응용통계연구 (2008)
21(4), 695–704

통계적 그래픽스 도구로서의 정다각기둥평행좌표그림

장대홍[1]

[1]부경대학교 수리과학부 통계학전공

(2008년 4월 접수, 2008년 5월 채택)

요 약

평행좌표그림은 다변량자료를 시각화하는 하나의 방법이다. 평행좌표그림은 4차원 이상의 직각좌표계 표시의 어려움을 극복할 수 있는 그림이다. 그러나 변수 축의 배열에 따라 같은 자료에 대하여도 다른 해석이 가능하다. 변수 선택 문제를 해결하는 한 가지 방법으로서 우리는 정다각기둥평행좌표그림을 제안할 수 있다.

주요용어: 평행좌표그림, 정다각기둥평행좌표그림, 다정다각기둥평행좌표그림.

1. 서론

자료의 크기를 n이라 하고 변수의 개수를 p라 할 때 자료행렬을 (x_{ij}), $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$로 나타낼 수 있다. 이러한 자료행렬에 대한 탐색적자료분석 단계에서 우리는 그래픽 방법들을 자주 이용하게 된다. grand tour, 체르노프얼굴(Chernoff face), glyph, Andrews 곡선, 레이더차트(radar chart, star diagram), 회전(rotation, spin plot), 산점도행렬, 평행좌표그림(parallel coordinate plot), 성좌그림(constellation graph), 원추그림(coneplot) 등이 있다. 이외에도, 수리적이고, 기하학적인 이론이 첨부된 그림인 biplot, 다차원척도법, 대응분석, 주성분그림 등이 있다. 이들 그림들 중 평행좌표그림은 Inselberg (1985)가 구체적으로 제안한 이후 최근까지도 많은 학문영역에서 다양하게 쓰이고 있다 (Gennings 등, 1990; Inselberg와 Dimsdale, 1990; Wegman, 1990; Madhavan 등, 1991; Miller와 Wegman, 1991; Lee 등, 1995; Bateson과 Curtiss, 1996; Keim과 Kriegel, 1996; Lee와 Ong, 1996; Weber와 Desai, 1996; Becker, 1997; Inselberg, 1998, 2002; Ankerst 등, 1998; Teppola 등, 1998; Chou 등, 1999; Fua 등, 1999a, 1999b; Goel 등 1999; Groller 등, 1999; King과 Harris, 1999; Hall과 Berthold, 2000; Siirtola, 2000; Theisel, 2000; Andrienko와 Andrienko, 2001; Falkman, 2001; Hauser 등, 2002; Heyden 등, 2002; Bert-hold와 Hall, 2003; Edsall, 2003; Graham과 Kennedy, 2003; Unwin 등, 2003; Hurley, 2004; Ye와 Lin, 2006; Albazzaz와 Wang, 2006; Moustafa와 Wegman, 2006; Huh와 Park, 2008; Kumasaka와 Shibata, 2008; Kwak와 Huh, 2008). 다음 그림 1.1은 iris 자료에 대한 평행좌표그림이다. 우리는 대략 두 개의 집락을 확인할 수 있다. 즉 type 1이 첫 번째 집락을 이루고 type 2와 type 3이 두 번째 집락을 이룸을 알 수 있다.

그러나 이 평행좌표그림에 대하여 다음과 같은 3가지 문제점이 발생한다.

1. 데이터가 많은 경우 그림의 뭉개짐(over-plotting)

이 논문은 2006학년도 부경대학교 기성회 학술연구비에 의하여 연구되었음.
[1](608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수.
E-mail: dhjang@pknu.ac.kr

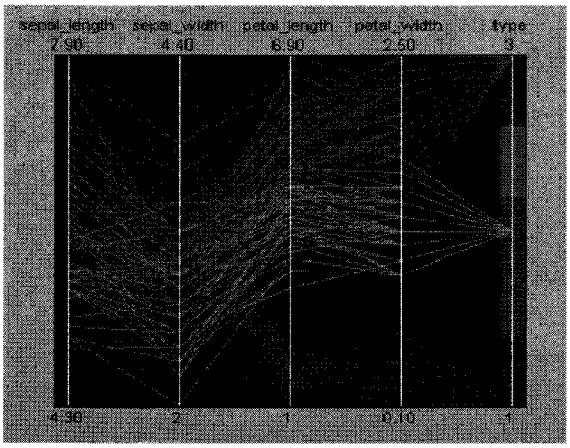

그림 1.1. IRIS 자료에 대한 평행좌표그림

2. 데이터의 식별
3. 변수의 배열

첫 번째 문제점의 해결책으로서 hierarchical parallel coordinate plot (Fua 등, 1999b), zooming (focus, Graham과 Kennedy, 2003 참조.), alpha-blending (Unwin 등, 2006) 그리고 parallel coordinate density plot (Miller와 Wegman, 1991)를 이용할 수 있고, 두 번째 문제점의 해결책으로서 smooth parallel coordinate plot (Theisel, 2000; Graham과 Kennedy, 2003; Moustafa와 Wegman, 2006; Kwak와 Huh, 2008; Huh와 Park, 2008)를 이용할 수 있다. 세 번째 문제점의 해결책으로서 축간 거리조정방법 (Huh와 Park, 2008), Endlink algorithm (Hurley, 2004) 그리고 정다각기둥평행좌표그림(regular polyprism parallel coordinate plot)을 이용할 수 있다. 정다각기둥평행좌표그림에 대한 아이디어는 McLean과 McEwen이 'milk carton plot'이라는 이름으로 제안하였다 (웹페이지 `www.scs.gmu.edu/ tmclean/parallel3d.html/`에 있었으나 2006년부터 사라짐). 본 논문에서는 이러한 정다각기둥평행좌표그림을 수학패키지 Maple을 이용하여 구현하여 보았다.

2. 정다각기둥평행좌표그림

다음 그림 2.1은 5개의 열이 있고 각 열이 연속적일양분포 $U(0,1)$에서 뽑은 랜덤데이터 50개로 구성된 50×5 자료행렬 X를 평행좌표그림으로 나타낸 그림이다. 표 2.1은 앞의 50×5 자료행렬 X에 대한 표본상관계수행렬을 나타낸다. 그림 2.1의 평행좌표그림은 변수축의 순차적인 배열 때문에 변수 1, 3 그리고 5 사이의 강한 상관관계를 나타내지 못한다. 그러므로 우리는 이웃하지 않는 변수들 사이의 관계도 알 수 있는 그래픽 수단이 필요하다. 이러한 수단으로 쓸 수 있는 그래픽도구가 정다각기둥평행좌표그림이다. 정다각기둥평행좌표그림에서 각 변수를 정다각기둥의 수직면의 모서리에 배당하고 자료는 정다각기둥의 수직면의 표면 상에 폐쇄직선으로 표시된다. 그림 2.2는 앞의 50×5 자료행렬 X에 대한 정다각기둥평행좌표그림을 나타낸다.

그림 2.2의 정다각기둥평행좌표그림에서 변수 1과 5 사이의 수직면을 통하여 변수 1과 5 사이의 상관관계가 -1인 것을 확인할 수 있다. 정다각기둥평행좌표그림에서 인접하여 있는 축 사이의 수직면을 회전시키면서 돌려 볼 수 있다. 이러한 회전을 통하여 인접하여 있는 변수들 사이의 관계를 알 수 있다. 또

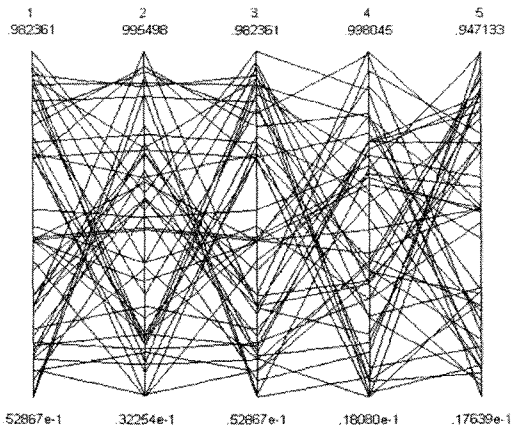

그림 2.1. 50×5 자료행렬 X에 대한 평행좌표그림

표 2.1. 50×5 자료행렬 X에 대한 표본상관계수행렬

$$\begin{pmatrix} 1 & 0.053 & 1 & -0.004 & -1 \\ 0.053 & 1 & 0.053 & 0.080 & -0.053 \\ 1 & 0.053 & 1 & -0.004 & -1 \\ -0.004 & 0.080 & -0.004 & 1 & 0.004 \\ -1.000 & -0.053 & -1 & 0.004 & 1 \end{pmatrix}$$

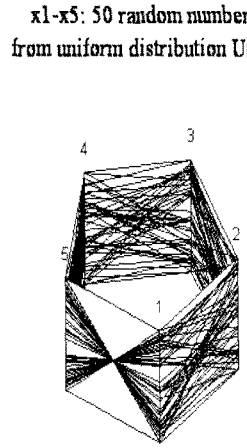

그림 2.2. 50×5 자료행렬 X에 대한 정다각기둥평행좌표그림

한 정다각기둥평행좌표그림에서 인접하여 있지 않은 축 사이의 수직면을 그려 볼 수 있다. 이를 통하여 인접하여 있지 않은 변수들 사이의 관계를 알 수 있다. 그림 2.3은 변수 1, 3 그리고 5 사이의 관계를 알기 위하여 변수 1과 3에 대응되는 수직면과 변수 3과 변수 5에 대응되는 수직면을 선택하여 나타낸 정다각기둥평행좌표그림이다. 정다각기둥평행좌표그림에서 인접하여 있지 않은 축 사이의 수직면을 회전

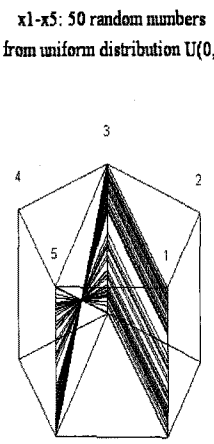

그림 **2.3.** 50×5 자료행렬 X에서 변수 1, 3, 5 사이의 관계를 알 수 있는 정다각기둥평행좌표그림

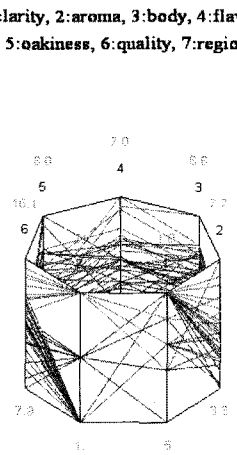

그림 **2.4.** 와인자료에 대한 정다각기둥평행좌표그림

시키면서 돌려 볼 수 있다. 이러한 회전을 통하여 인접하여 있지 않은 변수들 사이의 관계를 알 수 있다.

그림 2.3의 정다각기둥평행좌표그림에서 변수 1과 3에 대응되는 수직면과 변수 3과 변수 5에 대응되는 수직면을 통하여 변수 1과 3 사이의 상관관계가 1이고 변수 3과 5 사이의 상관관계가 -1인 것을 확인할 수 있다. 정다각기둥평행좌표그림은 위에서 살펴본 예를 통하여 알 수 있듯이 평행좌표그림의 단점인 변수의 배열 문제를 해결할 수 있는 유용한 그래픽 도구이다. 다음 그림 2.4는 통계패키지 Minitab에 나와 있는 자료인 와인자료에 대한 정다각기둥평행좌표그림이다. 이 와인자료는 38개의 관측치와 7개의 변수(clarity, aroma, body, flavor, oakiness, quality, region)를 갖고 있다.

그림 2.4에서 보는 것처럼 정다각기둥평행좌표그림에서는 정다각기둥의 각 모서리(축)에 각 변수를 대응시킨다. 즉, 7개의 변수 clarity, aroma, body, flavor, oakiness, quality, region를 정칠각기둥의 7개의 모서리에 배열하여 정칠각기둥의 각 면에 변수1-변수2, 변수2-변수3, 변수3-변수4, 변수4-변수5, 변

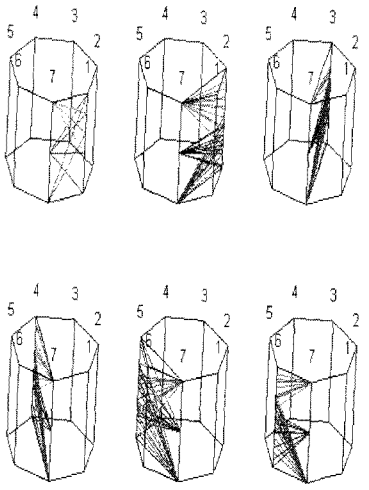

그림 **2.5.** REGION 변수에 대한 나머지 6개의 변수들 사이의 관계를 나타내는 정다각기둥평행좌표그림

1:clarity, 2:aroma, 3:body, 4:flavor,
5:oakiness, 6:quality, 7:region

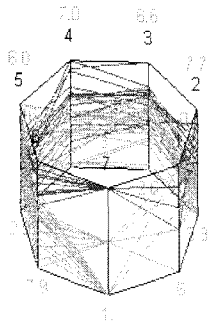

그림 **2.6.** 와인자료에 대하여 하이라이트 기능을 행한 정다각기둥평행좌표그림

수5-변수6, 변수6-변수7, 변수7-변수1를 대응시켜 평행좌표그림을 그린다. 하나의 데이터는 정칠각기둥의 각 면을 돌아가며 하나의 꺾은선으로 나타내어진다. 인접하여 있지 않은 변수들 사이의 관계는 관심있는 변수들을 가로질러 연결하여 만든 면에 데이터를 나타낸다. 예로 region이라는 변수에 대한 나머지 6개의 변수들 사이의 관계를 나타내면 다음 그림 2.5와 같다. aroma와 quality 변수에 대략 두 개의 집락이 나타남을 알 수 있다.

정다각기둥평행좌표그림에서 우리는 하이라이트(highlight) 기능과 변수선택(variable selection) 기능을 행할 수 있다. 그림 2.6은 와인자료에 대하여 하이라이트 기능을 행한 정다각기둥평행좌표그림을 나타낸다. region 변수값이 3인, 즉 세 개의 지역 중 세 번째 지역에서 산출된 와인자료만을 나타내었다. 세 번째 지역에서 산출된 와인자료만을 보라색으로 나타내고 기타 지역에서 산출되는 와인자료는 회색

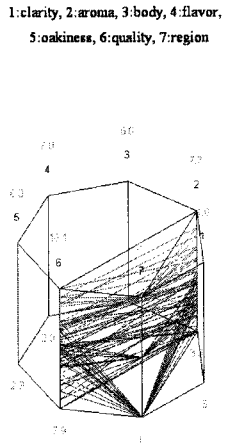

그림 **2.7.** 와인자료에 대하여 변수선택 기능을 행한 정다각기둥평행좌표그림

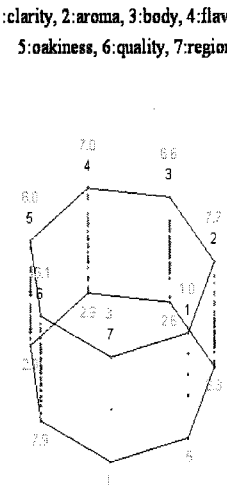

그림 **2.8.** 와인자료의 7개의 변수들에 대한 점도표를 나타내는 정다각기둥평행좌표그림

으로 나타내어 세 번째 지역에서 산출된 와인자료만을 부각시킬 수 있다.

그림 2.7은 집락이 비교적 잘 구분되는 2개의 변수(두 번째 변수 aroma와 여섯 번째 변수 quality) 만을 선택하는 변수선택 기능을 행한 정다각기둥평행좌표그림을 나타낸다. region 3에서 산출되는 와인은 aroma와 quality 값이 큰 값을 나타내고 region 1과 2에서 산출되는 와인은 aroma와 quality 값이 중간 이하의 값을 나타내고 있음을 알 수 있다.

정다각기둥평행좌표그림에서 우리는 각 기둥에 점도표를 그릴 수 있다. 그림 2.8는 와인자료의 7개의 변수들에 대한 점도표를 나타낸 정다각기둥평행좌표그림이다. 이 정다각기둥평행좌표그림을 회전시키면서 살펴보면 7개의 변수들에 대한 점도표를 자세히 볼 수 있어서 7개의 변수들 각각의 자료패턴을 확인할 수 있다.

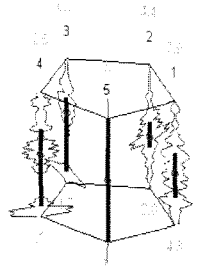

그림 **2.9.** IRIS 자료에 대한 바이올린그림

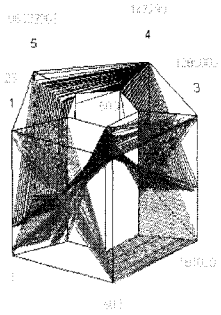

그림 **2.10.** 다정다각기둥평행좌표그림

다음 그림 2.9는 iris 자료에 대한 바이올린그림이다. 각 변수에서 확률밀도함수를 동시에 볼 수 있고 회전시키면서 살펴보면 4개의 변수들에 대한 확률밀도함수를 자세히 볼 수 있어서 4개의 변수들 각각의 확률밀도함수패턴을 확인할 수 있다. 이 바이올린그림에 iris 자료를 동시에 나타낼 수 있다. 그러면 Kumasaka와 Shibata (2008)이 제시한 textile plot과 같은 기능을 할 수 있게 된다.

위에서 살펴 본 여러 기능들을 통하여 평행좌표그림의 3차원 확장인 정다각기둥평행좌표그림은 평행좌표그림의 단점인 변수의 배열 문제를 해결할 수 있고 자료행렬의 복잡한 구조를 들여다 볼 수 있는 유용한 그래픽 도구임을 알 수 있다. 변수의 개수 p가 커지면 정다각기둥평행좌표그림은 점점 실린더 형태가 되고 변수 사이의 수직면의 폭이 점점 작아져서 자료의 패턴을 보기가 어렵게 된다. 그러므로 변수의 개수 p가 클 때는 평행좌표그림을 기본 그림으로 하고 관심대상이 되는 변수들만 뽑아 정다각기둥평행좌표그림을 그려 서로 비교하여 보는 것이 좋다고 생각한다. 정다각기둥평행좌표그림의 확장으로서 우리는 다음 그림 2.10과 같은 다정다각기둥평행좌표그림(multi-regular polyprism parallel coordinate

plot)을 제시할 수 있다. 이 그림은 2006년 7월 3일 현재 우리나라에서 주식가격이 제일 높은 3개의 대기업(롯데제과, 삼성전자, SK 텔레콤)에 대하여 2006년 7월 3일에서 2006년 8월 18일까지 날짜별로 5개의 변수(순위, 액면가격, 주식가격, 주식총량, 총시장가격)를 기준으로 작성된 정다각기둥평행좌표그림을 겹쳐 놓은 다정다각기둥평행좌표그림이다. 이 그림에서 빨간색은 롯데제과, 파랑색은 삼성전자, 녹색은 SK 텔레콤을 나타낸다. 우리는 이 삼차원그림을 회전시키면서 주식시장이 열린 33일 동안 3개의 기업이 5개의 변수에서 어떤 변화를 겪고 있는 지를 확인할 수가 있다.

3. 결론

탐색적자료분석 단계에서 이용되는 그래픽 방법들 중 평행좌표그림을 확장한 정다각기둥평행좌표그림은 원래의 평행좌표그림을 보완하는 그림으로서 탐색적자료분석시 유용한 그림도구가 될 수 있다.

참고문헌

Albazzaz, H. and Wang, X. Z. (2006). Historical data analysis based on plots of independent and parallel coordinates and statistical control limits, *Journal of Process Control*, **16**, 103–114.

Andrienko, G. and Andrienko, N. (2001). Exploring spatial data with dominant attribute map and parallel coordinates, *Computers, Environment and Urban Systems*, **25**, 5–15.

Ankerst, M., Berchtold, S. and Keim, D. A. (1998). Similarity clustering of dimensions for an enhanced visualization of multidimensional data, In *Proceedings of IEEE Symposium on Information Visualization*, 52–60.

Bateson, A. and Curtiss, B. (1996). A method for manual endmember selection and spectral unmixing, *Remote Sensing of Environment*, **55**, 229–243.

Becker, O. M. (1997). Representing protein and peptide structures with parallel-coordinates, *Journal of Computational Chemistry*, **18**, 1893–1902.

Berthold, M. R. and Hall, L. O. (2003). Visualizing fuzzy points in parallel coordinates, *IEEE Transactions on Fuzzy Systems*, **11**, 369–374.

Chou, S. Y., Lin, S. W. and Yeh, C. S. (1999). Cluster identification with parallel coordinates, *Pattern Recognition Letters*, **20**, 565–572.

Edsall, R. M. (2003). The parallel coordinate plot in action: Design and use for geographic visualization, *Computational Statistic & Data Analysis*, **43**, 605–619.

Falkman, G. (2001). Information visualisation in clinical odontology: Multidimensional analysis and interactive data exploration, *Artificial Intelligence in Medicine*, **22**, 133–158.

Fua, Y. H., Ward, M. O. and Rundensteiner, E. A. (1999a). Navigating hierarchies with structure-based brushes, In *Proceedings of IEEE Symposium on Information Visualization*, 58–64.

Fua, Y. H., Ward, M. O. and Rundensteiner, E. A. (1999b). Hierarchical parallel coordinates for exploration of large datasets, In *Proceedings of the Conference on Visualization '99*, 43–50.

Gennings, C., Dawson, K. S., Carter, W. H. Jr. and Myers, R. H. (1990). Interpreting plots of a multidimensional dose-response surface in a parallel coordinate system, *Biometrics*, **46**, 719–735.

Goel, A., Baker, C., Shaffer, C. A., Grossman, B., Haftka, R., Mason, W. H. and Watson, L. T. (1999). VizCraft: A multidimensional visualization tool for aircraft configuration design, In *Proceedings of Visualization '99*, 425–428.

Graham, M. and Kennedy, J. (2003). Using curves to enhance parallel coordinate visualizations, In *Proceedings of the Seventh International Conference on Information Visualization*, 10–16.

Groller, E., Loffelmann, H. and Wegenkittl, R. (1999). Visualization of dynamical systems, *Future Generation Computer Systems*, **15**, 75–86.

Hall, L. O. and Berthold, M. R. (2000). Fuzzy parallel coordinates, In *Proceedings of the International Conference of the North American Fuzzy Information Processing Society*, 74–78.

Hauser, H., Ledermann, F. and Doleisch, H. (2002). Angular brushing of extended parallel coordinates, In *Proceedings of the IEEE Symposium on Information Visualization*, 127–130.

Heyden, Y. V., Pravdova, V., Questier, F., Tallieu, L., Scott, A. and Massart, D. L. (2002). Parallel co-ordinate geometry and pricipal component analysis for the interpretation of large multi-response experimental designs, *Analytica Chimica Acta*, **458**, 397–415.

Huh, M. H. and Park, D. Y. (2008). Enhancing parallel coordinate plots, *Journal of the Korean Statistical Society*, **37**, 129–133.

Hurley, C. B. (2004). Clustering visualizations of multidimensional data, *Journal of Computational & Graphical Statistics*, **13**, 788–806.

Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, **1**, 69–91.

Inselberg, A. (1998). Visual data mining with parallel coordinates, *Computational Statistics*, **13**, 47–63.

Inselberg, A. (2002). Visualization and data mining of high-dimensional data, *Chemometrics and Intelligent Laboratory Systems*, **60**, 147–159.

Inselberg, A. and Dimsdale, B. (1990). Parallel coordinate: A tool for visualizing multi-dimensional geometry, In *Proceedings of Visualization '90*, 361–378.

Keim, D. A. and Kriegel, H. P. (1996). Visualization techniques for mining large databases: A comparison, *IEEE Transactions on Knowledge and Data Engineering*, **8**, 923–938.

King, R. K. and Harris, R. T. (1999). Parallel-coordinates visualization of capillary transport model analysis, In *Proceedings of BMES/EMBS Conference on Engineering in Medicine and Biology*, 1193.

Kumasaka, N. and Shibata, R. (2008). High-dimensional data visualization: The textile plot, *Computational Statistics & Data Analysis*, **52**, 3616–3644.

Kwak, I. Y. and Huh, M. H. (2008). Andrews' plots for extended uses, *The Korean Communications in Statistics*, **15**, 87–94.

Lee, H. and Ong, H. (1996). Visualization support for data mining, *IEEE Expert*, **11**, 69–75.

Lee, H., Ong, H., Toh, E. and Chan, S. (1995). A Multi-dimensional data visualization tool for knowledge discovery in databases, In *Proceedings of COMPSAC 95 Conference on computer Software and Applications*, 26–31.

Madhavan, P. G., Xu, B., Penna, M. A. and Low, W. C. (1991). Co-ordinate transformation in the hippocampal place cell phenomenon, In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, **3**, 1615–1620.

Miller, J. J. and Wegman, E. J. (1991). Construction of line densities for parallel coordinate plots, *Computing and Graphics in Statistics*, **36**, 107–123.

Moustafa, R. and Wegman, E. (2006). Multivariate continuous data-parallel coordinates, *Statistics and Computing, Graphics of Large Datasets: Visualizing a Million*, 143–156.

Siirtola, H. (2000). Direct manipulation of parallel coordinates, In *Proceedings of IEEE International Conference on Information Visualization 2000*, 373–378.

Teppola, P., Mujunen, S., Minkkinen, P., Puijola, T. and Pursiheimo, P. (1998). Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine's wet end, *Chemometrics and Intelligent Laboratory Systems*, **44**, 307–317.

Theisel, H. (2000). Higher order parallel coordinates, In *Proceedings of Vision, Modeling and Visualization 2000*, 119–125.

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

Unwin, A., Volinsky, C. and Winkler, S. (2003). Parallel coordinates for exploratory modelling analysis, *Computational Statistics and Data Analysis*, **43**, 553–564.

Weber, C. A. and Desai, A. (1996). Determination of paths to vendor market efficiency using parallel coordinates representation: A negotiation tool for buyers, *European Journal of Operational Research*, **90**, 142–155.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates, *Journal of the American Statistical Association*, **85**, 664–675.

Ye, H. and Lin, Z. (2006). Speed-up simulated annealing by parallel coordinates, *European Journal of Operational Research*, **173**, 59–71.

Regular Polyprism Parallel Coordinate Plot as a Statistical Graphics Tool

Dae-Heung Jang[1]

[1]Division of Mathematical Sciences, Pukyong National University

(Received April 2008; accepted May 2008)

Abstract

The parallel coordinate plot is a graphical data analysis technique for plotting multivariate data. The parallel coordinate plot overcomes the visualization problem of the Cartesian coordinate system for dimensions greater than 4. But, using different ordering of coordinate axes in the parallel coordinate plot of the same data may make different interpretations. Hence, we can use the regular polyprism parallel coordinate plot as an alternative for overcoming the variable arrangement problem of the parallel coordinate plot.

Keywords: Parallel coordinate plot, regular polyprism parallel coordinate plot, multi-regular polyprism parallel coordinate plot.

This work was supported by Pukyong National University Research Fund in 2006.

[1]Professor, Division of Mathematical Sciences, Pukyong National University, 599-1 Daeyeon-dong, Nam-gu, Busan 608-737, Korea. E-mail: dhjang@pknu.ac.kr