Enhanced Visualization of Time Series through Higher Fourier Harmonics

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

Visualization techniques can enable the exploration and detection of patterns and relationships in a complex data set by presenting the data in a graphical format in which the key characteristics become more apparent. A new visualization technique, the first Fourier harmonic projection (FFHP) was introduced to translate the multi-dimensional data into a two dimensional scatter plot where the spatial relationship of the points reflects the structure of the original data set. FFHP has been shown capable of visualizing various gene expression data sets. However, dimension arrangement is crucial for the effectiveness of FFHP and certain data, such as time series, prevents dimensions (time points) from being freely rearranged. In this paper, we present an alternative approach through higher Fourier harmonic projections to enhance the visualization. Our algorithm takes advantage of the theoretical meaning of the Fourier harmonics and "reshuffles" the dimensions of the data set without physically rearranging them. The experimental results demonstrated significant improvement of the visualizations.

Keywords

microarray, visualization, enhancement, time series, gene expression, higher Fourier harmonics

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 [8]. Extensive research has been conducted on the study of temporal patterns of gene expressions [1; 4; 11]. Visualization 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 among 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.

Visualization of microarray data is challenging because of its high dimensionality, noisy environment, and pattern varieties. We have presented a mapping for multi-dimensional data that is based on the first harmonic of the discrete Fourier transform – first Fourier harmonic projection, or FFHP in short [12; 14; 13]. It attempts to map high dimensional data points into a two dimensional space while preserving to the maximum possible extent the semantics of the data points and the intrinsic structure of the data set. FFHP has been applied on both sample and gene spaces. Our results indicated that it was capable of separating multiple types of samples [12]. Furthermore, temporal patterns were well reflected in the visualizations [13].

Dimension arrangement is crucial for the effectiveness of FFHP. Improper dimension ordering compromises the separation of substructures. For sample space visualization, a canonical dimension ordering algorithm was proposed [12]. However, certain data, such as time series, prevent dimensions (time points) from being freely rearranged. In this paper, we apply higher harmonic projections to enhance the visualization. Our algorithm takes advantage of theoretical meaning of the Fourier harmonics and "reshuffles" the dimensions of the data set without physically rearranging them. The proposed method was tested using three published, array-derived gene expression time series data sets. The results demonstrated significant improvement of the visualizations.

The remainder of this paper is organized as follows. Section 2 introduces the model of Fourier harmonic projections. The following section presents the higher harmonic approach. In Section 4, we show our experimental results. The last section discusses other issues in our approach.

2. FIRST FOURIER HARMONIC PROJECTION

Mapping

Mapping converts multi-dimensional data to two-dimensions for visualization. Time series data in its simplest form is merely a set of data $\{y_t, t = 0, \ldots, N-1\}$ where the subscript t indicates the time at which the datum y_t was observed [6]. On the other hand, a discrete-time real signal on N evenly distributed time points [3] is represented as an indexed sequence of N real numbers $0, \ldots, N-1$ denoted by $\mathbf{x}[n]$ and each term of $\mathbf{x}[n]$ is denoted by $x[n]$. The deno-

tation similarity between time series and digital signal suggests that we may 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 characterization for signals.

The frequency domain representation of discrete-time signals is through discrete-time Fourier transform, or DFT [10]. The DFT of a N-point signal $\mathbf{x}[n]$ is a frequency sequence with N complex values: $\mathcal{F}(\mathbf{x}[n]) = [\mathcal{F}_k(\mathbf{x}[n])]$, each term is called a harmonic:

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

$\mathbf{W}_N = e^{-i2\pi/N}$ is called twiddle factor. Each harmonic, $\mathcal{F}_k$, is a measurement of the kth sinusoidal frequency component in the signal: the zero harmonic 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 are complex numbers, they provide the two-dimensional point estimation for mapping a multi-dimensional signal. For this reason, we refer to the mapping as the Fourier harmonic projections. In particular, the first Fourier harmonic projection (FFHP) is:

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

The time complexity of Fourier harmonic projections is $O(N \log N)$. This is achieved by the fast Fourier transform algorithm (FFT), originally discovered by Cooley and Tukey [5]. 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 illustrated by Figure 1 [13].

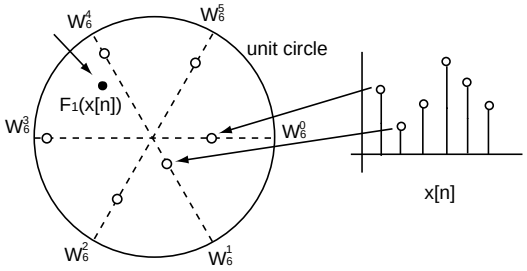


Figure 1: A geometric interpretation of the FFHP. A normalized 6-dimensional data point is shown by the stem plot. The twiddle factor divides the unit circle, centered at the origin, into 6 equal angles and each dimension of the data point is projected onto a different radial angle (open circle). The 6 projections are taken complex number sum to give a 2-dimensional image (filled circle).

For a normalized data point (the range of values of each dimension across the data set was 0 to 1) 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 kth 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. In this paper, all figures of FFP visualization are two dimensional scatter plot. The x-axis is labeled $Re[F1]$ and represents the real part of the first Fourier harmonic and the y-axis is labeled $Im[F1]$ because it represents the imaginary part of the first Fourier harmonic. The units for both axes are those of the input gene expression values.

Properties of FFHP

The FFHP has useful properties that preserve the correlation between dimensions in the multi-dimensional data point. We summarize them as propositions listed below. Equation (3) provides insight into substructure delineation capabilities of the FFHP. Points within the substructure are likely to map close to each other in the visualization. We will demonstrate that the relative locations of temporal profiles' mapping can be predicted by those propositions.

1. Data points with equal values for all the dimensions are mapped to the origin. If $\mathbf{x}[n] = [a, \ldots, a]$, then $\mathcal{F}_1(\mathbf{x}[n]) = \mathbf{0}$.

2. Data points with "amplitude-shift" by a constant are mapped to the same point. If $\mathbf{y}[n] = \mathbf{x}[n] + a$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])$. Illustrated in the left panel of Figure 2A.

3. 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 $\mathcal{F}_1(\mathbf{y}[n]) = a\,\mathcal{F}_1(\mathbf{x}[n])$. Illustrated in the right panel of 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 "time-shifted" by d dimensions relative to each other are mapped to the circumference of the circle concentric with the unit circle and the angle between them is $\phi = 2\pi d/N$. If $\mathbf{y}[n] = \mathbf{x}[n-d]$, then $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])W_N^d$. 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:

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

Detailed mathematical derivations of proposition 5 and 6 are listed in the appendix. For more discussions about those properties see [13].

3. HIGHER FOURIER HARMONIC PROJECTIONS

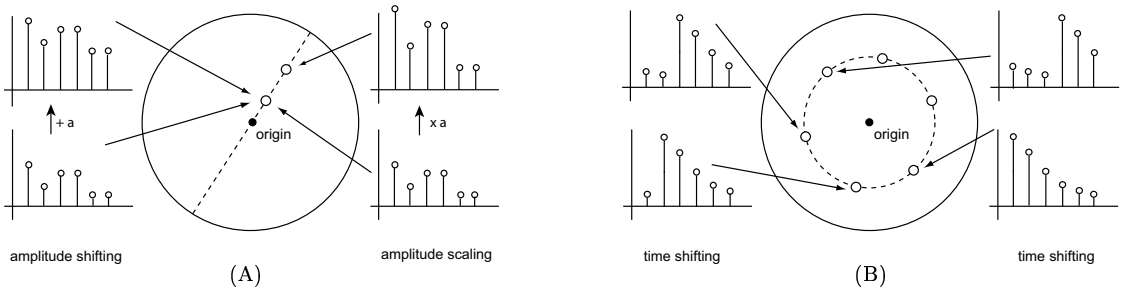


Figure 2: (A) Illustration of amplitude shifting and scaling effect of FFHP. (B) Illustration of time shifting effect of FFHP.

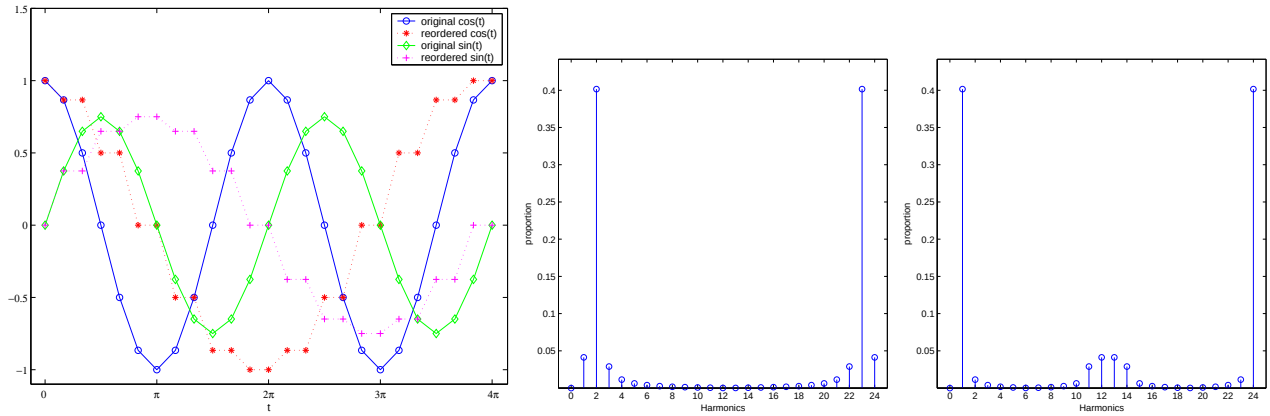


Figure 3: The effect of the second harmonic twiddle power index. (A) Two signals $\cos(t)$ and $\sin(t)$ on $t = [0, 4\pi]$ before and after being rearranged by the second harmonic twiddle power index. (B) Proportion of each of the harmonic for the original $\cos(t)$ signal. Clearly the second harmonic is the dominant component. (C) The proportion of each of the harmonic after the $\cos(t)$ signal being reordered by the second HTPI. It makes the first harmonic the most dominant component.

Harmonic Equivalency

The reason we focus specifically on the first Fourier harmonic projection is due to a so called "harmonic equivalency". It can be shown that for any harmonic (> 1), there exits an equivalent first harmonic of the original discrete signal being properly rearranged. The proof is based on a concept which we call *harmonic twiddle power index*: for an N-point signal, the k-th harmonic twiddle power index (HTPI in short) is a permutation of the N time indices from $0, \ldots, N-1$. It corresponds to the sequence that a particular time index mapped on the ascending sorted powers of twiddle factors $\mathbf{W}_N^0, \ldots, \mathbf{W}_N^{N-1}$, by the k-th harmonic.

For a simple illustration, given any 5-point signal, the first harmonic twiddle power index (HTPI) is $[0, 1, 2, 3, 4]$, the second HTPI is $[0, 3, 1, 4, 2]$, and the third HTPI is $[0, 2, 4, 1, 3]$. Take a closer look at the second HTPI: since $\mathcal{F}_2(\mathbf{x}[n]) = \sum_{n=0}^{N-1} x_n \mathbf{W}_N^{2k}$, we have $\mathcal{F}_2(\mathbf{x}[n]) = x[0]\mathbf{W}_5^0 + x[1]\mathbf{W}_5^2 + x[2]\mathbf{W}_5^4 + x[3]\mathbf{W}_5^6 + x[4]\mathbf{W}_5^8 = x[0]\mathbf{W}_5^0 + x[1]\mathbf{W}_5^2 + x[2]\mathbf{W}_5^4 + x[3]\mathbf{W}_5^1 + x[4]\mathbf{W}_5^3$. Sort by the power of $\mathbf{W}_5$, we have $\mathcal{F}_2(\mathbf{x}[n]) = x[0]\mathbf{W}_5^0 + x[3]\mathbf{W}_5^1 + x[1]\mathbf{W}_5^2 + x[4]\mathbf{W}_5^3 + x[2]\mathbf{W}_5^4$. The sequence of the time index $[0, 3, 1, 4, 2]$ is the second harmonic twiddle power index.

HTPI is always a permutation of $0, \ldots, N-1$ even if certain harmonic does not use all powers of $\mathbf{W}_N$ for the calculation. For example, let $N = 4$, $\mathcal{F}_2(\mathbf{x}[n]) = x[0]\mathbf{W}_4^0 + x[1]\mathbf{W}_4^2 + x[2]\mathbf{W}_4^0 + x[3]\mathbf{W}_4^2 = x[0]\mathbf{W}_4^0 + x[2]\mathbf{W}_4^0 + x[1]\mathbf{W}_4^2 + x[3]\mathbf{W}_4^2$. The second HTPI is $[0, 2, 1, 3]$ even though $\mathbf{W}_4^1$ and $\mathbf{W}_4^3$ were never used.

The relationship between the k-th harmonic of the original signal and the first harmonic of the rearranged signal can be concluded easily: *any k-th harmonic of a signal ($1 < k < N$) is equivalent to the first harmonic of the the original signal whose time index be rearranged by the k-th harmonic twiddle power index.*

Effects of Harmonic Twiddle Power Index

Each harmonic is the measurement of the corresponding frequency component in the signal. The "shape" of a signal determines the contribution of its harmonics. Intuitively, very flat signal is dominated by the zero harmonic while a signal with 2 cycles, or roughly 2 peaks, is the sign of second harmonic dominance. The rearrangement of the time indices will reshape the signal and thus redistribute the contribution of its harmonics. However, a theorem discovered by Parseval stated that signal's total energy was preserved under DFT [9].

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} \|\mathcal{F}_k(\mathbf{x}[n])\|^2 \qquad (4)$$

Since $\sum_{n=0}^{N-1} |x[n]|^2$ is invariant to the order of n, Parseval's theorem suggests that signal's total sum of harmonic norms is fixed regardless the arrangement of its time indices. The

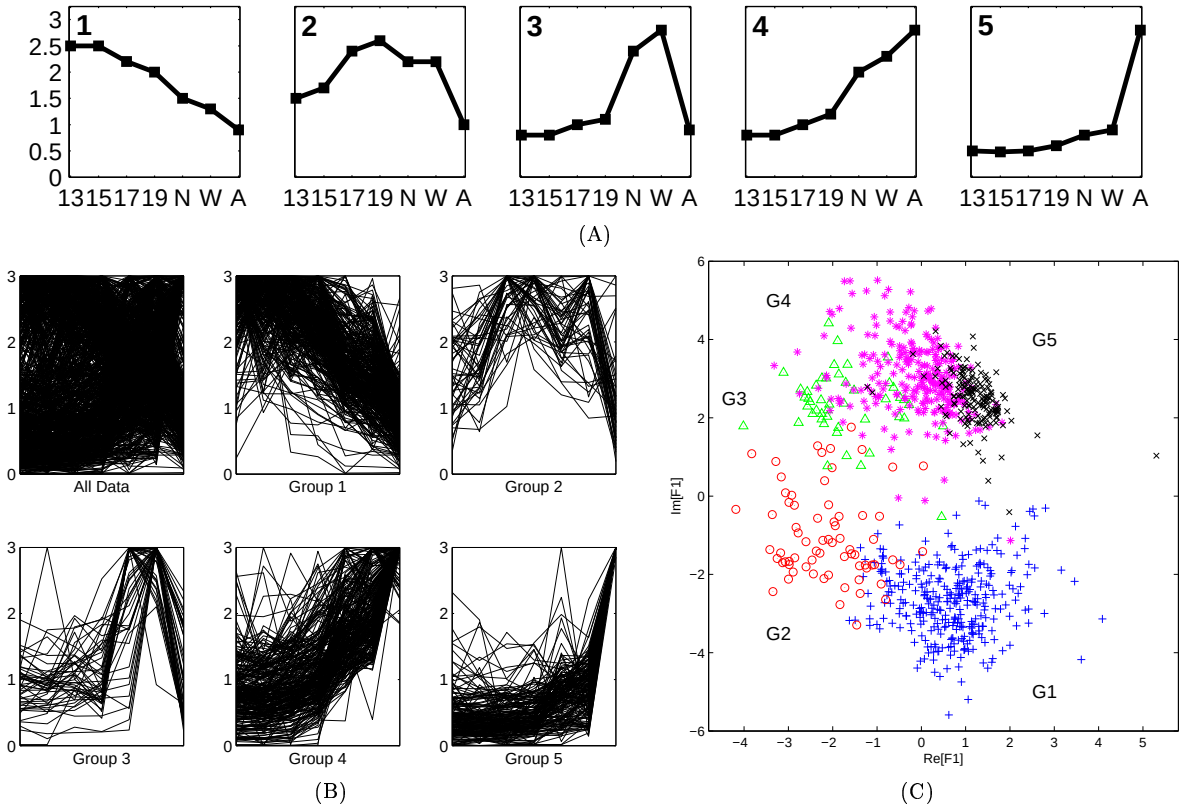


Figure 4: The *rat kidney* data set. **(A)** Idealized temporal gene expression profiles. The groups were named 1 through 5 based on the timing of their peak expression during development. 7 time points were 13, 15, 17, 19 embryonic days; N, newborn; W, 1 week old; A, adult. **(B)** Visualization in parallel coordinates for the entire data set and each of the gene groups. Patterns of genes in each group comply to the profiles depicted in (A). **(C)** Visualization under FFHP. Five gene groups were represented by blue plus symbols, red circles, green triangles, magenta stars, and black cross symbols.

arrangement of time indices affects only the proportion of each harmonic.

Harmonic equivalency suggests that higher harmonics can be considered as a systematical "reshuffling" of the dimensions rather than physically rearranging them. Figure 3 gives an insightful view of the effects of second harmonic twiddle power index. A 25-point signal simulating $\cos(t)$ on $t = [0, 4\pi]$ is depicted by a blue curve with circle marks in Figure 3A. This signal has a dominant second harmonic component (Figure 3B). Rearranging this signal by the second harmonic twiddle index results in a signal resembling $\cos(t/2)$ which is drawn as a red curve with star marks. The most dominant component in this newly generated signal is the first harmonic. Figure 3C verifies this. Since $\cos(t) = \sin(t + \pi/2)$, the situation is similar for $\sin(t)$.

Figures 3B-C give a concise view of the contribution of each harmonic in one signal. For a set of signals, we take the histogram of such distribution to measure the harmonic distribution in this set. We call it the *harmonic spectrum*.

We have stated that points within substructure are likely to map close to each other in the visualization. However, overlapping may occur due to the "opposite cancellation" effect. Recall in Figure 1, each dimension value is sequentially mapped onto the vector defined by the sequential powers of the twiddle factor. A multiple cycle-liked pattern will cause the dimension values mapped cancel each other. The first harmonic projection can minimize this "opposite cancellation" thus maximize the separation.

We conclude this subsection with one more fact: due to symmetry, not all harmonic projections have a distinct layout. In fact, half of them do not: $\mathcal{F}_k(\mathbf{x}[n])$ and $\mathcal{F}_{N-k}(\mathbf{x}[n])$ are conjugate each other. In other words, they are symmetric to the real axis, $\mathcal{F}_k(\mathbf{x}[n]) = \overline{\mathcal{F}_{N-k}(\mathbf{x}[n])}$.

Higher Harmonic Projection Approach

Our visualization paradigm is finding two-dimensional point characterization for a signal, in this case, discrete Fourier harmonics. From the above discussion we conclude that applying first Fourier harmonic projection is more appropriate if the proportion of the first harmonic is significant and if possible, a canonical dimension ordering should be adopted to make the first harmonic dominant.

There are situations where index order has to be fixed, such as in a time series. Our approach uses in three steps: (1) calculate the harmonic spectrum of the data set, (2) find the dominant harmonic k, and (3) apply the kth Fourier harmonic projection for the visualization.

4. RESULTS

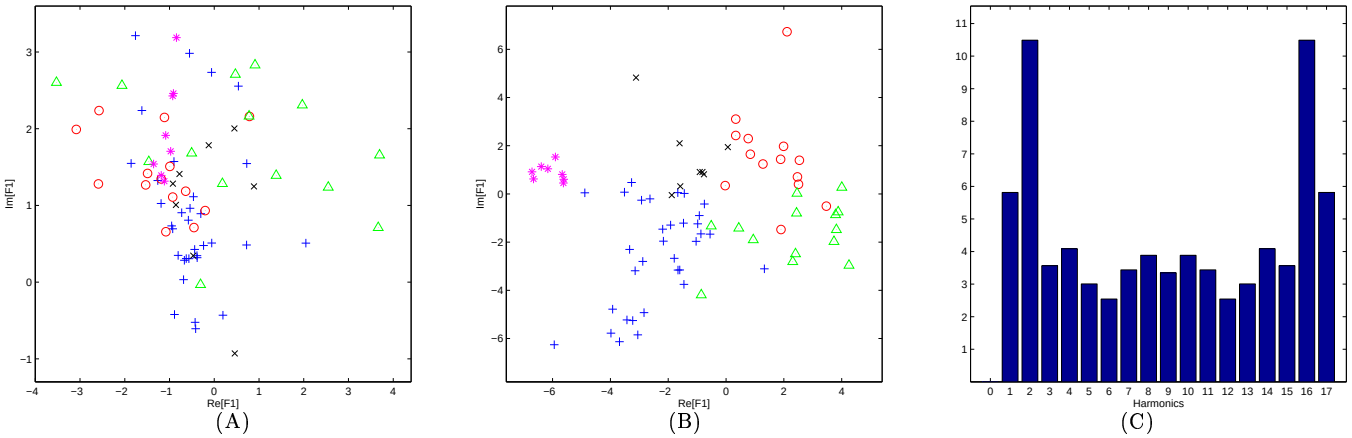


Figure 5: The *Yeast-A* data. (A) Visualization using first Fourier harmonic projection. Five temporal patterns are represented by blue plus symbols, red circles, green triangles, magenta stars, and black cross symbols. (B) Visualization using second Fourier harmonic projection. (C) The harmonic spectrum reveals the dominance of the second harmonic.

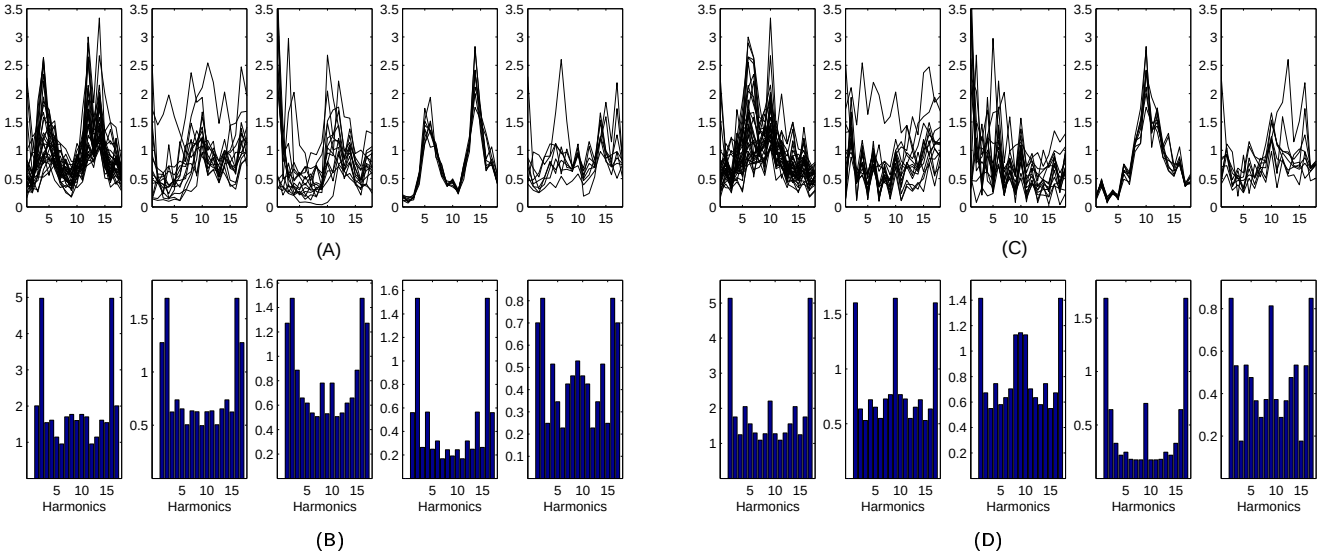


Figure 6: The detail of *Yeast-A* data. **(A)** Parallel coordinates for the 5 gene groups of the original data set. **(B)** The harmonic spectrum of those gene groups in (B). **(C)** Parallel coordinates for the gene groups in (A) reordered by the second harmonic twiddle power index. **(D)** The harmonic spectrum of corresponding gene groups in (C).

Data Sets for Visualization

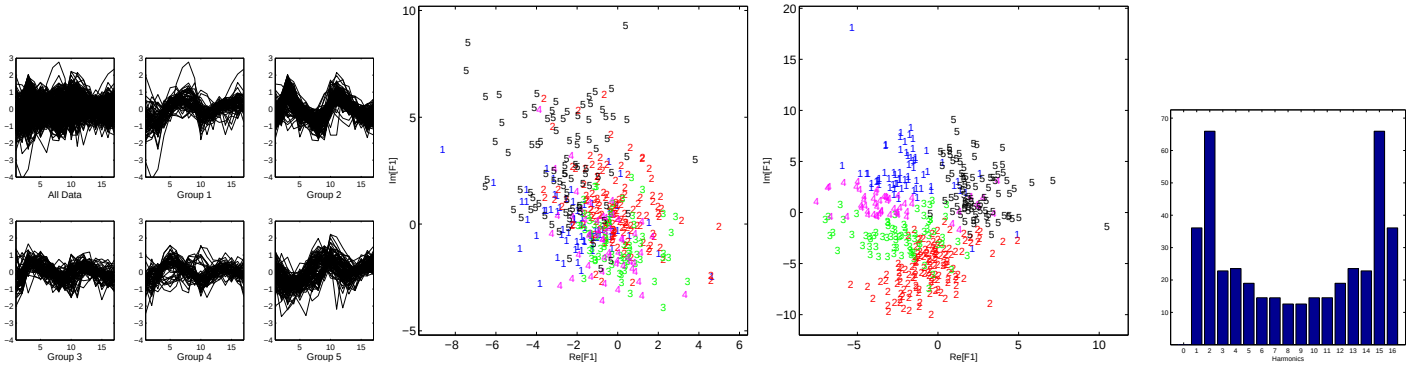
Our approach was tested using three published array-derived data sets. The *rat kidney* array data set of Stuart et al. [11] contains measurements of gene expressions during rat kidney organogenesis. The data were downloaded from *http://organogenesis.ucsd.edu/data.html*. It consists of 873 genes which 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 *yeast-A* data set of Alter et al. [1] is the result of a study of the yeast S. cerevisiae over two cell-cycle periods at 7-min intervals for 119 min (18 time points). We used a subset of 77 genes classified by traditional methods into five cell-cycle stages (by the author): MyG1, G1, S, SyG2, and G2yM. The data set were downloaded from *http://genome-www.stanford.edu/GSVD/htmls/pnas.html*.

The *yeast-B* data set from the report of Cho et al. [4] is about genome-wide characterization of mRNA transcript levels during the cell cycle of the budding yeast S. cerevisiae. It consists of 416 genes containing 17 time points. Five groups of genes are reported by the author. The data were download from *http://171.65.26.52/yeast_cell_cycle/cellcycle.html*.

Rat Kidney Dataset

To illustrate the visualization under FFHP and the roles its propositions played, we applied a 7-time point *rat kidney* data set. There are 5 discrete patterns or substructures of gene groups. Figures 4A-B show the idealized and actual gene expression profiles. Figures 4C shows the visualization under FFHP: colorized scatter plot reflecting the structure of the data. There are 5 sets of colored symbols for each of the 5 gene groups. Each symbol represents one gene across



(A) (B) (C) (D)

Figure 7: The *Yeast-B* data. (A) Parallel coordinates for the entire data set and each of the groups. (B) Visualization using first Fourier harmonic projection. Five colored letters represent genes in each of the 5 groups. (C) Visualization using second Fourier harmonic projection. (D) The harmonic spectrum reveals the dominance of the second harmonic.

7 time points. As proposition 6 suggested, genes from each group are aggregated.

Two large clusters, symmetric to the real axis, are clearly apparent from the visualization. Propositions of FFHP suggest the reason of their formation. 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 of genes that have a relatively steady increase in expression throughout development. 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 should 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. Similar arguments can be applied to the case of group 1 vs. group 2, or group 3 vs. group 4.

Yeast-A Dataset

When the first harmonic is not the dominant component, applying FFHP may yield undesirable result. This is illustrated in Figure 5A: overlapping occurred in the visualization and the separation of different temporal patterns was very poor. The harmonic spectrum in Figure 5C indicated that the dominant component was the second harmonic. Closer inspection of the temporal patterns in Figure 6A revealed that evenly spread 2-peak shapes were apparent (especially in the first and fourth panel). This was the signature of a signal dominated by the second harmonic. The observation was confirmed by the harmonic spectrum of those gene groups (Figure 6B).

Based on the harmonic spectrum, the second Fourier harmonic projection was applied to the *Yeast-A* data set, shown in Figure 5B. Compared with Figure 5A, the group separation improved significantly. Figures 6B-C gave an insightful view of the "scene behind". Recall in the previous section, we have shown that applying the second harmonic projection is equivalent to applying the first harmonic projection on the data set whose dimensions reordered by the second HTPI. Figure 6C showed the *Yeast-A* data set with rearranged dimensions (time points) by the second HTPI. Previous 2-peak shapes turned into 1-peak like shapes. This was the sign of the first harmonic dominance. As confirmed in Figure 6D, the first harmonic indeed became much more dominant.

Yeast-B Dataset

A similar situation occurs in *Yeast-B* data set. It exhibits second harmonic dominant patterns shown in Figure 7A. The harmonic spectrum graph (Figure 7D) confirms this. As expected, second harmonic projection greatly improves the gene group separation illustrated in Figures 7B-C.

Third Harmonic Projection

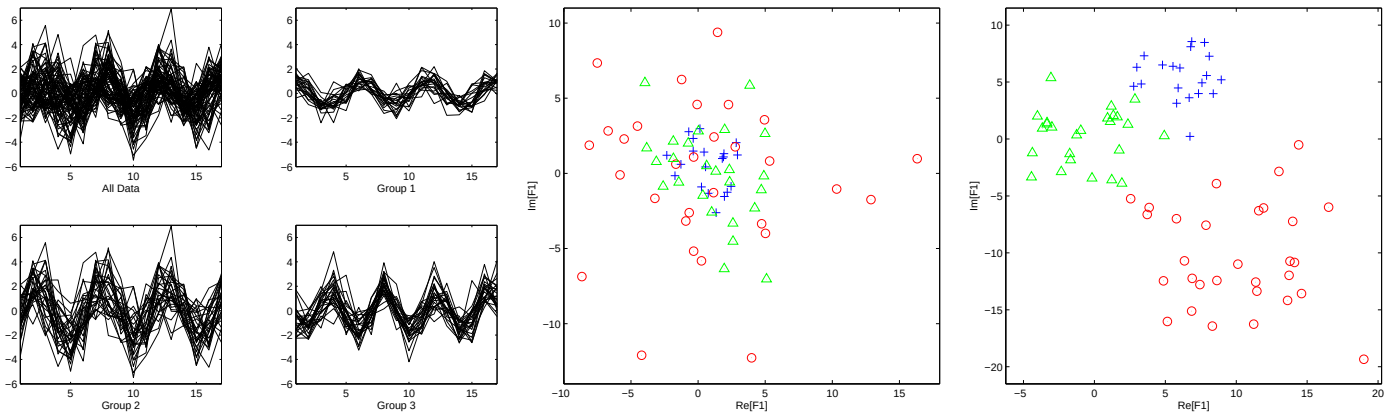
We conclude the experiments with a situation using the third harmonic projection. A 17-time point synthetic data set is generated which has 3 patterns (20, 30, and 25 data points each). Figure 8A shows the data set in parallel coordinates. Applying the first harmonic projection (Figure 7B), the visualization has noticeable overlapping. However, applying the third harmonic projection, shown in Figure 8C, the separation is improved dramatically.

5. DISCUSSION

Dimension arrangement affects a large number of visualization techniques. Previous work demonstrated that the desirable condition for applying FFHP to achieve maximized separation of substructures was when the data set had low fluctuation – the first harmonic was the dominant component. If possible, dimensions should be reordered to make the first harmonic dominant.

Ankerst et al. have shown that the general problem of finding the optimal one- and two-dimensional arrangement is NP-complete [2]. In this paper, we propose using higher Fourier harmonic projections to enhance visualization of time-series. Our method can be viewed as a heuristic approach for optimal dimension arrangement with a sound theoretical basis even though dimensions are not physically moved.

The first harmonic projection (FFHP) does not require an implicitly underlying assumption such that data set has some periodicity. In fact, it works better when no such periodicity exists. As long as patterns are relatively distinct, the first harmonic is dominant, and with little noise, FFHP can yield good visualization layout. Higher harmonic projections significantly improve the substructure separation when



(A) (B) (C)

Figure 8: Synthetic data set. (A) Parallel coordinates for the entire data set and each of the groups. Data has 3 peaks indicating third harmonic is the dominant component. (B) Visualization using first Fourier harmonic projection. Blue plus symbols, red circles, and green triangles represent 3 groups. (C) Visualization using third Fourier harmonic projection.

patterns are basically periodic. Even though the improvements were not measured numerically, the visualization layouts provided overwhelming evidence.

The results illustrate some of the weakness of Fourier harmonic projections. When the data set has large number of patterns, the sheer complexity of the data would make the visualization difficult to interpret. In more complex situations such as a set of data is made of different experiments and each experiment has a different time scale, or a subset of the data is a time series and other subset are not, or data are not necessarily measured in evenly separated time period, directly applying FHPs may not produce satisfactory results. More preprocessing steps are needed.

Two-dimensional visualizations under FFHP mapping are identical to those of radial coordinate visualization techniques, e.g., RadViz [7]. However, rather than the vector notation and the *spring paradigm* of RadViz, substantive reformulation of the mapping provides valuable theoretical insights not only allows properties of the mapping to be easily derived but also offers possible extensions.

Our experiments demonstrated that Fourier harmonic projections offers an alterative format of visualization. We believe that using projections alone or combining with heat plot or parallel coordinates would give biologist more powerful tools for analyzing and visualizing microarray data sets.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Science Foundation.

6. REFERENCES

[1] Alter, O., Brown, P. O., and Botstein, D. Generalized Singular Value Decomposition for Comparative Analysis of Genome-Scale Expression Data Sets of Two Different Organisms. *Proc. Natl. Acad. Sci. USA*, Vol. 100(6):3351–3356, March 2003.

[2] Ankerst M., Berchtold S., Keim D. A. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In *Proc. Information Visualization*, Phoenix, AZ, 1998.

[3] Cadzow, J. A., Landingham, H. F. *Signals, Systems, and Transforms*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1985.

[4] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, Vol. 2(1):65–73, July 1998.

[5] Cooley, J. W. and Tukey, J. W. An Algorithm for The Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 19(90):297–301, 1965.

[6] Diggle, P. J. *Time Series: A Biostatistical Introduction*. Oxford University Press, Oxford OX2 6DP, 1990.

[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] 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.

[9] Moon, T. K., and Stirling, W. C. *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, Inc., Upper Saddle River, NJ, 2000.

[10] Morrison, N., editor. *Introduction to Fourier Analysis*. John Wiley & Sons, Inc., New York, NY, 1994.

[11] 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.

[12] Zhang, L., Zhang, A., and Ramanathan, M. Visualized Classification of Multiple Sample Types. In *BIOKDD 2002, 2nd Workshop on Data Mining in Bioinformatics, in conjunction with 8th ACM SIGKDD*, pages 55–62. Edmonton, Alberta, Canada, July 2002.

[13] Zhang, L., Zhang, A., and Ramanathan, M. Fourier Harmonic Approach for Visualizing Temporal Patterns of Gene Expression Data. In *IEEE Computer Society Bioinformatics Conference (CSB2003)*. Stanford, CA, August 2003.

[14] 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 of Proposition 5 and 6

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. 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 3. 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 5 (TIME SHIFTING). *Two 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 $\mathcal{F}_1(\mathbf{y}[n]) = \mathcal{F}_1(\mathbf{x}[n])W_N^d$.*

Proof: Assume $0 \le n < N$, 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),

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

$$= \sum_{l=-d}^{N-1-d} x[l]e^{-i2\pi l/N}e^{-i2\pi d/N} = \mathbf{W}_N^d \sum_{l=-d}^{N-1-d} x[l]e^{-i2\pi l/N}$$

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,

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

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

Definition 1. The mean of a signal $\mathbf{x}[n]$ is defined as $\widehat{x} = \sum_{n=0}^{N-1} x[n]/N$. The k-th sample autocovariance coefficient of a signal $\mathbf{x}[n]$ is defined as $g_k = \sum_{n=0}^{N-1-k} (x[n] - \widehat{x})\,(x[n+k] - \widehat{x})/N$. 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: From Eq. (2),

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

Let $\omega = 2\pi/N$, by Lemma 2, we have $\sum_{n=0}^{N-1}\cos(n\omega) = \sum_{n=0}^{N-1}\sin(n\omega) = 0$. Now add a term $\widehat{w}$, the mean of $\mathbf{w}[n]$,

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

Expending each squaring term by Lemma 1,

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

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

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

□