Visualization Techniques for the Integration of Rank Data

Michael G. Schimek[1] Eva Budinská[2]

[1]Medical University of Graz, Graz, Austria

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

COMPSTAT 2010, Paris, France, August 22-27, 2010

Motivation

- In various fields of application we are confronted with lists of distinct objects in rank order
- The ordering might be due to a measure of strength of evidence or to an assessment based on expert knowledge or a technical device
- The ranking might also represent some measurement taken on the objects which might not be comparable across the lists, for instance, because of different assessment technologies or levels of measurement error

Our aim is

- to **consolidate such lists of common objects**
- to **provide computationally tractable solutions**, hence appropriate algorithms and graphs

Medizinische Universität Graz

General assumptions

- Let us assume ℓ assessors or laboratories ($j = 1, 2, \ldots, \ell$) assigning rank positions to the same set of N distinct objects
- Assessment of N distinct objects according to the extent to which a particular attribute is present
- All assessors, independently of each other, rank the same objects between 1 and *N* on the basis of relative performance
- The ranking is from 1 to *N*, without ties
- Missing assessments are allowed
- The ℓ assessors produce ℓ rank lists τ_j
- There are $(\ell^2 - \ell)/2$ possible pairs of such lists τ_j

Medizinische Universität Graz

The problem

- In most applications, especially for large or huge numbers N of objects, it is unlikely that consensus prevails
- As result only the top-ranked objects matter (the remainder ones show random ordering)
- Quite often we observe a general decrease, not necessarily monotone, of the probability for consensus rankings with increasing distance from the top rank position

Typically there is reasonable **conformity in the rankings for the first, say k, elements of the lists**: **notion of *top-k rank lists***

Tasks: **Consensus in preference and voting, integration of search engine results, meta-analysis of microarray experiments**

Medizinische Universität Graz

A motivating example: U.S. college preference data

- Avery et al. (2005) developed a statistical model which allows the construction of a ranking of U.S. undergraduate programs based on students' revealed preferences
- Data from 1357 high achieving students (90th percentile of all SAT takers) seeking admission
- $N = 110$ colleges and universities taking part in the national ranking (matriculation tournaments)
- For each college/university there are two rankings of interest: **matriculation rank** (MR) and **preference rank** (PR)
- There are no missing assignments
- **Question: Is there a top list of conforming rank assignments?**

Medizinische Universität Graz

A motivating example: U.S. college preference data

College Name	MR	PR
Harvard University (o_1)	1	1
California Inst. of Technology (o_2)	2	7
Yale University (o_3)	3	5
Massachusetts Inst. of Technology (o_4)	4	3
Stanford University (o_5)	5	2
Princeton University (o_6)	6	4
Brown University (o_7)	7	6
Columbia University (o_8)	8	8
Amherst College (o_9)	9	13
Dartmouth College (o_{10})	10	11
Wellesley College (o_{11})	11	33
University of Pennsylvania (o_{12})	12	12
University of Notre Dame (o_{13})	13	14
Swarthmore College (o_{14})	14	10
Cornell University (o_{15})	15	15
Georgetown University (o_{16})	16	9
⋮	⋮	⋮

Medizinische Universität Graz

The data stream input

- The **indicator variable** takes $I_j = 1$ if the ranking given by the second assessor to the object ranked j by the first is not distant more than δ from j, and $I_j = 0$ otherwise
  $\Rightarrow$ **data stream**
- **Concordance** is assumed for an arbitrary object o when its rank in τ_i is maximal δ index positions apart from its rank in τ_j
- The data stream **depends on the distance parameter** δ
- δ is defined by the shift in index positions of a particular object o in one list, say τ_i, with respect to the other list, say τ_j
- A sequence of data streams ordered according to δ represents the **reduction of discordance**

Medizinische Universität Graz

U.S. college data: data streams for $\delta = 0$ to 5

Object	*MR*	*PR*	$\delta = 0$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$	$\delta = 5$
o_1	1	1	1	1	1	1	1	1
o_2	2	7	0	0	0	0	0	1
o_3	3	5	0	0	1	1	1	1
o_1	4	3	0	1	1	1	1	1
o_1	5	2	0	0	0	1	1	1
o_1	6	4	0	0	1	1	1	1
o_1	7	6	0	1	1	1	1	1
o_1	8	8	1	1	1	1	1	1
o_1	9	13	0	0	0	0	1	1
o_{10}	10	11	0	1	1	1	1	1
o_{11}	11	33	0	0	0	0	0	0
o_{12}	12	12	1	1	1	1	1	1
o_{13}	13	14	0	1	1	1	1	1
o_{14}	14	10	0	0	0	0	1	1
o_{15}	15	15	1	1	1	1	1	1
o_{16}	16	9	0	0	0	0	0	0
#(zeros)			12	8	6	5	3	2

Medizinische Universität Graz

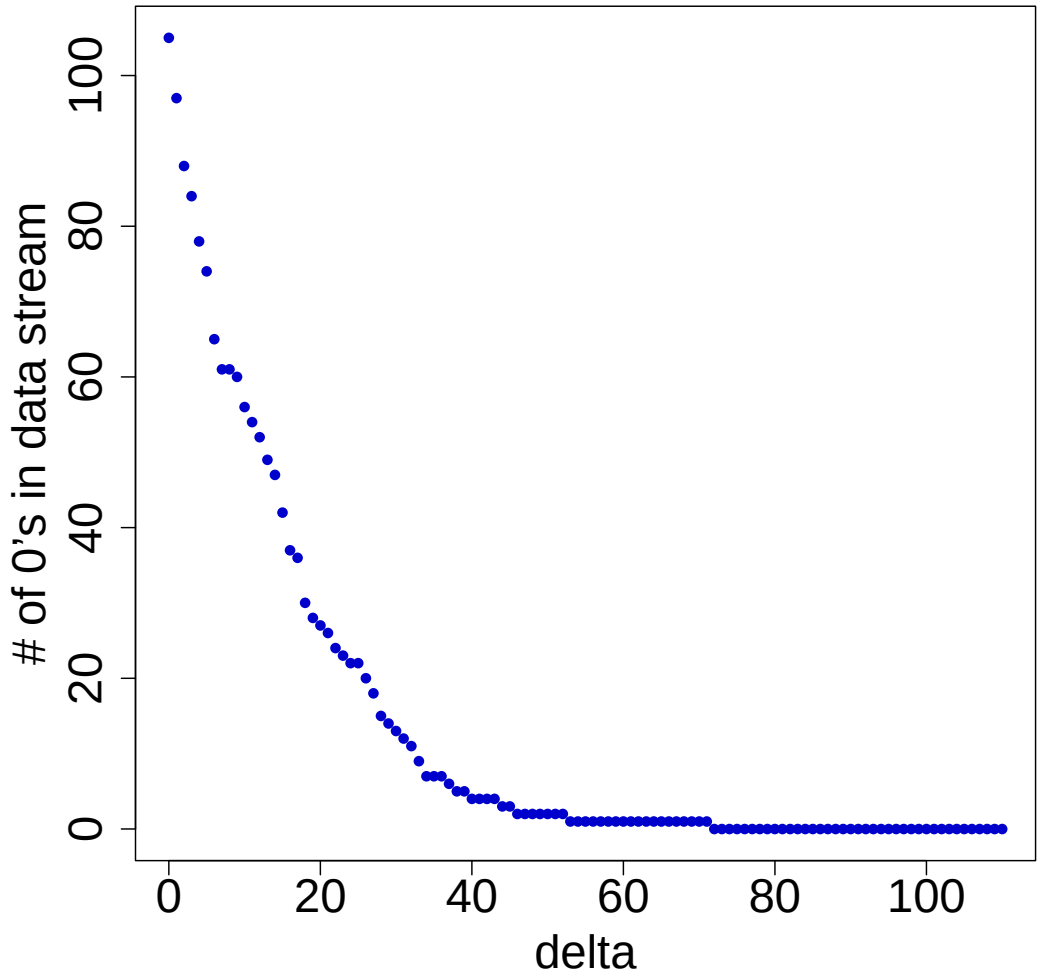
Selection of $\hat{k}$ for list truncation

Moderate deviation-based inference for random degeneration in paired rank lists (Hall and Schimek, 2008, 2010)

- For the estimation of the point of degeneration j_0 into noise independent **Bernoulli random variables** are assumed
- A general **decrease of the probability** p_j (need not be monotone) **for concordance of rankings** with increasing distance j from the top rank is assumed
- A **distance parameter** δ and a **tuning parameter** ν are required to account for the **closeness of the assessors' rankings and the degree of randomness in the assignments**
- The **algorithm** represents a simplified mathematical model;
- It is embedded in an **iterative scheme** to account for irregular rankings

Medizinische Universität Graz

Δ-plot for matriculation rank and preference rank of U.S. colleges



Medizinische Universität Graz

U.S. college preference data: inference results

- δ choice based on Δ-plot
- Sharp decline of $\#(zeros)$', especially for δ's up to about 20 (around $\delta = 45$ almost no discordance left)
- Pilot sample size $\nu \geq 4$ (functions as smoothing parameter)
- For $\delta = 10$ and $\nu = 4$ we obtain the smallest of all stable results: $\hat{j}_0 = 16$ (**15 top ranking colleges**)
- For $\delta = 20$ and $\nu = 28$ we obtain $\hat{j}_0 = 71$ (**70 top ranking colleges**)
- **Both results make sense** and depend on the goal of the study (more than one result because of modest separability)

An application: meta-analysis of microarray data

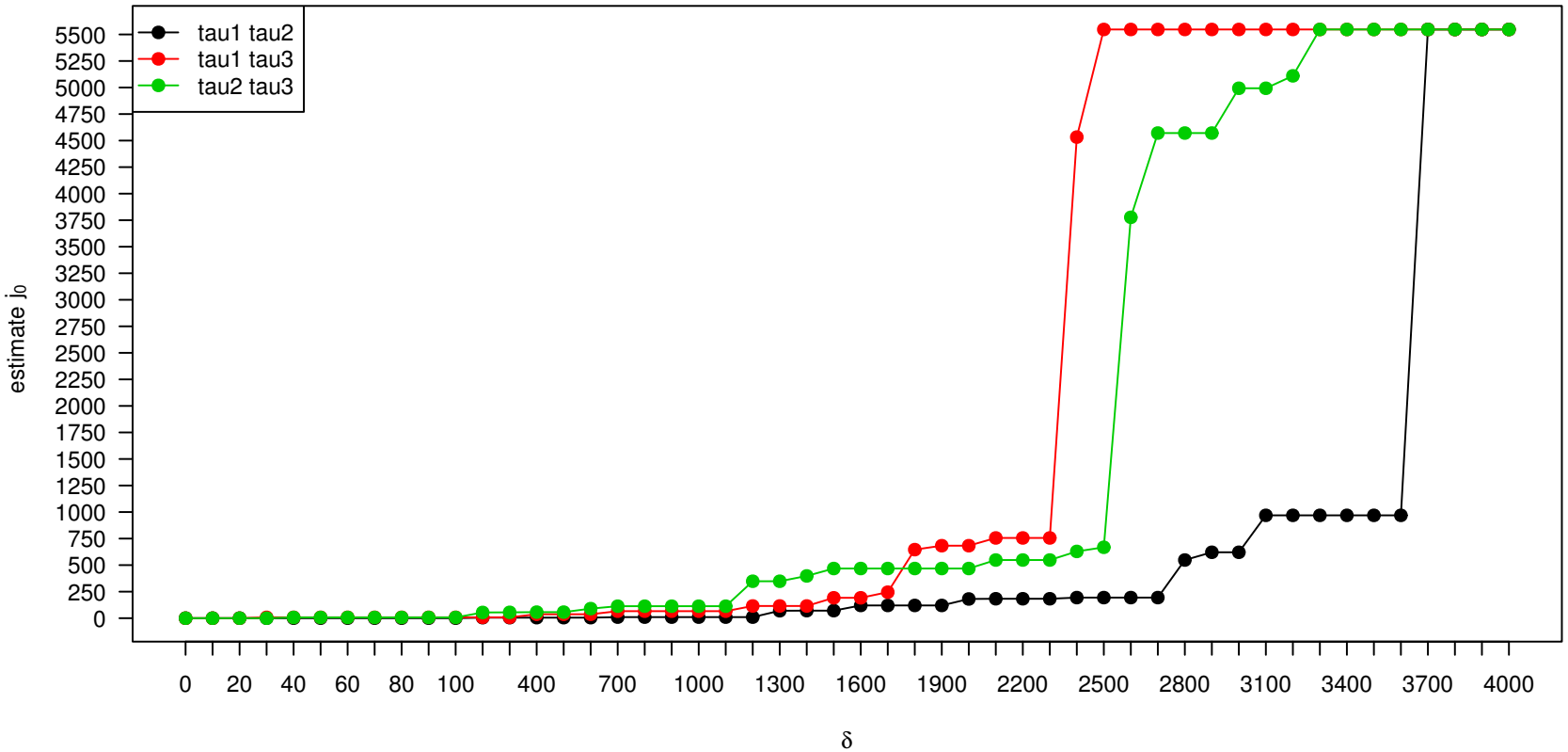
Breast cancer data due to Sørlie et al. (2003)

- Study goal: Identification of breast tumor subtypes from gene expression measured by microarrays
- Here we consider selected expression data from three independent patient cohorts called *Norway*, *Norway FU*, and *Stanford*, hybridized on different platforms
- Only genes (unique gene symbols) common to all platforms are analyzed
- 3 ranked lists, τ_1, τ_2, and τ_3, each of length $N = 5812$

Our task:
Identification of a subset of genes supported by all 3 cohorts that can be used for further unsupervised analysis of subtypes of breast cancer

Medizinische Universität Graz

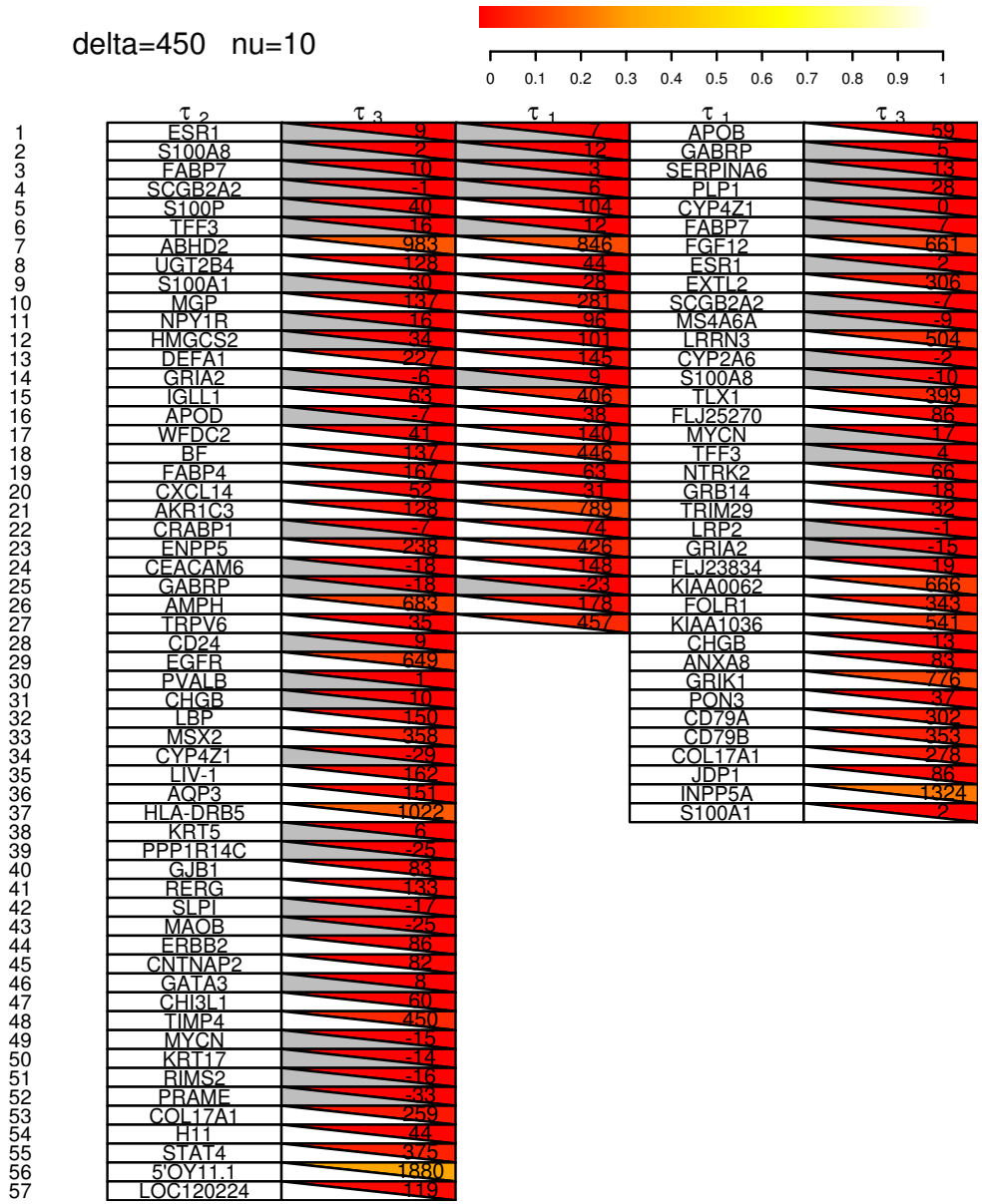
Estimates of j_0 for a range of δ values, combining pairwise the lists τ_1, τ_2, and τ_3 ($r = 1.2$, $C = 0.4$)



Aggregation map: Graphical integration of paired ranked lists

- Define a partial **reference list** τ_1^0; anyone of the 2 lists with $\max_j(\hat{k}_j)$ objects among all pairwise comparisons (τ_1^0 gives the ordering of the objects o_i on the vertical axis of the plot)
- The **partial lists** $\tau_2, \tau_3, \ldots, \tau_\ell$ are ordered from highest to lowest by their individual k_j when compared to the reference list τ_1^0 (one column per list)
- In each cell we represent: (1) **top-k membership**, '*yes*' is denoted by color 'grey' and '*no*' by 'white',
(2) **distance** of a current object $o_i \in \tau_1^0$ from its position in the other list, color scale from 'red' *identical* to 'yellow' *far distant* (integer value denotes distance with negative sign if to the left, and positive sign if to the right)
- **Implemented in R utilizing the grid add-on package of Murrell (2006)**

Aggregation map for $\delta = 450$, combining τ_1, τ_2, and τ_3



Summary and conclusions

- **Irregularities**, typical for empirical ranked lists, can be **well represented by means of data streams**
- Data streams are distance-dependent: distance can be evaluated via the **Δ-plot**
- **Data stream input is sufficient for** (1) **inference** on the degradation of information and for (2) the **graphical integration** of top-ranked objects
- The **aggregation map**, a new graphical tool, provides additional insight into a top-k set of objects
- The approach is **computationally tractable and efficient**
- The procedures will soon be available in the **R-package TopKLists**
- **The approach has already demonstrated its practical value**

Medizinische Universität Graz