Graphical Editing for Multivariate Continuous Data

박 진 우, 수원대학교 통계정보학과
이 은 경, 서울대학교 통계학과

2005. 12. 9
조사연구학회 추계 학술발표대회

< 목 차 >

1. Introduction

* What? (Granquist, 1995; Kovar & Granquist, 1997)

  - the procedure for detecting and adjusting individual errors in data records resulting from data collection and capture

  - tidy up the data

* Fellegi and Holt (1976) : F-H system

1. Introduction

* Problem

  - Graphical Editing for Multivariate Continuous Data

* Example : Size Korea 2004 (기술표준원) 3차원 측정 데이터

  표본 : 전국 성인남녀 4934명
  측정 : 3차원 스캐너 활용 측정 후 신체 128 부위 측정값 자동 계산

  스캐너 및 자동계산 소프트웨어의 오류 가능성
  서로 깊은 연관관계를 지니는 연속형 다변량 변수

1. Introduction

* Graphical editing (Granquist, 1995)

  - convenient, user-friendly, parameter free
  - easy to understand, flexible

  - univariate : histogram, box plot
  - bivariate : scatter plot
  - multivariate : scatter plot matrix ?

* Study Objective

  - graphical tool for multivariate continuous data editing
  - GGobi (Swayne, 2003)

2. Traditional Approaches of Graphical Editing

2.1 Editing with univariate graph

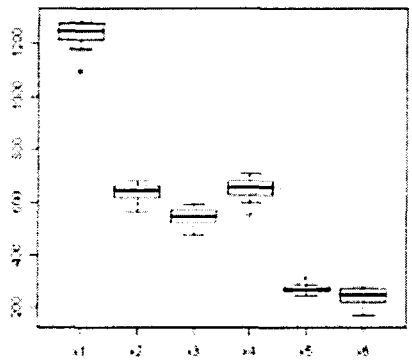
<Example> 8 year-old girls data

id	키	가슴둘레	허리둘레	엉덩이둘레	팔둘레	몸무게
1	1268	660	575	676	309	26.7
2	1285	665	557	694	280	28.0
3	1091	563	475	550	244	16.7
4	1222	648	570	678	264	25.8
5	1213	584	497	630	264	21.7
6	1176	592	518	601	245	20.3
7	1210	659	592	693	285	27.3
8	1279	623	515	648	255	25.4
9	1263	625	532	679	266	27.5
10	1284	664	565	691	280	27.1
11	1244	636	532	630	267	23.4
12	1216	606	490	597	264	21.4
13	1222	630	580	653	274	24.5
14	1235	610	564	658	268	24.1
15	1179	670	542	625	252	22.2
16	1268	678	549	672	280	27.4
17	1274	648	526	647	268	24.6
18	1279	644	537	648	265	25.0
19	1253	644	567	673	274	25.9
20	1283	664	586	705	288	18.4

2. Traditional Approaches of Graphical Editing

2.1 Editing with univariate graph

* 각 변수에 독립적으로 box plot을 적용 : 3개의 의심점 발견

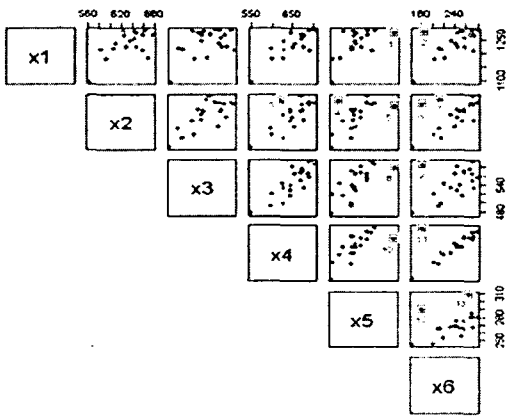


- *문제점 : 변수들 간의 연관성을 무시*

2. Traditional Approaches of Graphical Editing

2.2 Editing with scatterplot matrix

* 2차원 그래프를 사용하는 경우 ⇒ *5개 의심점*

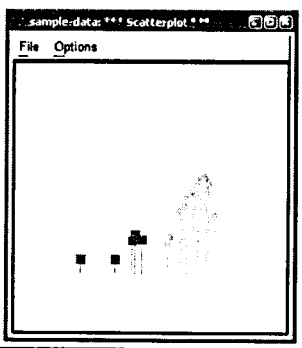


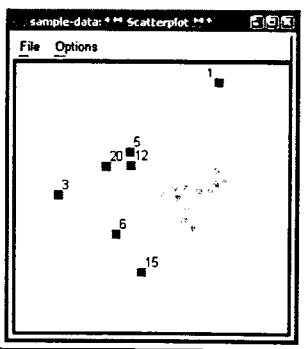
2. Traditional Approaches of Graphical Editing

- 2차원 그래프를 사용하는 경우의 문제점
  - *변수의 수가 적은 경우에만 작업이 가능*
  - *변수가 많고, 데이터 수가 많을 때 작업이 용이하지 않음*

3. Editing with Dynamic Graphical Methods

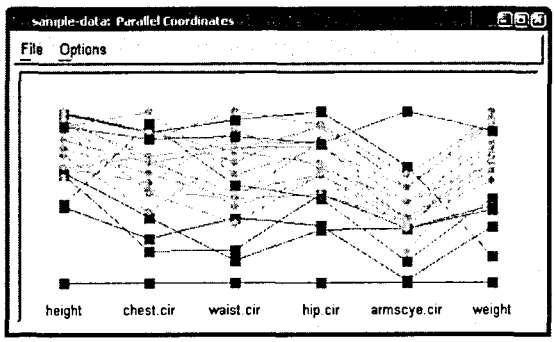
- GGobi : interactive and dynamic graphical system, freeware
- Projection Pursuit Guided Tour
  - 다차원의 점을 1차원 또는 2차원의 좌표에 투영 : 1D tour, 2D tour
  - Central Mass





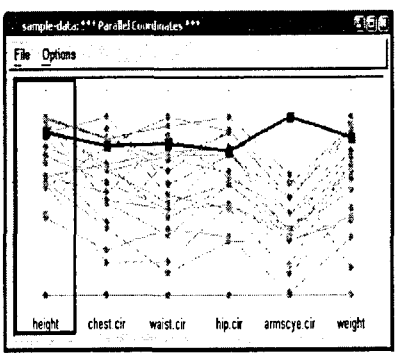
3. Editing with Dynamic Graphical Methods

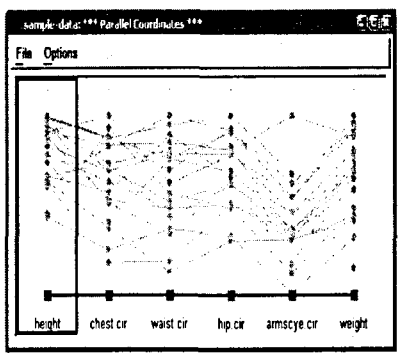
- Parallel Coordinate Plot
  - 이상점으로 판명된 ID에서 어떤 변수에 문제가 있는지 판단하는데 도움



3. Editing with Dynamic Graphical Methods

- Parallel Coordinate Plot





4. Application

- 2004 Size Korea : 3차원 측정 데이터

  표본 : 전국 성인남녀 4934명

  측정 : 3차원 스캐너 활용 측정 후 신체 128 부위 측정값 자동 계산

- Data Editing 전략

  1단계 : Logical Editing ⇒ 데이터 확인 및 수정

  2단계 : Editing with Dynamic Graphs ⇒ 데이터 확인 및 수정

4. Application

- Logical Editing

  - 총 54가지의 편집 규칙 마련

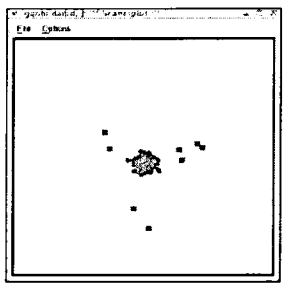
  (예) {눈높이 > 키 } ⇒ 오류 !

  - 299개(6.06%)의 논리적 오류 발견

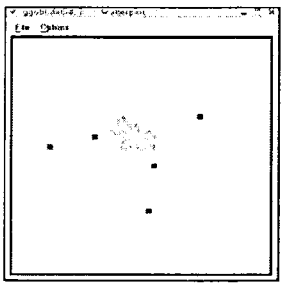
  - 점검 결과: 스캐닝 이미지에서 표식(landmark) 손상

4. Application

- Editing with Dynamic Graphs
  - 15개 중요 관심변수 선정
  - 성별, 연령그룹별로 2D optimal projection,
  - Parallel Coordinate Plot



남자 **8-10**세



여자 **8-10**세

4. Application

남 자			여 자		
연령그룹	표본수	오류수 (%)	연령그룹	표본수	오류수 (%)
8-10	148	8 (5.41%)	8-10	142	5 (3.52%)
11-13	197	4 (2.03%)	11-13	204	5 (2.45%)
14-16	209	6 (2.87%)	14-16	209	4 (1.91%)
17-19	200	12 (6.0%)	17-19	202	0 (0.00%)
20-29	517	8 (1.55%)	20-29	518	6 (1.16%)
30-39	512	0 (0.00%)	30-39	523	3 (0.57%)
40-49	254	3 (1.18%)	40-49	265	10 (3.77%)
50-59	251	5 (1.99%)	50-59	232	4 (1.72%)
60-69	172	2 (1.16%)	60-69	179	5 (2.79%)
Total	2460	48 (1.95%)	Total	2474	42 (1.70%)

5. Conclusion

- 다변량 연속형 데이터베이스
  - 데이터 품질 보장을 위한 Data Editing은 필수적

- 다변량 연속형 데이터 편집
  - Classical Graphical Tool은 한계가 있음
  - Outlier Detection을 위한 통계량 : 어려움, 통계적 분포 가정

  - Dynamic Graphical Tools를 Data Editing에 적용 !
    : 매우 효과적인 도구로 활용 가능함.

- ISO/DIS 20685, 3D scanning methodologies for internationally compatible anthropometric databases, 2003.
- Granquist, L., Kovar, J. G., 1997. Editing of Survey Data: How Much Is Enough?, In: Survey Measurement and Process Quality, (eds. Lyberg, et al.), 415-435, John Wiley & Sons.
- Granquist, L., 1995. Improving the Traditional Editing Process, In: Business Survey Methods (eds. Cox et al.), 177-199, John Wiley & Sons.
- Unwin, A., 2000. Using your eyes - making statistics more visible with computers, CSDA 32, 303-312.
- Ghosh, B. D., Schafer, J. L., 2003. Multiple Edit/Multiple Imputation for Multivariate Continuous Data, JASA 98(464), 807-817.
- Fellegi, I. P, Holt, C. 1976. A Systematic Approach to Automatic Edit and Imputation, JASA , 17-35.
- Swayne, D. F., Lang, D. T., Buja, A., and Cook, D. 2003. GGobi: evolving from XGobi into an extensible framework for interactive data visualization, CSDA 43, 423-444.
- Becker, C., and Gather, U., 2001. The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. CSDA 36, 119-127.
- Jones, P. R. M., and Rioux, M. 1997. Three-dimensional Surface Anthropometry: Applications to the Human Body, Optics and Lasers in Engineering 28, 89-117.