The Korean Communications in Statistics **Vol. 15** No. 1, 2008, pp. 65–72

Andrews' Plots for Extended Uses

Il Youp Kwak[1] and Myung-Hoe Huh[2]

Abstract

Andrews (1972) proposed to combine trigonometric functions to represent n observations of p variates, where the coefficients in linear sums are taken from the values of corresponding observation's respective variates. By viewing Andrews' plot as a collection of n trajectories of p-dimensional objects (observations) as a weighting point loaded with dimensional weights moves along a certain path on the hyper-dimensional sphere, we develop graphical techniques for further uses in data visualization. Specifically, we show that the parallel coordinate plot is a special case of Andrews' plot and we demonstrate the versatility of Andrews' plot with a projection pursuit engine.

Keywords: Andrews' plot; graphics; parallel coordinate plot; projection pursuit.

1. Introduction

For graphical representation of p-dimensional n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of which $\mathbf{x}_i = (x_{i1}, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, Andrews (1972) proposed to use

$$f_{\mathbf{x}_i}(t) = x_{i1}\, 2^{-1/2} + x_{i2} \sin t + x_{i3} \cos t + x_{i4} \sin 2t + x_{i5} \cos 2t + \cdots$$

for $0 \leq t < 2\pi$. Figure 1.1 is the Andrews' plot for a subset of the Olive Oil data with eight variables (Unwin *et al.*, 2006; Cook and Swayne, 2007).

[1] Graduate Student in Master's Course, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : iykwak@korea.ac.kr

[2] Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
Correspondence : stat420@korea.ac.kr

* Authors are willing to send R scripts for drawing pictures contained in this paper to anyone interested.

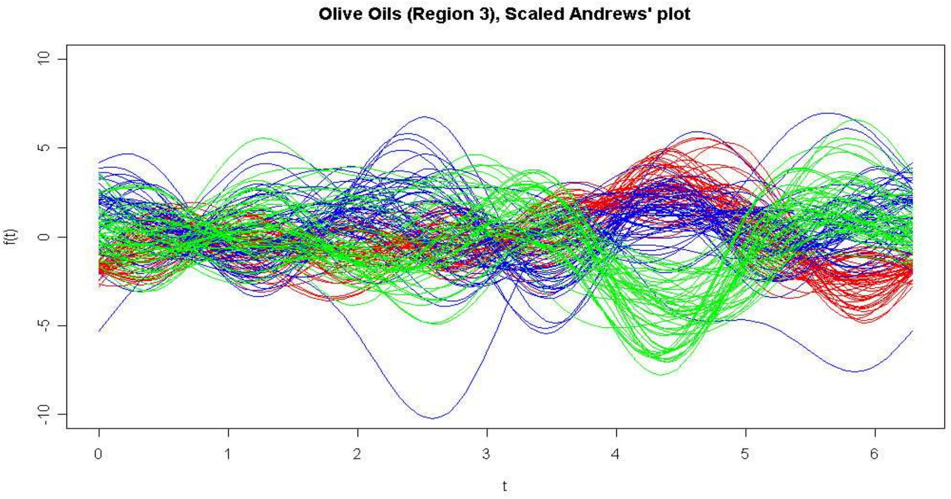

Figure 1.1: Andrews' plot for Region 3 subset of the Olive Oil data that has $n = 151$ observations with $p = 8$ variables (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic). Variables are standardized to have mean 0 and standard deviation 1.

Andrews' scheme possesses at least two nice properties (Embrechts and Herzberg, 1991). First, Andrews' function for the mean vector $\bar{\mathbf{x}}$ is equal to the average of individual Andrews' functions. That is,

$$f_{\bar{\mathbf{x}}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(t).$$

Second, Euclidean distance between observations are preserved in the functional space in the sense that

$$\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 = \frac{1}{\pi}\int_0^{2\pi} (f_{\mathbf{x}_i}(t) - f_{\mathbf{x}_{i'}}(t))^2 dt.$$

But, in practice, the plot of Andrews' functions $f_{\mathbf{x}_i}(t), i = 1, \ldots, n$ for $t \in [0, 2\pi)$ have several shortcomings in visualization of the multivariate dataset. Above all, the plot's outlook appears somewhat differently depending on the input order of p variables. Also, the plot may not reveal interesting features hidden in the data.

In this study, we view Andrews' plot as a collection of n trajectories of p-dimensional objects (observations) as a weighting point loaded with dimensional weights moves along a certain designated path on the hyper-dimensional sphere.

Such conceptualization leads to extended uses in displaying the multi-dimensional dataset. For instance, we generate the parallel coordinate plots as a special case of Andrews' plots and enhance Andrews' plots by adding a projection pursuit engine to move along better scenic paths on the sphere.

2. Andrews' Paths and Parallel Coordinate Plot

Define a "basic" Andrews' path by

$$\mathbf{a}_{2k+1}(t) = \left(\frac{1}{\sqrt{2}}, \sin t, \cos t, \sin 2t, \cos 2t, \ldots, \sin kt, \cos kt\right)'$$

for $0 \leq t < 2\pi$. Then $\|\mathbf{a}_{2k+1}(t)\|^2 = (2k+1)/2$, not varying with t. Therefore, Andrews' path $\mathbf{a}_{2k+1}(t)$ for $0 \leq t < 2\pi$ is a continuous smooth curve on a $2k+1$ dimensional sphere with fixed radius. Also, we may note that, for odd p $(= 2k+1)$, Andrews' functions are given by

$$f_{\mathbf{x}_i}(t) = \langle \mathbf{a}_p(t), \mathbf{x}_i \rangle \text{, for } i = 1, \ldots, n,$$

where $\langle \mathbf{a}, \mathbf{x} \rangle$ denotes the inner product between two p-dimensional vectors $\mathbf{a}$ and $\mathbf{x}$.

For even p $(= 2k)$, we augment p-dimensional observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to $p + 1$ dimensional $\mathbf{x}_1^+, \cdots, \mathbf{x}_n^+$, by inserting one blank component to the front of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. That is,

$$\mathbf{x}_i^+ = (0, x_{i1}, x_{i2}, \ldots, x_{ip}), \text{ for } i = 1, \ldots, n.$$

Hence, for even p $(= 2k)$,

$$f_{\mathbf{x}_i}(t) = \left\langle \mathbf{a}_{p+1}(t), \mathbf{x}_i^+ \right\rangle \text{, for } i = 1, \ldots, n.$$

Therefore, Andrews' functions are projections of multi-dimensional observation vectors on equal-length weighting vectors $\mathbf{a}_p(t)$ in Andrews' path. Hereafter, we assume that p is odd and we use notations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for $\mathbf{x}_1^+, \ldots, \mathbf{x}_n^+$.

Now, we question how many pairwise orthogonal vectors can be found in Andrew's path? The answer is p (which is assumed to be odd) and one set of orthogonal vectors are $\mathbf{a}_p(t_k)$, where $t_k = 2\pi(k-1)/p$ for $k = 1, \ldots, p$. Since it can be shown without much difficulty, we will omit the proof (However, it does not hold for even p). Therefore, for odd p, if properly scaled,

$$\mathbf{a}_p(t_1), \ldots, \mathbf{a}_p(t_p)$$

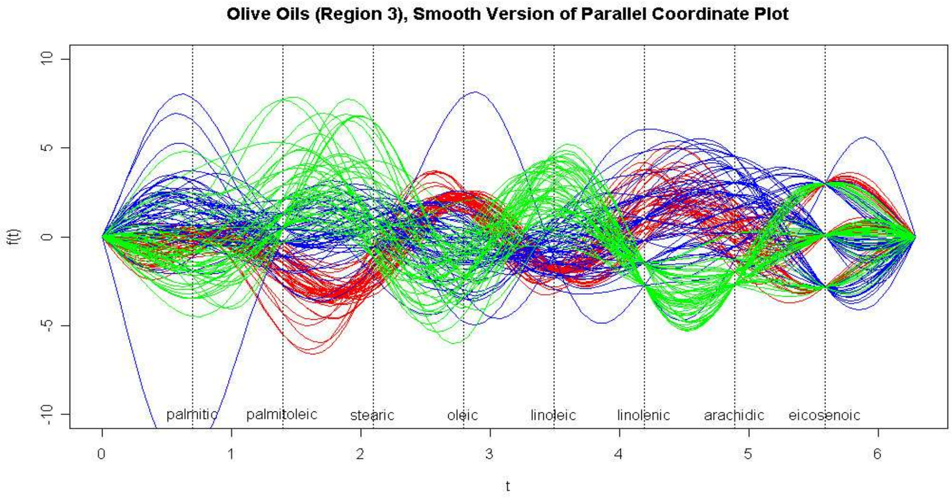

Figure 2.1: Andrews' plot as a smooth version of parallel coordinate plot for the Olive Oil (Region 3) data.

forms an orthonormal basis for the p-dimensional Euclidean space.

Now, we consider a "rotated" Andrews' path by

$$\mathbf{b}_p(t) = U_p \ \mathbf{a}_p(t)\,, \tag{2.1}$$

where

$$U_p = \begin{pmatrix} a_p(t_1)'/\sqrt{p/2} \\ \vdots \\ a_p(t_p)'/\sqrt{p/2} \end{pmatrix}$$

is a $p \times p$ orthonormal matrix. Then, $\mathbf{b}_p(t)$ for $0 \le t < 2\pi$ is an orthonormal transform of the basic Andrews' path and

$$(\mathbf{b}_p(t_1), \ldots, \mathbf{b}_p(t_p)) = \sqrt{p/2}\ I_p\,,$$

where I_p denotes the $p \times p$ identity matrix. Therefore, Andrews' functions along the rotated Andrews' path (2.1)

$$f_{\mathbf{x}_i}(t) = \langle \mathbf{b}_p(t), \mathbf{x}_i \rangle\,, \text{ for } \ i = 1, \ldots, n$$

yield parallel coordinate plot (PCP) of p input variables. Figure 2.1 shows a "rotated" Andrews' plot that is almost identical to PCP. There are two differences compared to conventional PCP. First, we now have smoothly connected curves instead of piecewise linear lines. In PCP, the former type is considered the better one compared to the latter type (Huh and Park, 2008). Second, all curves meet at $t = 0$ and $t = 2\pi$, since the "zero" variable is inserted into the dataset (of eight variables) to make p, the number of variables, odd.

3. Projection Pursuit via Andrews' Plots

In this section, we utilize Andrews' plot more fully to produce a "projection pursuit" diagram. Now we sketch our technique.

To extend the idea of a rotated path of Section 2, define a "randomly rotated" Andrews' path by

$$\mathbf{c}_p(t) = V_p \ \mathbf{a}_p(t) \, ,$$

where V_p is a $p \times p$ randomly generated orthonormal matrix. And, define "randomly rotated" Andrews' functions by

$$f_{\mathbf{x}_i}(t) = \langle \mathbf{c}_p(t), \mathbf{x}_i \rangle \, , \ \text{for} \ \ i = 1, \ldots, n. \tag{3.1}$$

Graphs of Andrews' functions (3.1) over $0 \leq t < 2\pi$, which we call a "randomly rotated" Andrews' plots, may reveal various features of the multivariate dataset.

"Randomly rotated" Andrews' plot shows the projected trajectories (3.1) over t. For each specific randomly generated Andrews' plot, we may select t_0 between 0 and 2π, at which the distribution of projections $f_{\mathbf{x}_1}(t), \ldots, f_{\mathbf{x}_n}(t)$ becomes the most "interesting". Following a standard convention of projection pursuit methodology (Cook and Swayne, 2007), we will adapt the holes index I_{Holes} to one dimensional projections. It is expressed as

$$I_{Holes}(y_1, \ldots, y_n; V_p) = \frac{1 - \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \exp\left(-\dfrac{1}{2} y_i^2\right)}{1 - \exp\left(-\dfrac{p}{2}\right)},$$

where $y_i = f_{\mathbf{x}_i}(t), \ i = 1, \ldots, n$. Once we determined t_0 at which the index is maximized over t, save $\mathbf{c}_p(t_0) = V_p \ \mathbf{a}_p(t_0)$ for later use.

We repeat the above process over m $(= 100)$ times, to exploit more. We denote $\mathbf{c}_p^*$ for the winner among $\mathbf{c}_p(t_0)$'s and write

$$y_i^* = \left\langle \mathbf{c}_p^*, \mathbf{x}_i \right\rangle , \ \text{for} \ i = 1, \ldots, n.$$

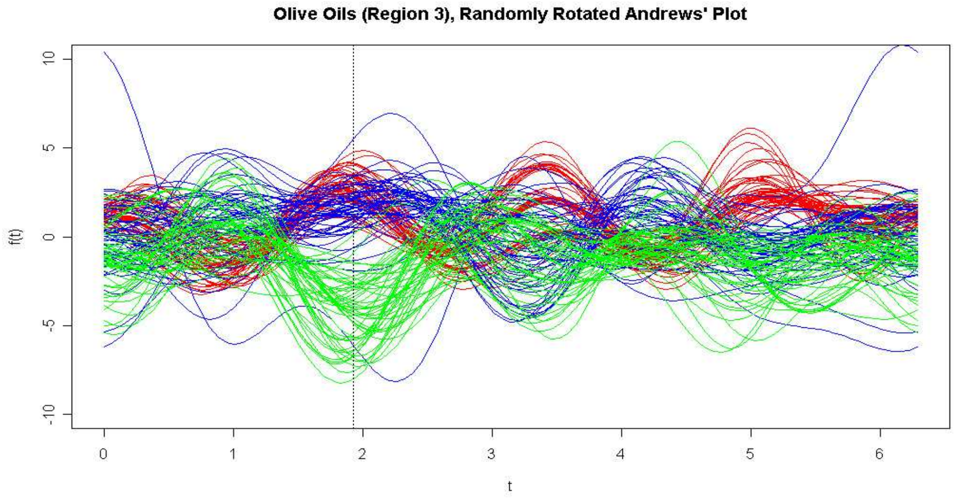

Figure 3.1: “Randomly rotated” Andrews’ plot for Olive Oil (Region 3) data. The dotted line shows the optimal point at which the holes index is maximized.

Figure 3.1 is a “randomly rotated” Andrews’ plot for Olive Oil data (Region 3). The value at dotted line is the optimal point $\mathbf{c}_p^*$ at which the holes index is maximized.

Put

$$\mathbf{x}_i^* = \left(I_p - \mathbf{c}_p^* \mathbf{c}_p^{*t} / \left\|\mathbf{c}_p^*\right\|^2\right) \mathbf{x}_i \ (= (I_p - H_p)\, \mathbf{x}_i\,),\ i = 1, \ldots, n$$

and consider an independent “randomly rotated” Andrews’ path

$$\mathbf{d}_p(t) = W_p\ \mathbf{a}_p(t),$$

where W_p is a $p \times p$ randomly generated orthonormal matrix. Then, it holds

$$\langle\, (I_p - H_p)\, \mathbf{d}_p(t), \mathbf{x}_i \rangle = \langle \mathbf{d}_p(t), (I_p - H_p)\, \mathbf{x}_i \rangle\,,\ \text{for } i = 1, \ldots, n.$$

Hence “randomly rotated” Andrews’ function for each $\mathbf{x}_i$ along the path $\mathbf{d}_p(t)$ orthogonal to $\mathbf{c}_p^*$ is identical to that for $\mathbf{x}_i^*$ along the path $\mathbf{d}_p(t)$. Thus, we repeat the process with residuals $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ from the first round to obtain the second dimensional visualization. This second round produces the optimal projection vector $\mathbf{d}_p^*$ and projections of $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ on $\mathbf{d}_p^*$. We denote them by

$$z_i^* = \left\langle \mathbf{d}_p^*, \mathbf{x}_i^* \right\rangle\,,\ \text{for } i = 1, \ldots, n\,.$$

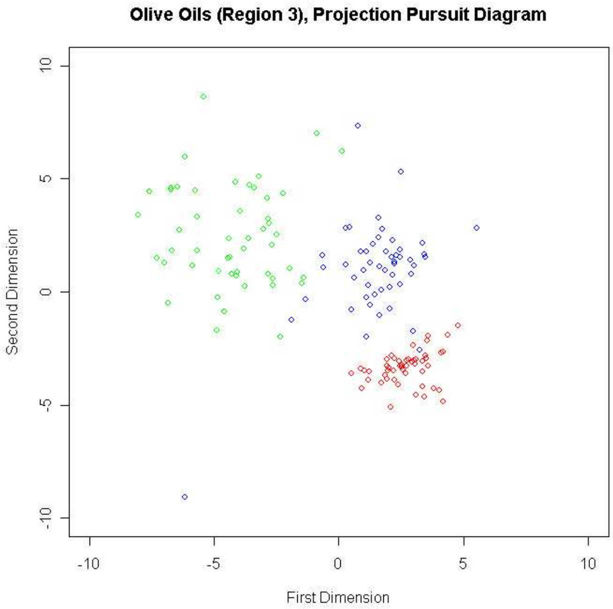

Figure 3.2: Projection pursuit diagram of the Olive Oil (Region 3) data via Andrews' plot. Colors represent three subregions of sample origins.

Finally, plot (y_i^*, z_i^*), $i = 1, \ldots, n$. The picture may be considered as a "projection pursuit" for p-dimensional n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Figure 3.2 is a projection pursuit diagram of the Olive Oil data (Region 3) via Andrews' plot. Colors represent three subregions of sample origins. We now see that three subregions are well separated.

4. Concluding Remarks

This study is aimed to broaden the use of the Andrews' plot which relies on Fourier orthogonal functions rather than Euclidean axes. By extending the concept of Andrews' path, we have shown that the parallel coordinate plot can be obtained as a special case of Andrews' plot and we even obtained a "projection pursuit" diagram via Andrews' plots. In generation of such plots, we used a *brute force* Monte Carlo optimization which needs to be refined.

References

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, **28**, 125–136.

Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*. Springer, New York.

Embrechts, P. and Herzberg, A. M. (1991). Variations of Andrews' plot. *International Statistical Review*, **59**, 175–194.

Huh, M. H. and Park, D. Y. (2008). Enhancing parallel coordinates plots, To appear in *Journal of the Korean Statistical Society*.

Unwin, A. Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets: Visualizing a Million*. Springer, New York.

[Received 08 2007, Accepted 11 2007]