature of VizStruct that allows the user to interact with the data via dynamic animations. It is analogous to the grand tour [2] and the user interacts with the visualization by changing the dimension parameter associated with each dimension. The default value for each dimension parameter is 0.5 and the individual parameters can be changed over the range -1 to $+1$ either manually or systematically using the program. Because no two-dimensional mapping can capture all the interesting properties of the original multi-dimensional space, two points that are close in the visualization can theoretically be far apart in the multi-dimensional space. The dimension tour, which creates animations that explore dimension parameter space, can reveal structures in the multi-dimensional input that were hidden due to overlap with other points in the visualization.

Figure 3 illustrates the capabilities of the dimension tour using a synthetic dataset containing 25,000 points in three dimensions. At the default settings of dimensional parameters (Figure 3A), 5 clusters are apparent. However, during the course of the animations (Figures 3B and 3C), the multi-layered structures of the original 5 clusters become increasingly apparent.

3 RESULTS

Data Sets for Visualization

Our approach was tested using two published array-derived data sets. The *rat kidney* array dataset of Stuart et al. [16] contains measurements of gene expressions during rat kidney organogenesis. The data were downloaded from <http://organogenesis.ucsd.edu/data.html>. It consists of 873 genes which vary significantly during kidney development at 7 different time points: gestational day 13, 15, 17, 19; newborn (N); 1 week (W); and nonpregnant adult (A).

The *fibroblasts* dataset of Iyer et al. [10] is the result of a study of the response of human fibroblasts to serum. It consists of gene expressions measuring the temporal changes in mRNA levels of 517 human genes at 13 time points, ranging from 15 minutes to 24 hours after serum stimulation. The data were downloaded from <http://genome-www.stanford.edu/serum>.

Rat Kidney Dataset

In the *rat kidney* dataset, there are 5 discrete patterns or groups of gene expression during nephrogenesis. Figure 4 illustrates these temporal profiles characterized by the idealized gene expressions.

Figure 5A shows how genes are classified by a hierarchical clustering algorithm. It copies the Figure 3 in [16]. Figure 5B shows the parallel coordinates of the dataset. Patterns of genes in each group comply to the profiles depicted in Figure 4 (with some noise).

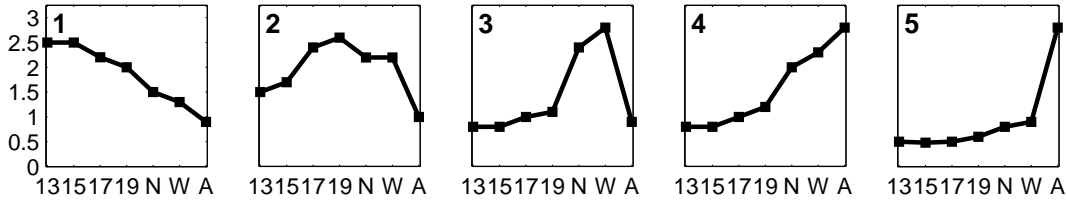


Figure 4. Idealized temporal gene expression profiles during kidney development. The groups were named 1 through 5 based on the timing of their peak expression during development. Seven time points were 13, 15, 17, 19 embryonic days; N, newborn; W, 1 week old; A, adult.

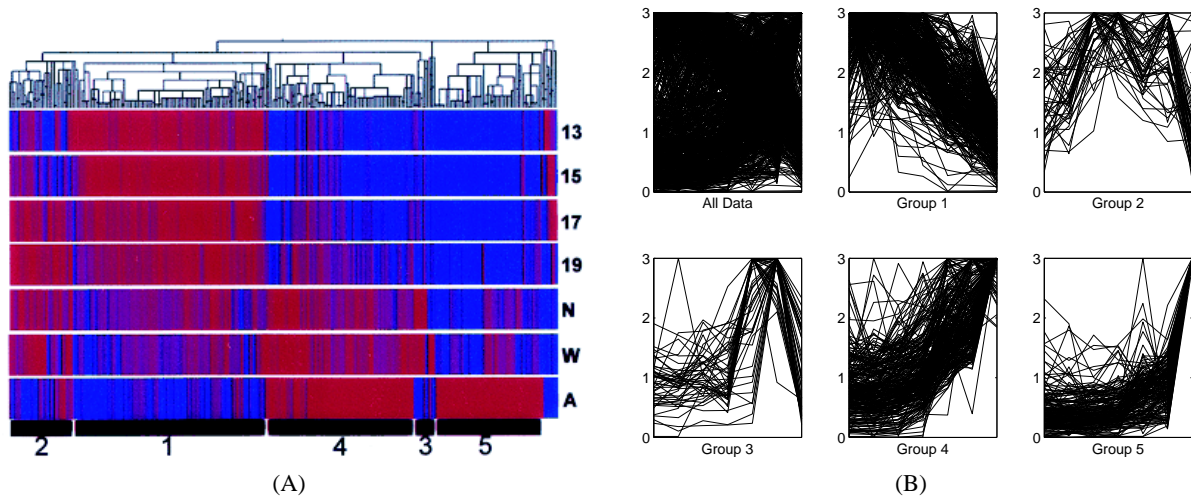


Figure 5. Visualization of the *rat kidney* dataset in heat plot and parallel coordinates. (A) Dendrogram and a heat plot from hierarchical clustering algorithm. (B) Parallel coordinates for the entire dataset and each of the gene groups.

Figures 6A-B show the visualization of the *rat kidney* dataset in VizStruct under the first Fourier harmonic projection for two dimension parameter settings. There are 5 sets of colored symbols for each of the 5 gene groups. Each symbol represents one gene across 7 time points. In Figure 6A, two big clusters are clearly apparent from the visualization. The top cluster consists of genes from groups 3, 4, and 5. The bottom cluster is comprised of genes from groups 1 and 2. Furthermore, genes from each group are aggregated.

The formation of two large clusters can be interpreted by the temporal profiles. Groups 1 and 2 with genes which have very high relative levels of expression in early development are quite different from groups 3, 4, and 5 for genes that have a relatively steady increase in expression throughout development. The visualization also indicates that points in the upper cluster are symmetric to the points in the lower cluster. Properties of FFHP may suggest the reason. Temporal profiles of groups 1 and 4 suggest that

they are somewhat symmetric to the middle time point (gestational day 19). By Proposition 4, they would be mapped to points symmetric to the real axis. On the other hand, groups 4 and group 5 are mapped closely since they have similar profiles except for the significantly up-regulated in the last time point. The same arguments can be applied in the case of group 1 vs. group 2, or group 3 vs. group 4.

Due to the noise, most boundaries between groups are not very clear. However, the separation between group 4 and group 5 improves in Figure 6B compared to Figure 6A.

Fibroblast Dataset

Temporal patterns are slightly complicated in the *fibroblast* dataset. Data has been classified into 10 groups using the hierarchical clustering algorithm by the original author [10]. Figure 7A shows the result of the hierarchical clustering. It is a copy of Figure 3 from [10]. Figure 7B gives the

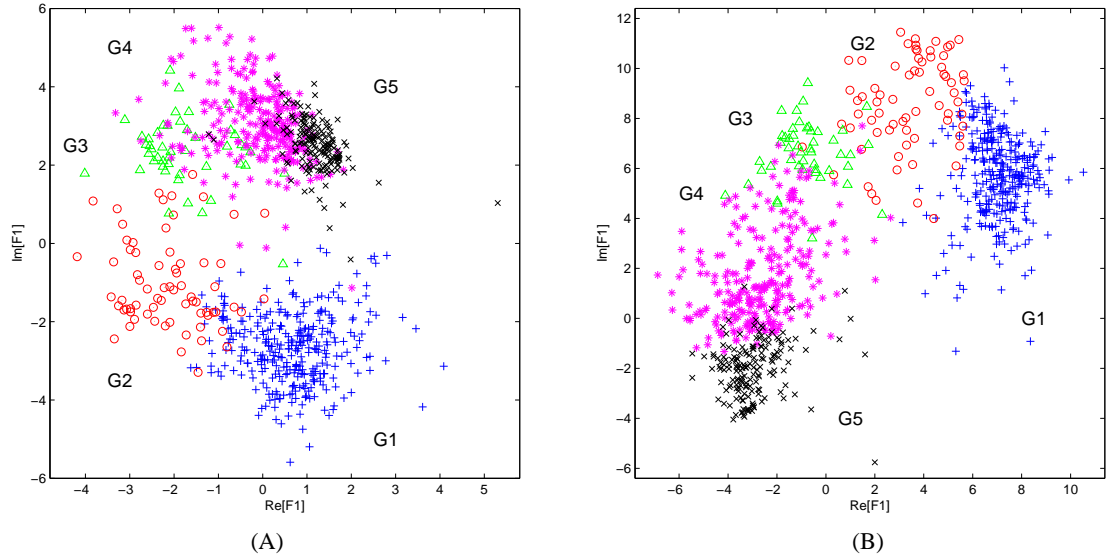


Figure 6. Visualization of the *rat kidney* dataset in VizStruct. Five gene groups were represented by blue plus symbols, red circles, green triangles, magenta stars, and black cross symbols. The dimension parameters for (A) and (B) were $\langle 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5 \rangle$ and $\langle 1, 0.5, -1, -0.5, 0.5, 1, -1 \rangle$, respectively.

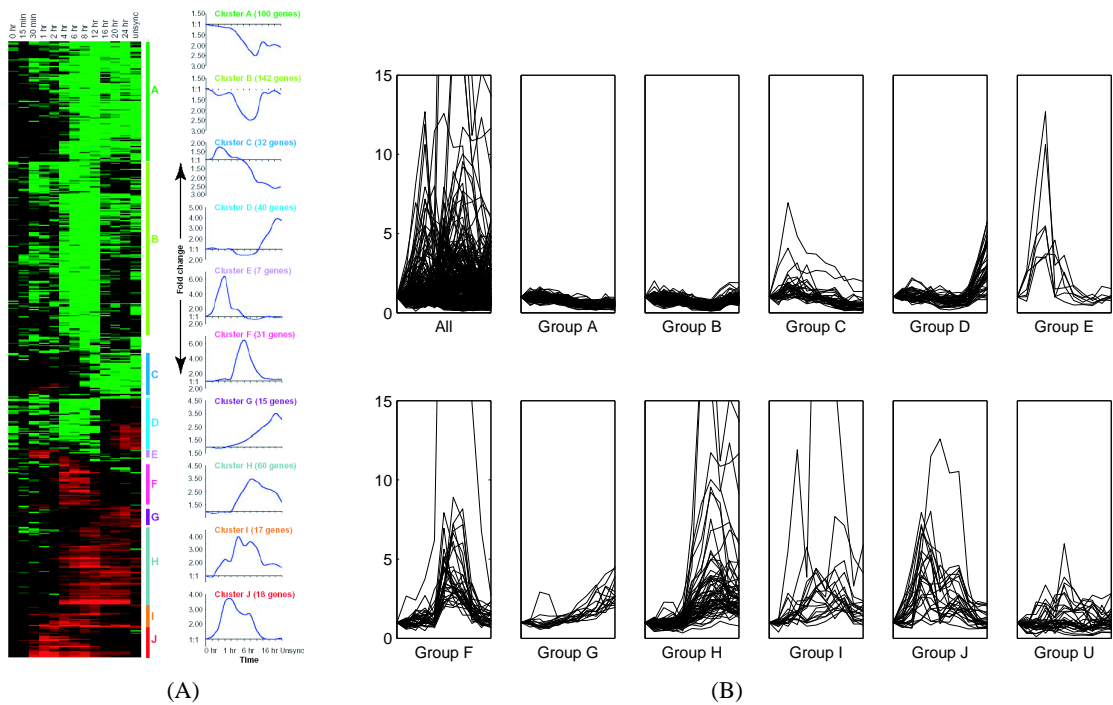


Figure 7. Visualization of the *fibroblast* dataset in heat plot and parallel coordinates. (A) Dendrogram and a heat plot from hierarchical clustering algorithm. (B) Parallel coordinates for the entire dataset and each of the gene groups. The gene group labelled *U* consists genes without label after hierarchical clustering.

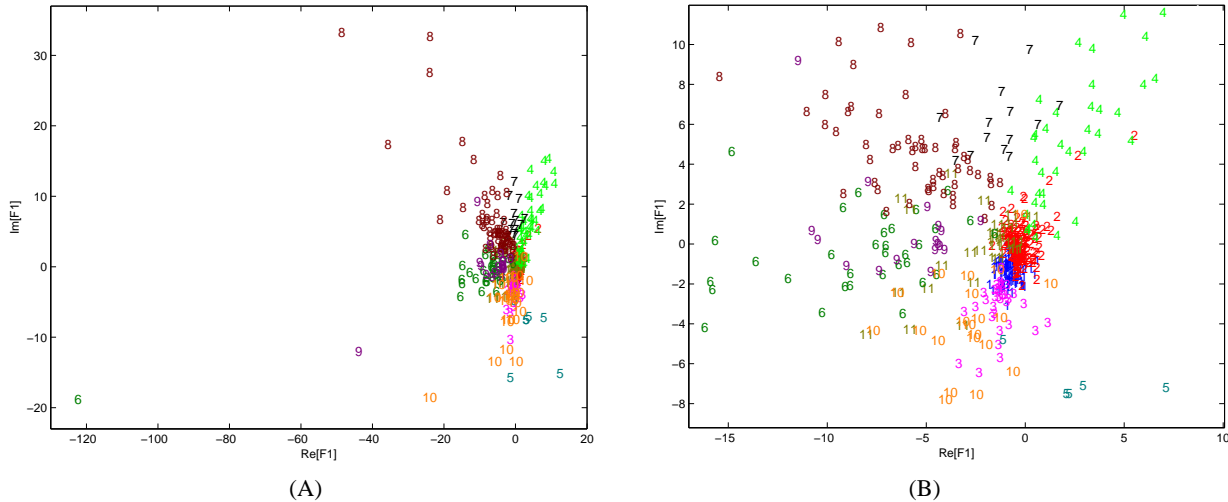


Figure 8. Visualization of the *fibroblast* dataset in VizStruct. 11 colored numbers were used for each of the gene groups. (A) The mapping of the entire dataset. (B) Enlarged portion of the visualization in (A).

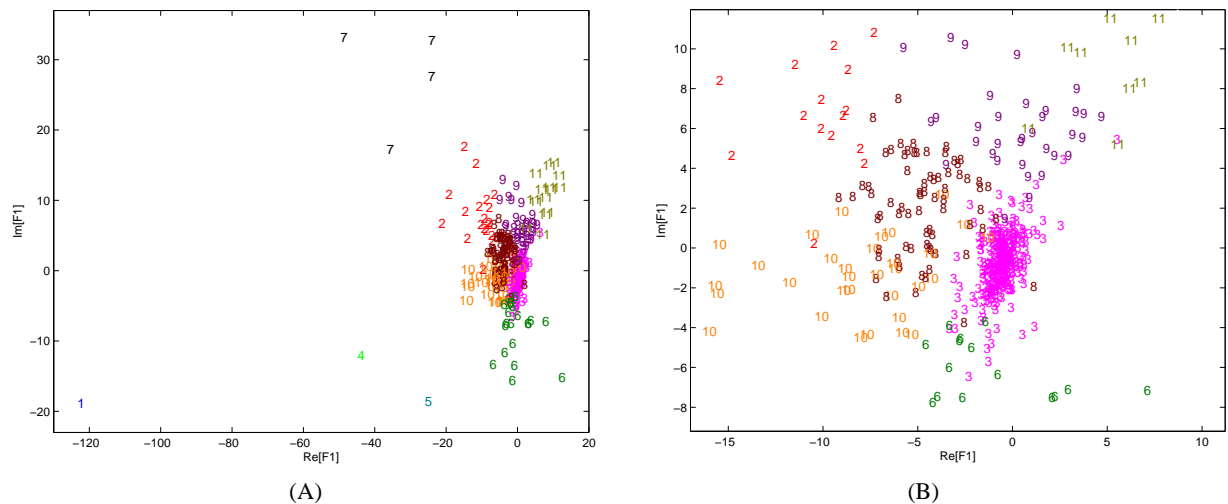


Figure 9. Visualization of the *fibroblast* dataset in VizStruct. Genes were grouped by the k-means clustering method. 11 colored numbers were used for each gene groups. (A) The mapping of the entire dataset. (B) Enlarged portion of the visualization in (A).

parallel coordinates the entire dataset and each gene group.

Figure 8A shows the visualization of the *fibroblasts* dataset in VizStruct under the first Fourier harmonic projection. Colored numbers 1 through 11 were used for each gene group. The visualization reveals several outliers and no distinct clusters. The majority of data are too dense to be seen. By zooming technology, the enlarged detail is shown in Figure 8B. More numbers spread out, but most blue numbers (1) were still covered by numbers 2 and 3.

Two observations can be made: (1) a large number of genes are close to the center. (2) most clusters have a radial shape emitting from the center. They can be interpreted by the FFHP properties. (1) Temporal patterns of group *A* and group *B* have very flat shape and relative smaller values. By Propositions 1 and 2, they tend to be mapped closed to the origin. (2) In hierarchical clustering, Pearson's correlation coefficient was used to measure the similarity. Genes whose time values differ due to the amplitude shifting or multiply-

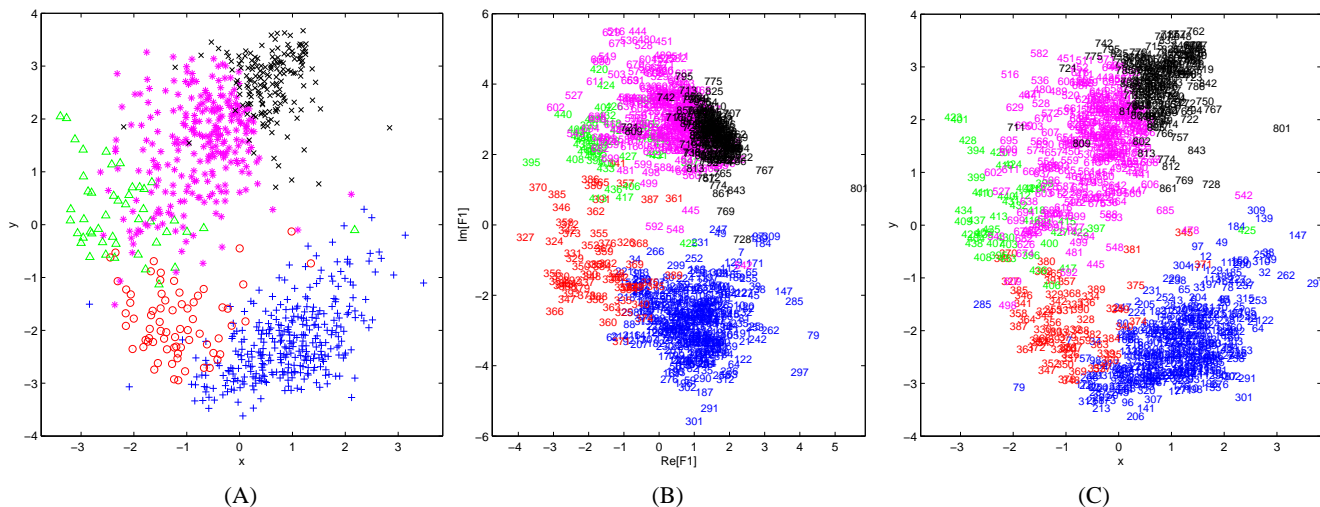


Figure 10. Comparison of Sammon's mapping and the first Fourier harmonic projection. (A) The result from Sammon's mapping of the *rat kidney* dataset. Five gene groups were represented by 5 colored symbols the same as were used in Figure 6. (B)–(C) Comparison of relative gene locations between (B) VizStruct and (C) Sammon's mapping. Five colors were used for each of the gene groups. The color schema is the same as in (A). Notice that the visualization layout of (A) and (C) is identical. The purpose of panel (C) is to compare the gene by gene location to the panel (B).

ing a constant have very high coefficient values. By Propositions 2 and 3, they tend to be mapped closely along a line though the origin.

Some relative gene group locations can be predicted. For example, by Proposition 4, groups 4 and 5 are mapped symmetric to the real axis. Close inspection suggests that temporal pattern of group 8 is somewhat a 2 or 3 right time-shift from the pattern of group 6. By Proposition 5, each time point shift corresponds to roughly $2/13\pi \approx 30^\circ$ rotation. In Figure 8B, genes in group 8 are mapped about 60° clockwise to genes in group 6.

Lacking clear clusters in the visualization indicates that different clustering methods may yield different results. Figure 9 shows the grouping aspect of k-means clustering. Euclidean distance was used in k-means as the distance measure. The visualization reveals that genes closed to the origin are grouped as one.

Comparison of FFHP to Other Visualizations

Heat plot and parallel coordinates display all individual dimensional information. In parallel coordinates, a polyline represents a gene over time. This format is a concise and intuitive for displaying temporal profiles. However, the parallel coordinate has obvious drawbacks: when the data size becomes larger it becomes increasingly unreadable as indicated in the first panel of Figure 5 and Figure 7. Heat plot uses a color mosaic instead of a polyline to overcome over-

lapping. Combined with the dendrogram from hierarchical clustering, it gives a global view as well as individual clusters of the dataset by grouping genes with similar patterns together. Cluster relationships in heat plot is one dimensional: clusters of genes are listed one by one as shown in Figure 5A and Figure 7A.

The first Fourier harmonic projection takes a different approach. It uses a two dimensional point to represent each gene over time. By doing so, it takes advantage of the spatial relationship of the points to reflect the structure of the original dataset. Thus the cluster relationships become two dimensional and FFHP has a better capability of displaying outliers than heat plot. Unlike heat plot, which requires algorithms to group similar genes to make a meaningful visualization, FFHP directly mapped the multi-dimensional data onto a two dimensional space without any prior knowledge of the dataset.

Multidimensional scaling (MDS) [3] is a competing approach for visualizing multi-dimensional data. We compared FFHP to Sammon's mapping [15], a variant of MDS that optimizes the following stress function \mathcal{E} :

$$\mathcal{E} = \frac{1}{\sum_i \sum_{j < i} d_{ij}^*} \sum_i \sum_{j < i} \frac{(d_{ij} - d_{ij}^*)^2}{d_{ij}^*}, \quad (4)$$

where d_{ij}^* is the distance between points i and j in the N -dimensional space and d_{ij} is the distance between i and j in the visualization.

Figure 10 shows Sammon’s mapping of the *rat kidney* dataset. A comparison of Figure 10A to Figure 6A reveals the extensive similarities between FFHP and Sammon’s mapping. The relative locations of individual samples are also remarkably similar. This is indicated in Figure 10B and Figure 10C.

Sammon’s mapping has some drawbacks: (1) it provides a single final result and the user cannot intervene interactively during visualization, (2) the incremental addition of even a single point requires a complete repetition of the optimization procedure and possible extensive reorganization of all the previously mapped points to new locations, and (3) it requires time-consuming optimization procedures of time complexity $O(N^2)$ or greater.

Our results illustrate some of the strengths and weaknesses of FFHP. The visualization reflects the structure of the dataset: outliers, clusters and their relationships. Comprehending the structure of the dataset can facilitate the choice and understanding the results of different clustering methods. Although the FFHP uses an approach to mapping multi-dimensional data that is distinctively different from Sammon’s mapping, it yields results that are consistently comparable. However, when the dataset contains a large number of patterns, it becomes difficult to separate different patterns and the visualization may be difficult to interpret.

4 DISCUSSION

Visualization of microarray data is a challenge because of its high dimensionality. In this paper, we have explored use of the first Fourier harmonic projection for visualizing multi-dimensional time course array datasets. Unlike parallel coordinate or heat plot approach which display all dimensional information, FFHP uses a two dimensional point to represent each data point (gene in this case). Our results indicated that temporal patterns were well reflected by spatial relationships: genes with similar pattern were aggregated and relative locations of gene groups can be predicted. Moreover, the first Fourier harmonic projection was shown to yield results that were similar to those from Sammon’s mapping.

Achieving two dimensional mapping requires a trade-off. The mapping is lossy for detailed dimensional information. Our approach attempts to preserve to the maximum semantics of the data points via Fourier harmonic aspect. In particular, characterizing the data using two descriptive measurements: the real and imaginary portions of the first discrete Fourier harmonic. A similar approach uses principal component analysis (PCA) [11]. This visualization deploys another two descriptive measurement: the first and second principal component.

In addition to the first Fourier harmonic projection, higher Fourier harmonics can also be used as mappings.

It can be shown that for any harmonic (> 1), there exists an equivalent first harmonic of the original discrete signal whose time indices are systematically rearranged. At certain conditions (such as temporal patterns with high frequency variations, i.e. multiple cycles), higher harmonic projections enhance substructure separation in the visualization. Detailed discussion is beyond the scope of this paper due to a length constraint.

The FFHP mapping results in two-dimensional visualizations that are identical to those of radial coordinate visualization techniques, e.g., RadViz [7]. However, rather than vector notation and the *spring paradigm* of RadViz, we have used a complex number notation. This substantive reformulation of the mapping provides valuable theoretical insights and allows important properties of mapping, including its relationship to the DFT, to be easily derived. It can also create possible extensions such as higher Fourier harmonic projections.

The first Fourier harmonic projection requires data without missing values. This requirement can be easily met because filling in missing values is a mature research field [17].

Gene expression data can be studied in either sample space or gene space. Here, we have reported only the visualization on gene space. In a separate report [19], we applied the first Fourier harmonic projection on the sample space and performed visualization-driven classifications. Our experiments demonstrated that FFHP offered an alternative format of visualization. We believe that using FFHP alone or in combination with heat plot or parallel coordinates would give a biologist additional powerful tools for analyzing and visualizing microarray data sets.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Science Foundation. We also thank the anonymous reviewers for their constructive comments on the manuscript.

References

- [1] Cadzow, J. A., Landingham, H. F. *Signals, Systems, and Transforms*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.
- [2] Cook, C., Buja, A., Cabrera, J., and Hurley, C. Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, 2(3):225–250, 1995.
- [3] Davison, M. L. *Multidimensional Scaling*. Krieger Publishing, Inc., Malabar, FL, 1992.
- [4] Diggle, P. J. *Time Series: A Biostatistical Introduction*. Oxford University Press, Oxford OX2 6DP, 1990.
- [5] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95:14863–14868, December 1998.

- [6] Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., Ruiz, J., Mousses, S., and Kallioniemi, O. Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps, 2003. To appear in Machine Learning: Special Issue on Methods in Functional Genomics.
- [7] Hoffman, P. E., Grinstein, G. G., Marx, K., Grosse, I., and Stanley, E. DNA Visual and Analytic Data Mining. In *IEEE Visualization '97*, pages 437–441, Phoenix, AZ, 1997.
- [8] Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., and Fedoroff, N. V. Fundamental Patterns Underlying Gene Expression Profiles: Simplicity from Complexity. *Proc. Natl. Acad. Sci. USA*, Vol. 97(15):8409–8414, July 2000.
- [9] Ideker, T., Galitski, T., and Hood, L. A New Approach to Decoding Life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.*, 2:343–372, July 2001.
- [10] Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J. R., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science*, Vol. 283(1):83–87, January 1999.
- [11] K. Y. Yeung and W. L. Ruzzo. Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics*, Vol. 17(9):763–774, 2001.
- [12] Kohonen, T. *Self-Organizing Maps*, Springer Series in Information Sciences, volume Vol. 30. Springer, Berlin, Heidelberg, New York, 1995.
- [13] Morrison, N., editor. *Introduction to Fourier Analysis*. John Wiley & Sons, Inc., New York, NY, 1994.
- [14] Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L., editor. *The Analysis of Gene Expression Data: Methods and Software*. Springer-Verlag New York, Inc, New York, NY, 2003.
- [15] Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, 1969.
- [16] Stuart, R. O., Bush, K. T., and Nigam, S. K. Changes in Global Gene Expression Patterns During Development and Maturation of the Rat Kidney. *Proc. Natl. Acad. Sci. USA*, Vol. 98(10):5649–5654, May 2001.
- [17] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, Vol.17(6):520–525, 2001.
- [18] Ward, M. O. XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. In *IEEE Visualization 1994*, pages 326–336, 1994.
- [19] Zhang, L., Zhang, A., Ramanathan, M. et al. VizCluster and Its Application on Clustering Gene Expression Data. *International Journal of Distributed and Parallel Databases*, 13(1):79–97, January 2003.

Appendix: Proof for Propositions

Lemma 1

$$\left(\sum_{n=0}^{N-1} a_n \right)^2 = \sum_{n=0}^{N-1} a_n^2 + 2 \sum_{k=1}^{N-1} \sum_{t=0}^{N-k-1} a_t a_{t+k}.$$

Lemma 2 For any two complex numbers z and w , (1) $\overline{z+w} = \overline{z} + \overline{w}$, (2) $\overline{z\overline{w}} = \overline{z} w$, (3) $\overline{\overline{z}} = z$.

Lemma 3 Let $j \in \mathbb{N}$, then $\sum_{n=0}^{N-1} e^{-i2\pi jn/N}$
 $= \sum_{n=0}^{N-1} \cos(2\pi jn/N) = \sum_{n=0}^{N-1} \sin(2\pi jn/N) = \mathbf{0}$.

Lemma 4 FFHP is homomorphic: $\mathcal{F}_1(a\mathbf{x}[n] + b\mathbf{y}[n]) = a\mathcal{F}_1(\mathbf{x}[n]) + b\mathcal{F}_1(\mathbf{y}[n])$.

Proposition 1 (Cancellation) If $\mathbf{x}[n] = [a, \dots, a]$, then $\mathcal{F}_1(\mathbf{x}[n]) = \mathbf{0}$.

Proof: From the formula in Eq. (2), we get

$$\mathcal{F}_1(\mathbf{x}[n]) = \sum_{n=0}^{N-1} a e^{-i2\pi n/N} = a \sum_{n=0}^{N-1} e^{-i2\pi n/N}.$$

By Lemma 3, let $j = 1$. $\mathcal{F}_1(\mathbf{x}[n]) = \mathbf{0}$. \square

Proposition 2 (Amplitude Shifting) If $\mathbf{y}[n] = \mathbf{x}[n] + a$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])$

From the formula in Eq. (2), we get

$$\begin{aligned} \mathcal{F}_1(\mathbf{y}[n]) &= \sum_{n=0}^{N-1} (x[n] + a) e^{-i2\pi n/N} = \sum_{n=0}^{N-1} x[n] e^{-i2\pi n/N} \\ &+ \sum_{n=0}^{N-1} a e^{-i2\pi n/N} = \mathcal{F}_1(\mathbf{x}[n]) + \mathbf{0} \end{aligned}$$

The second summation is $\mathbf{0}$ by Proposition 1. \square

Proposition 3 (Amplitude Multiplying) If $\mathbf{y}[n] = a\mathbf{x}[n]$, then $\mathcal{F}_1(\mathbf{y}[n]) = a\mathcal{F}_1(\mathbf{x}[n])$.

From the formula in Eq. (2), we get

$$\begin{aligned} \mathcal{F}_1(\mathbf{y}[n]) &= \sum_{n=0}^{N-1} a x[n] e^{-i2\pi n/N} \\ &= a \sum_{n=0}^{N-1} x[n] e^{-i2\pi n/N} = a \mathcal{F}_1(\mathbf{x}[n]) \end{aligned}$$

\square

Proposition 4 (Transposing) $\mathbf{y}[n] = \mathbf{x}[N - n - 1]$, then $\mathcal{F}_1(\mathbf{y}[n]) = \overline{\mathcal{F}_1(\mathbf{x}[n])}$.

By Lemma 2, we have

$$\begin{aligned}\mathcal{F}_1(\bar{\mathbf{x}}[n]) &= \sum_{n=0}^{N-1} x[n] e^{-i2\pi n/N} = \sum_{n=0}^{N-1} \overline{x[n] e^{i2\pi n/N}} \\ &= \overline{\sum_{n=0}^{N-1} x[n] e^{i2\pi n/N}} = \overline{\mathcal{F}_1(\mathbf{x}[-n])}\end{aligned}$$

However, when $\mathbf{x}[n]$ is a real signal, $\bar{\mathbf{x}}[n] = \mathbf{x}[n]$. Then we have

$$\mathcal{F}_1(\mathbf{x}[n]) = \overline{\mathcal{F}_1(\mathbf{x}[-n])}. \text{ i.e., } \overline{\mathcal{F}_1(\mathbf{x}[n])} = \mathcal{F}_1(\mathbf{x}[-n]).$$

Since $\mathbf{x}[N-n-1] = \mathbf{x}[-n]$ then $\mathcal{F}_1(\mathbf{y}[n]) = \overline{\mathcal{F}_1(\mathbf{x}[n])}$ \square

Proposition 5 (Time Shifting) If $\mathbf{y}[n] = \mathbf{x}[n-d]$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathbf{W}_N^d \mathcal{F}_1(\mathbf{x}[n])$.

Proof: Assume $0 \leq n < N$, let $l = n-d$, then $n = l+d$. When $n=0$, $l=-d$ and when $n=N-1$, $l=N-1-d$. From the formula in Eq. (2), we get

$$\begin{aligned}\mathcal{F}_1(\mathbf{y}[n]) &= \mathcal{F}_1(\mathbf{x}[n-d]) = \sum_{l=-d}^{N-1-d} x[l] e^{-i2\pi(l+d)/N} \\ &= \sum_{l=-d}^{N-1-d} x[l] e^{-i2\pi l/N} e^{-i2\pi d/N} = \mathbf{W}_N^d \sum_{l=-d}^{N-1-d} x[l] e^{-i2\pi l/N}\end{aligned}$$

However, $e^{i2\pi n/N} = e^{i2\pi(n+N)/N}$ and $x[n] = x[n+N]$,

$$\begin{aligned}\sum_{l=-d}^{N-1-d} x[l] e^{-i2\pi l/N} &= \sum_{l=-d}^{-1} x[l+N] e^{-i2\pi(l+N)/N} \\ &+ \sum_{l=0}^{N-1-d} x[l] e^{-i2\pi l/N}\end{aligned}$$

Let $t = l+N$ for the first summation and $t = l$ for the second summation, we get

$$\begin{aligned}\sum_{l=-d}^{N-1-d} x[l] e^{-i2\pi l/N} &= \sum_{t=N-d}^{N-1} x[t] e^{-i2\pi t/N} \\ &+ \sum_{t=0}^{N-1-d} x[t] e^{-i2\pi t/N} \\ &= \sum_{t=0}^{N-1} x[t] e^{-i2\pi t/N} = \mathcal{F}_1(\mathbf{x}[n])\end{aligned}$$

Therefore, $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n]) \mathbf{W}_N^d$. \square

Definition 1 The mean of a signal $\mathbf{x}[n]$ is defined as $\hat{x} = \sum_{n=0}^{N-1} x[n]/N$. The k -th sample autocovariance coefficient of a signal $\mathbf{x}[n]$ is defined as $g_k = \sum_{n=0}^{N-1-k} (x[n] - \hat{x})(x[n+k] - \hat{x})/N$. g_0 is called the variance of $\mathbf{x}[n]$. The k -th sample autocorrelation coefficient is defined as $r_k = g_k/g_0$.

Proposition 6 (General Distance) Let $\mathbf{w}[n] = \mathbf{x}[n] - \mathbf{y}[n]$ be the difference between $\mathbf{x}[n]$ and $\mathbf{y}[n]$. The distance between $\mathcal{F}_1(\mathbf{x}[n])$ and $\mathcal{F}_1(\mathbf{y}[n])$ is

$$\|\mathcal{F}_1(\mathbf{w}[n])\|^2 = g_0 N \left(1 + 2 \sum_{k=1}^{N-1} r_k \cos(2\pi k/N) \right).$$

Proof: By Lemma 4, the distance between $\mathcal{F}_1(\mathbf{y}[n])$ and $\mathcal{F}_1(\mathbf{x}[n])$ is $\|\mathcal{F}_1(\mathbf{w}[n])\|$. From Eq. (2), we get

$$\begin{aligned}\|\mathcal{F}_1(\mathbf{w}[n])\| &= \left\| \sum_{n=0}^{N-1} w[n] e^{-i2\pi n/N} \right\| \\ &= \left\| \sum_{n=0}^{N-1} w[n] \cos(2\pi n/N) - iw[n] \sin(2\pi n/N) \right\|\end{aligned}$$

Let $\omega = 2\pi/N$, by Lemma 3, we have $\sum_{n=0}^{N-1} \cos(n\omega) =$

$\sum_{n=0}^{N-1} \sin(n\omega) = 0$. Now add a term \hat{w} , the mean of $\mathbf{w}[n]$,

$$\begin{aligned}\|\mathcal{F}_1(\mathbf{w}[n])\|^2 &= \left(\sum_{n=0}^{N-1} w[n] \cos(n\omega) \right)^2 + \left(\sum_{n=0}^{N-1} w[n] \sin(n\omega) \right)^2 \\ &= \left(\sum_{n=0}^{N-1} (w[n] - \hat{w}) \cos(n\omega) \right)^2 \\ &+ \left(\sum_{n=0}^{N-1} (w[n] - \hat{w}) \sin(n\omega) \right)^2\end{aligned}$$

Expanding each squaring term by Lemma 1, we get

$$\begin{aligned}\sum_{n=0}^{N-1} (w[n] - \hat{w})^2 (\cos^2(n\omega) + \sin^2(n\omega)) \\ + 2 \sum_{k=1}^{N-1} \sum_{t=0}^{N-1-k} [(w[t] - \hat{w})(w[t+k] - \hat{w}) \Omega]\end{aligned}$$

where $\Omega = \cos(t\omega) \cos((t+k)\omega) + \sin(t\omega) \sin((t+k)\omega)$.

By trigonometry identity $\cos \theta \cos \phi + \sin \theta \sin \phi = \cos(\phi - \theta)$, we have $\Omega = \cos(k\omega)$. Now

$$\begin{aligned}\|\mathcal{F}_1(\mathbf{w}[n])\|^2 &= \sum_{n=0}^{N-1} (w[n] - \hat{w})^2 \\ &+ 2 \sum_{k=1}^{N-1} \sum_{t=0}^{N-1-k} [(w[t] - \hat{w})(w[t+k] - \hat{w}) \cos(k\omega)] \\ &= N(g_0 + 2 \sum_{k=1}^{N-1} g_k \cos(k\omega)) \\ &= g_0 N \left(1 + 2 \sum_{k=1}^{N-1} r_k \cos(2\pi k/N) \right). \quad \square\end{aligned}$$