혼합물실험을 위한 평행좌표그림의 활용

장대홍[1)]

요약

혼합물실험에서 성분의 개수가 많고 성분에 대하여 제한조건이 있는 경우 흥미영역을 도시하는 작업은 성분의 개수가 4개 이상일 때 힘든 작업이 된다. 또한 반응변수가 여러 개인 경우 최적화문제도 까다롭게 된다. 이때, 탐색적 자료분석의 한 도구로서 평행좌표그림을 이용하면 통계자료분석시 많은 도움을 받을 수 있다.

주요용어: 혼합물실험, 제한영역, 최적화, 평행좌표그림

1. 서론

혼합물실험에서는 하나의 제품이 q개의 성분의 혼합으로 되어 있어서 각 성분의 혼합비가 문제가 된다. 그러므로, 혼합물실험에서는 다음과 같은 성분들에 대한 제약조건이 붙는다.

$$\text{모든 성분 } x_i(i=1,2,\ldots,q)\text{에 대하여 } 0\le x_i\le 1 \text{ 이고, } \sum_{i=1}^{q} x_i=1 \text{ 이다.} \quad (1)$$

또한, 우리는 종종 경제적인 이유나 실험 전 성분들에 대한 정보가 상당부분 있는 경우 등 여러 가지 이유로 위에서 정의된 심플렉스의 전 영역이 아닌 제한된 영역에서만 실험을 하게 된다. 우리는 이런 제한된 영역을 다음과 같이 나타낼 수 있다.

$$0\le a_i\le x_i\le b_i\le 1(i=1,2,\ldots,q)$$
$$0\le A_j\le \sum_{i=1}^{q} c_{ij}\,x_i\le B_j\le 1(j=1,2,\ldots,s) \quad (2)$$

혼합물실험을 위한 회귀모형으로서 우리는 상수항이 없고, 이차항으로서 순수이차항이 없고 혼합이차항만 있는 다음과 같은 이차회귀모형을 주로 사용한다.

$$E(y_u)=\sum_{i=1}^{q}\beta_i x_{iu}+\sum_{i<j}^{q}\beta_{ij}x_i x_j \quad (3)$$

혼합물실험에서는 (1)식을 만족하는 흥미영역 내에서 어떠한 성분이 반응변수에 유의한 영향을 미치는 지를 알고자 하고, 반응량을 최대 또는 최소로 만드는 최적혼합비율을 찾고자 한다. 위의 제한된 영역은 q가 4인 경우 3차원의 불규칙한 다면체가 된다. 그리고, q가 5 이상인 경우 직각좌표계로는 나타내기 어렵게 된다. 추정회귀식을 이용한 제한된 영역에서의 등고선도를 그리는 것도 불가능하게 된다. 더군다나, 반응변수가 2개 이상인 경우는 최적화가 다양한 모습을 띄게 된다. 이러한 때, 탐색적 자료분석의 한 도구로서 평행좌표그림을 이용하면 혼합물실험의 통계분석 시 많은 도움을 받을 수 있다. 평행좌표그림은 Inselberg(1985)가 구체적으로 제안한 이후 최근까지도 많은 학문영역에서 다양하게 쓰이고 있다.(Gennings외 3인(1990), Inselberg와 Dimsdale(1990), Wegman(1990), Madhavan외 3인(1991), Miller와 Wegman(1991), Lee외 3인(1995), Bateson과 Curtiss(1996), Keim과 Kriegel(1996), Lee와 Ong(1996), Weber와 Desai(1996), Becker(1997), Inselberg(1998, 2002), Ankerst외 2인(1998), Teppola외 4인(1998), Chou외 2인(1999), Fua외 2인(1999 a, b), Goel외 6인(1999), Groller외 2인(1999), King과

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 자연과학대학 수리과학부 통계학전공 교수

Harris(1999), Hall과 Berthold(2000), Siirtola(2000), Andrienko와 Andrienko(2001), Falkman(2001), Hauser외 2인(2002), Berthold와 Hall(2003)) 2003년도 'Computational Statistics and Data Analysis'라는 통계관련 저널의 vol. 43, no. 4호는 'Data Visualization'라는 제목으로 특집호로 꾸며졌는 데 여기에 평행좌표그림에 대한 여러 편의 논문들이 수록되어 있다.

2. 평행좌표그림의 활용

혼합물실험에서 성분의 개수가 많고 성분에 대하여 제한조건이 있는 경우 흥미영역을 도시하고 반응변수들의 동시최적화문제를 풀기 위하여 평행좌표그림을 이용할 수 있다. 실험계획에 의한 실험이 완성된 후 얻어지는 데이터에 대하여 평행좌표그림을 그림으로써 제한영역의 모습도 확인하여 보고 반응변수들 사이의 관계도 알아볼 수 있다. 또한, 추정회귀식의 회귀계수에 대한 평가도 평행좌표그림을 통하여 한 눈에 알아볼 수 있다. 잔차를 통한 회귀진단도 평행좌표그림을 통하여 다양한 측도들을 비교하여 볼 수 있다. 반응변수들에 대한 추정회귀식을 이용하여 제한조건 (2)를 만족하는 적당한 크기의 $(x_1, x_2, \ldots, x_q, \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_r)$의 조합들을 시뮬레이션을 통하여 구하면 평행좌표그림을 구할 수 있고, 이 평행좌표그림을 통하여 경험적으로 최적화문제를 풀 수 있고 반응변수에 대한 조건부분포를 브러싱기법을 이용하여 확인하여 볼 수 있다.

앞으로의 논의를 위하여 Martin, Platts, Seddon, 그리고 Stillman(2003)의 논문에 나오는 자료를 이용하기로 한다. 이 혼합물실험은 세라믹이나 금속제품의 표면을 보호하거나 장식하기 위하여 입히는 유리제품에서 납성분을 대신하여 8개의 산화물들 (x_1:SiO_2, x_2: Al_2O_3, x_3:ZrO_2, x_4:TiO_2, x_5: B_2O_3, x_6: Na_2O+K_2O(1:1), x_7:CaO, x_8:ZnO)을 첨가하는 실험으로서 반응변수는 3개이고, 각각 중성용액(증료수), 알칼리성용액(NaOH), 산성용액(HCl)에서의 젖빛유리의 중량감소량의 비율 (y_1, y_2, y_3)이다. 그러므로, 반응변수들의 값을 최소화시키는 최적혼합비율을 찾는 것이 중요하다. 8개 성분들에 대한 제한된 조건들은 다음과 같다.

$$\begin{gathered} 0.40 \le x_1 \le 0.53, 0 \le x_2 \le 0.04, 0 \le x_3 \le 0.04, 0 \le x_4 \le 0.04, \\ 0.17 \le x_5 \le 0.30, 0.10 \le x_6 \le 0.20, 0.02 \le x_7 \le 0.15, 0 \le x_8 \le 0.10, \\ 0.45 \le x_1 + x_2 + x_3 + x_4 \le 0.53,\ x_2 + x_3 + x_4 \le 0.08, \\ 0.37 \le x_5 + x_6 \le 0.45,\ x_7 + x_8 \le 0.15 \end{gathered} \tag{4}$$

4개의 batch로 나누어 실시된 63개의 실험계획점들 중 9개는 제한영역의 중심점이고, 54개는 GOSSET software로 구한 실험점 90개 짜리 D-최적화실험계획에서 임의로 선택된 실험점들이다. 63개의 실험을 평행좌표그림으로 나타내면 다음 그림 1과 같다. 평행좌표그림에서는 각 데이터가 하나의 꺾은 선으로 나타내어진다.

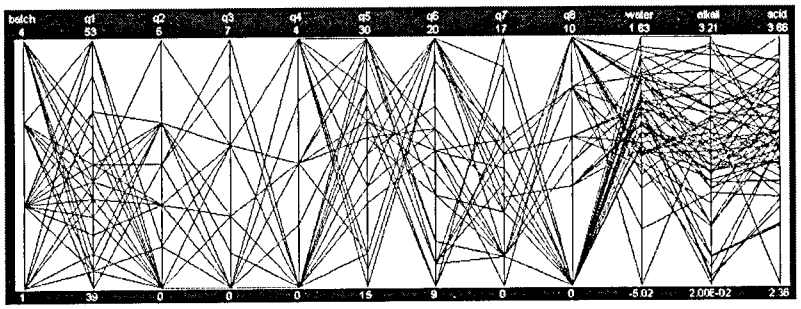

그림 1. 63개 실험데이터

브러싱기법을 이용하면 제한영역의 중심점을 나타낼 수 있다. 8개의 성분들에 대하여 표준화를 시행하면 다음 그림 2와 같이 나타난다. 우리는 (4)식의 제한조건을 만족하는 제한영역을 그림

2에서 볼 수 있다. 특이한 모양을 갖고 있음을 알 수 있다. 그림 2에서 y_1(중성)의 최소값과 y_3(산성)의 최소값은 일치하고 이 때 y_2(알칼리성)의 값은 큰 값을 갖는 반면, y_2 (알칼리성)의 최소값에 대하여 y_1(중성)과 y_3(산성)의 값은 작은 값이 아님을 알 수 있다. 즉, 최적화(최소화) 입장에서는 y_1(중성)의 값과 y_3(산성)의 값이 비례 관계에 있고, y_2 (알칼리성)의 값은 y_1(중성)과 y_3(산성)의 값과 반비례 관계에 있음을 알 수 있다.

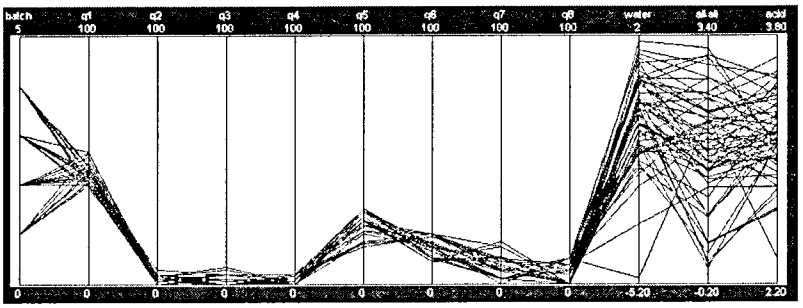

그림 2. 표준화 후의 63개 실험데이터

이 63개의 실험데이터를 이용하여 통계분석을 하여 보면 batch라는 블록효과는 없음을 알 수 있다. 성분의 개수가 8개이기 때문에 혼합이차항의 개수는 무려 28개가 된다. 회귀식으로 8개의 일차항은 일단 포함시키고 혼합이차항의 포함 여부를 따지기 위하여 각 반응변수에 대하여 변수선택방법으로서 단계적 방법(stepwise method)을 사용하여 변수를 선택하여 보면 다음 그림 3과 같이 평행좌표그림으로 나타낼 수 있다. 여기서 크기는 분산분석표 상의 F값을 나타낸다. 유의한 혼합이차항이 반응변수 별로 많이 다름을 알 수 있다.

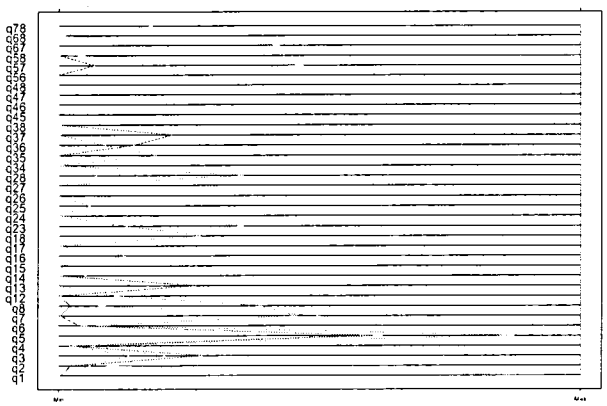

그림 3. 회귀계수에 대한 F값을 나타내는 평행좌표그림

특이값과 영향이 큰 관측값을 찾기 위한 회귀진단용 여러측도들을 평행좌표그림을 이용하면 한 눈에 이러한 측도들을 비교하여 볼 수 있다. 그림 4는 반응변수 y_1(중성)에 대하여 회귀진

단용 여러측도들을 나타낸 평행좌표그림이다. 다른 반응변수에 대해서도 이러한 평행좌표그림을 그려 특이값과 영향이 큰 관측값을 찾아낼 수 있다.

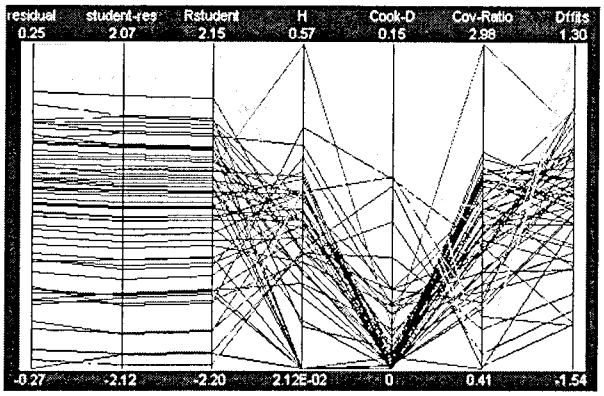

그림 4. 반응변수 y_1(중성)에 대한 회귀진단용 여러 측도들을 나타내는 평행좌표그림

추정된 회귀식들을이용하여 제한조건 (4)를 만족하는 285개의 (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$)의 조합들을 시뮬레이션을 통하여 구하면 다음 그림 5와 같은 평행좌표그림을 구할 수 있다.

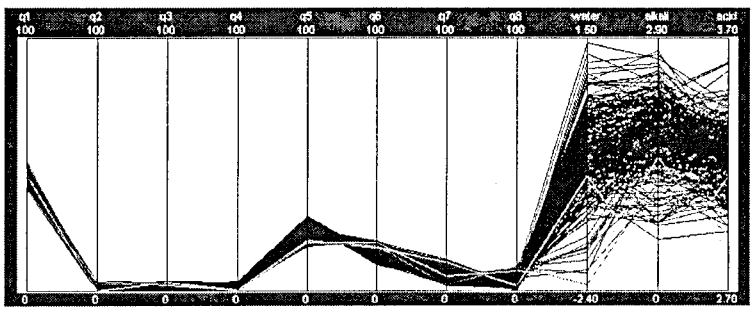

그림 5. 285개의 시뮬레이션 데이터

이 그림에서 (4)식의 제한조건을 만족하는 제한영역을 볼 수 있다. 최적화(최소화) 입장에서는 y_1(중성)의 값과 y_3(산성)의 값이 비례 관계에 있고, y_2 (알칼리성)의 값은 y_1(중성)과 y_3(산성)의 값과 반비례 관계에 있음을 알 수 있다. 반응변수 y_1(중성)에 대한 조건부분포를 브러싱기법을 이용하여 구하여 보면 그림 6과 같다. 반응변수 y_1(중성)에 대하여 내림차순으로 7개의 범위로 나누어 브러싱을 이용하여 각 범위에 해당되는 점들을 하이라이트시키면 반응변수 y_1(중성)를 조건부로 하는 데이터들의 조건부분포를 파악할 수 있다.

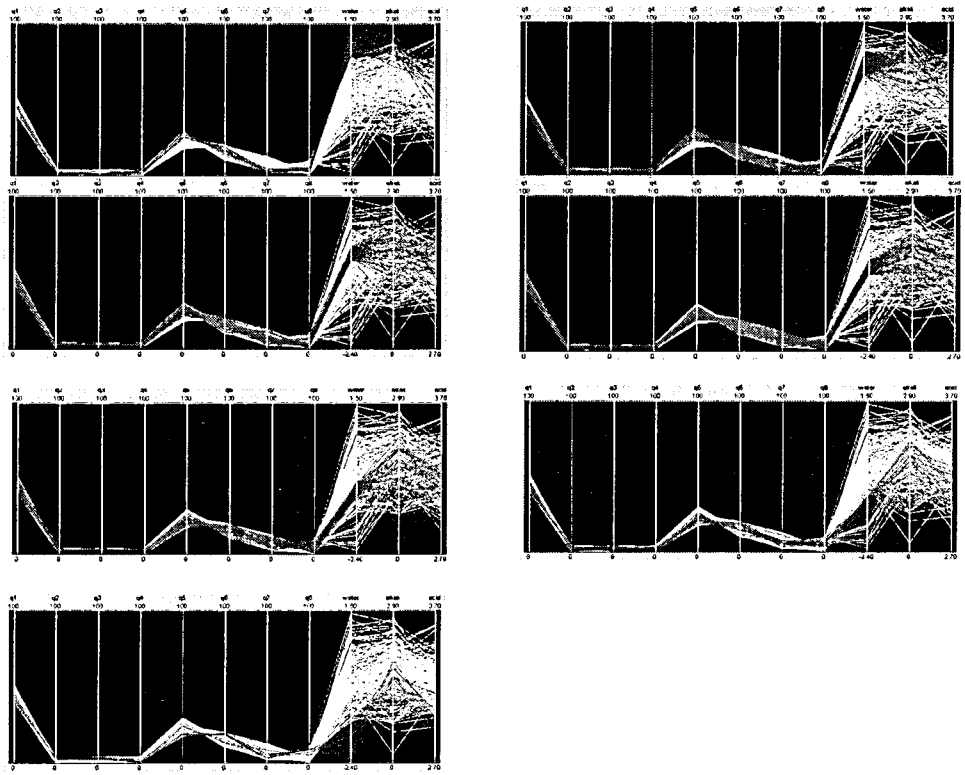

그림 6. 반응변수 y_1(중성)에 대한 조건부분포

3. 결론

혼합물실험에서 성분의 개수가 많고 성분에 대하여 제한조건이 있는 경우 흥미영역을 도시하고 반응변수들의 동시최적화문제를 풀기 위하여 평행좌표그림을 이용할 수 있다. 또한, 평행좌표그림을 통하여 회귀계수들의 유의성이나 회귀진단을 한 눈에 확인하여 볼 수 있다. 이런 작업들을 통하여 탐색적 자료분석의 한 도구로서 평행좌표그림을 이용하면 통계자료분석시 많은 도움을 받을 수 있다.

참고문헌

[1] Andrienko, G. and Andrienko, N. (2001), Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates, *Computers, Environment and Urban Systems*, 25, 5-15.

[2] Ankerst, M., Berchtold, S. and Keim, D. (1998), Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data, *Proceedings of IEEE Symposium on Information Visualization*, 52-60.

[3] Bateson, A. and Curtiss, B. (1996), A Method for Manual Endmember Selection and Special Unmixing, *Remote Sensing of Environment*, 55, 229-243.

[4] Becker, O. M. (1997), Representing Protein and Peptide Structures with Parallel-Coordinates, *Journal of Computational Chemistry*, 18, 1893-1902.

[5] Berthold, M. R. and Hall, L. O. (2003), Visualizing Fuzzy Points in Parallel Coordinates, *IEEE Transactions on Fuzzy Systems*, 11, 369-374.

[6] Chou, S., Lin, S. and Yeh, C. (1999), Cluster Identification with Parallel Coordinates, *Pattern Recognition Letters*, 20, 565-572.

[7] Falkman, G. (2001), Information Visualisation in Clinical Odontology: Multidimentional Analysis and Interactive Data Exploration, *Artificial Intelligence in Medicine*, 22, 133-158.

[8] Fua, Y., Ward, M. O. and Rundensteiner, E. A. (1999 a), Navigating Hierarchies with Structure-based Brushes, *Proceedings of IEEE Symposium on Information Visualization*, 58-64.

[9] _____(1999 b), Hierarchical Parallel Coordinates for Exploration of Large Datasets, *Proceedings of Visualization '99*, 43-50.

[10] Gennings, C., Dawson, K. S., Carter, W. H. and Myers, R. H. (1990), Interpreting Plots of Multidimensional Dose-Response Surfaces in a Parallel Coordinate System, *Biometrics*, 46, 719-735.

[11] Goel, A., Baker, C., Shaffer, C. A., Grossman, B., Haftka, R. T., Mason, W. H. and Watson, L. T. (1999), VizCraft: A Multidimensional Visualization Tool for Aircraft Configuration Design, *Proceedings of Visualization '99*, 425-428.

[12] Groller, E., Loffelmann, H. and Wegenkittl, R. (1999), Visualizations of Dynamical Systems, *Future Generation Computer Systems*, 15, 75-86.

[13] Hall, L. O. and Berthold, M. R. (2000), Fuzzy Parallel Coordinates, *Proceedings of 19th International Conference of Fuzzy Information Processing Society*, 74-78.

[14] Hauser, H., Ledermann, F. and Doleisch, H. (2002), Angular Brushing of Extended Parallel Coordinates, *Proceedings of the IEEE Symposium on Information Visualization*, 127-130.

[15] Inselberg, A. (1985), The Plane with Parallel Coordinates, *The Visual Computer*, 1, 69-91.

[16] _____(1998), Visual Data Mining with Parallel Coordinates, *Computational Statistics*, 13, 47-63.

[17] _____(2002), Visualization and Data Mining of High-dimensional Data, *Chemometrics and Intelligent Laboratory Systems*, 60, 147-159.

[18] Inselberg, a. and Dimsdale, B. (1990), Parallel Coordinate: A Tool for Visualizing Multi-dimensional Geometry, *Proceedings of Visualization '90*, 361-378.

[19] Keim, D. A. and Kriegel, H. (1996), Visualization Techniques for Mining Large Databases: A Comparison, *IEEE Transactions on Knowledge and Data Engineering*, 8,923-938.

[20] King, K. and Harris, T. (1999), Parallel-coordinates Visualization of Capillary Transport Model Analysis, *Proceedings of BMES/EMBS Conference on Engineering in Medicine and Biology*, 1193.

[21] Lee, H. and Ong, H. (1996), Visualization Support for Data Mining, *IEEE Expert*, 11, 69-75.

[22] Lee, H., Ong, H., Toh, E. and Chan, S. (1995), A Multi-Dimensional Data Visualization Tool for Knowledge Discovery in Databases, *Proceedings of COMPSAC 95 Conference on computer Software and Applications*, 26-31.

[23] Madhavan, P. G., Xu, B., Penna, M. A. and Low, W. C. (1991), Co-ordinate Transformation in the Hippocampal Place Cell Phenomenon, *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, 1615-1620.

[24] Miller, J. J. and Wegman, E. J. (1991), Construction of Line Densities for Parallel Coordinate Plots, in *Computing and Graphics in Statistics* edited by A. Buja and P. A. Tukey. Springer-Verlag, New York, NY.

[25] Siirtola, H. (2000), Direct Manipulation of Parallel Coordinates, *Proceedings of IEEE International Conference on Information Visualization*, 373-378.

[26] Teppola, P., Mujunen, S., Minkkinen, P., Puijola, T. and Pursiheimo, P. (1998), Principal Component Analysis, Contribution Plots and Feature Weights in The Monitoring of Process Data From A Paper Machine's Wet End, *Chemometrics and Intelligent Laboratory Systems*, 44, 307-317.

[27] 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.

[28] Wegman, E. J. (1990), Hyperdimensional Data Analysis using Parallel Coordinates, *Journal of the American Statistical Association*, 85, 664-675.