A test procedure for random degeneration of paired rank lists

Michael G. Schimek[1,3] Peter Hall[2] Eva Budinská[3]

[1]Medical University of Graz
Institute for Medical Informatics, Statistics and Documentation
8036 Graz, Austria, Europe

[2]The University of Melbourne, Australia

[3]Masaryk University, Czech Republic, Europe

MCP 2007
9–11 July 2007, Vienna, Austria

Introduction

The statistical problem

- Assume two assessors (e.g. laboratories, search engines)
- The first assessor ranks N distinct objects according to the extent to which a particular attribute is present
- The ranking is from 1 to N, without ties

There are two different situations of interest:

1. The second assessor assigns each object to the one or the other of two categories (0-1-decision)
2. The second assessor also ranks the objects from 1 to N

An **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 m, say, from j,
and $I_j = 0$ otherwise
The I_j's form a **data stream** (input of our algorithm)

Medizinische Universität Graz

Introduction continued

The goal

- In both situations we wish to determine how far into the two rankings one can go before the **differences between them degenerate into noise**
- This allows us to identify a **sequence of objects** that is characterized by a **high degree of assignment conformity**

Typical applications are:

- Data integration from various 'omic' platforms in molecular research (e.g. microarrays, SNP arrays, microRNA arrays)
- Construction of a meta-search engine for Web applications
- Top-k-list comparisons from data logs on the Web across time (e.g. automatic screening for emerging trends)

Medizinische Universität Graz

Examples of data streams

EXAMPLE 1: DATA STREAM WITH PERFECT OVERLAP IN THE TOP-SET (N=100 **OBJECTS**)

Assignments	Objects																					
Ranking 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	- - - -	100
Ranking 2	1	2	3	4	5	6	11	9	8	10	18	12	28	17	90	13	21	23	19	14	- - - -	87
Indicator	1	1	1	1	1	1	0*	0	0	1	0	1	0	0	0	0	0	0	1	0	- - - -	0
* Point of degeneration																						

EXAMPLE 2: DATA STREAM WITHOUT PERFECT OVERLAP IN THE TOP-SET (N=100 **OBJECTS**)

Assignments	Objects																					
Ranking 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	- - - -	100
Ranking 2	2	1	3	5	4	6	11	9	8	10	18	12	28	17	90	13	21	23	19	14	- - - -	87
Indicator	0	0	1	0	0	1	0*	0	0	1	0	1	0	0	0	0	0	0	1	0	- - - -	0
* Is this the point of degeneration?																						

An 'omic' real data example

In **differential gene expression** we usually have results from **various biological experiments and/or platforms** for prespecified conditions such as tumor vs. normal (e.g. in tissue)
These results come in **rank lists** of genes (due to criteria such as fold change, multiplicity-controlled p-values, SAM-output)
Our goal is to consolidate these results
The problem:

- Usually such **lists have limited 'overlap'**
- There is **no operational concept of list conformity** when combining rank lists
- The theory of nonparametric order statistics (random variables) does not really help (due to lack of distributional models)

Medizinische Universität Graz

Some technical complications

Let us assume a pairwise consideration of order statistics (e.g. characterizing the rank order of genes resulting from various selection procedures)

- Considerable **overlap only for top-ranking objects**
- **Number of top-ranking objects very small compared to overall number** of rank positions
- Even the top-ranked positions can show **stochastic fluctuation** (no 'perfect' overlap)
- Hard to define an **end condition** for the set of top-ranking objects
- Need a formal concept to characterize **overlap degeneration**

Overall goal and attempts to achieve it

Aggregation of input information with respect to high conformity

Technically hard to achieve because

- no or **little prior information** (e.g. in molecular research)
- **lack of stochastic concepts** that are appropriate
- **numerically extremely expensive** (combinatorial approaches can be NP hard; see Fagin et al., 2003, SIAM JDM 17, 134-160)

Approaches taken so far:

- A **similarity score** for ordered lists (Yang et al., 2005, JBCB 4, 693-708; Bioconductor package `OrderedList`)
- Computer-intensive statistics (Cross Entropy Monte Carlo) for **rank aggregation** (Lin et al., forthcoming)

Medizinische Universität Graz

Our approach

We apply a **statistical approach** taking advantage of the theoretical concept of 'moderate deviations' (for this mathematical concept see e.g. Donoho et al., 1995, JRSS, B 57, 301-369)

- We are not aiming at an overall score or global rank aggregation
- We reduce the input information to a **sequence of indicator variables** with respect to the concordance of paired ranks
- Our focus is the **selection of a set of top-ranked objects** based on paired information
- We allow for the **realistic setting** of **irregular rankings** (**fuzzy instead of perfect overlap**)
- We **even allow for very small numbers of top-ranking objects** compared to the overall number of objects

Medizinische Universität Graz

Our methodology

We develop a methodology that allows us to **test for random degeneration of paired rank information**
This is equivalent to the **identification of that point where the data stream degenerates into noise**
Basic assumptions

- For the estimation of the point of degeneration into noise we assume independent **Bernoulli random variables**
- There are **no ties** in the rankings
- Under the condition of a general **decrease of the probability for concordance of rankings with increasing distance from the top rank** a formal inference model can be developed
- **Moderate deviation arguments are suitable for testing in a sequence of indicator values** (data stream)

Medizinische Universität Graz

Our methodology continued

- **The 2nd assessor, like the 1st, ranks the objects** $o_1, \ldots, o_N$ from 1 to N ($j = 1, \ldots, N$)
- **Indicator** $I_j = 1$ if the ranking given by the 2nd assessor to the object ranked j by the 1st **is not distant more than** m **from** j**, and** $I_j = 0$ **otherwise**
- Can symmetrise this definition by asking that both, or at least one, of the two distances not exceed 1
- Taking $m = 0$, symmetry already prevails, but then we have to adjust for irregular rankings
- Several **'regularization' (tuning) parameters** are introduced to account for the **closeness of the assessors' rankings and the degree of randomness in the assignments**

Medizinische Universität Graz

A simplified model

We shall assume that the Bernoulli random variables I_j are **independent**, somewhat a simplification
Note: The additional complication of modelling dependence in the nonstationary sequence of 0's and 1's would not pay off for our purposes

Our model:
Independent Bernoulli random variables $I_1, \ldots, I_N$ are observed, with $p_j \geq \frac{1}{2}$ for each $j \leq j_{0-2}$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$ for $j \geq j_0$
From this information we wish to estimate the value of j_0 (point of degeneration into noise)
The **'general decrease' of p_j for increasing j**, implied by this condition, need not be monotone
In some circumstances the threshold $\frac{1}{2}$ should be replaced by a different value

Algorithm

- Consists of an ordered sequence of **'test stages'** $s_1, s_2, \ldots$
- Stage s_k terminates a distance J_{s_k} into the sequence $I_1, \ldots, I_N$
- When k is odd, J_{s_k} is a potential lower bound to j_0
- Can show that when $k = 1$, the probability that $J_{s_{2k-1}}$ is a lower bound for j_0 is approximately equal to 1 under our model (analogous for each $k \geq 1$)
- Stage s_k starts by drawing a **'pilot sample' of size** ν, consisting of the set of values I_j for which j is among the first ν indices to the right of $J_{s_{k-1}} - r\nu$, if k is odd, or to the left of $J_{s_{k-1}} + r\nu$, if k is even ($r > 1$ fixed)
- The sequence of consecutive steps that leads from $J_{s_k} \pm r\nu$ to J_{s_k} is called the **'test stream'** for stage s_k

Algorithm continued: inference

- Pilot sample size to construct
  $\hat{p}_j^+ = \frac{1}{\nu} \sum_{\ell=j}^{j+\nu-1} I_\ell$ and $\hat{p}_j^- = \frac{1}{\nu} \sum_{\ell=j-\nu+1}^{j} I_\ell$
- **These quantities represent estimates of** p_j computed from the ν data pairs I_ℓ for which ℓ lies immediately to the right of j, or immediately to the left, respectively
- Pilot sample size ν **can be interpreted as smoothing parameter** (choice is critical; see simulations)
- Choose the **constant** $C > 0$ so that $z_\nu \equiv \left(C\,\nu^{-1}\log\nu\right)^{\frac{1}{2}}$ is a **moderate-deviation bound for testing the null hypothesis** H_0 **that** $p_k = \frac{1}{2}$ **for** ν **consecutive values of** k**, versus the alternative** H_1 **that** $p_k > \frac{1}{2}$ **for at least one of the values of** k
- Assuming that H_0 applies to the ν consecutive values of k in the respective series we reject H_0 if and only if $\hat{p}_j^{\pm} - \frac{1}{2} > z_\nu$

Algorithm continued: inference and tuning

- Under H_0, the variance of $\hat{p}_j^{\pm}$ equals $(4\nu)^{-1}$
- Therefore we should take $C > \frac{1}{4}$ **if we are to control moderate deviations**
- Can be compared with the construction of thresholds for wavelet expansions under moderate-deviation arguments
- **The test results depend on the choice of the tuning parameters**, i.e. r, C, ν, and m
- Because our mathematical model can never be more than an approximation to the complex decision problem an **iterative algorithm** was developed (adjusting for irregularity)
- A prototype is implemented **in the statistical computing environment R**

Medizinische Universität Graz

Simulation outline and results

Goal was to study the role of tuning parameters

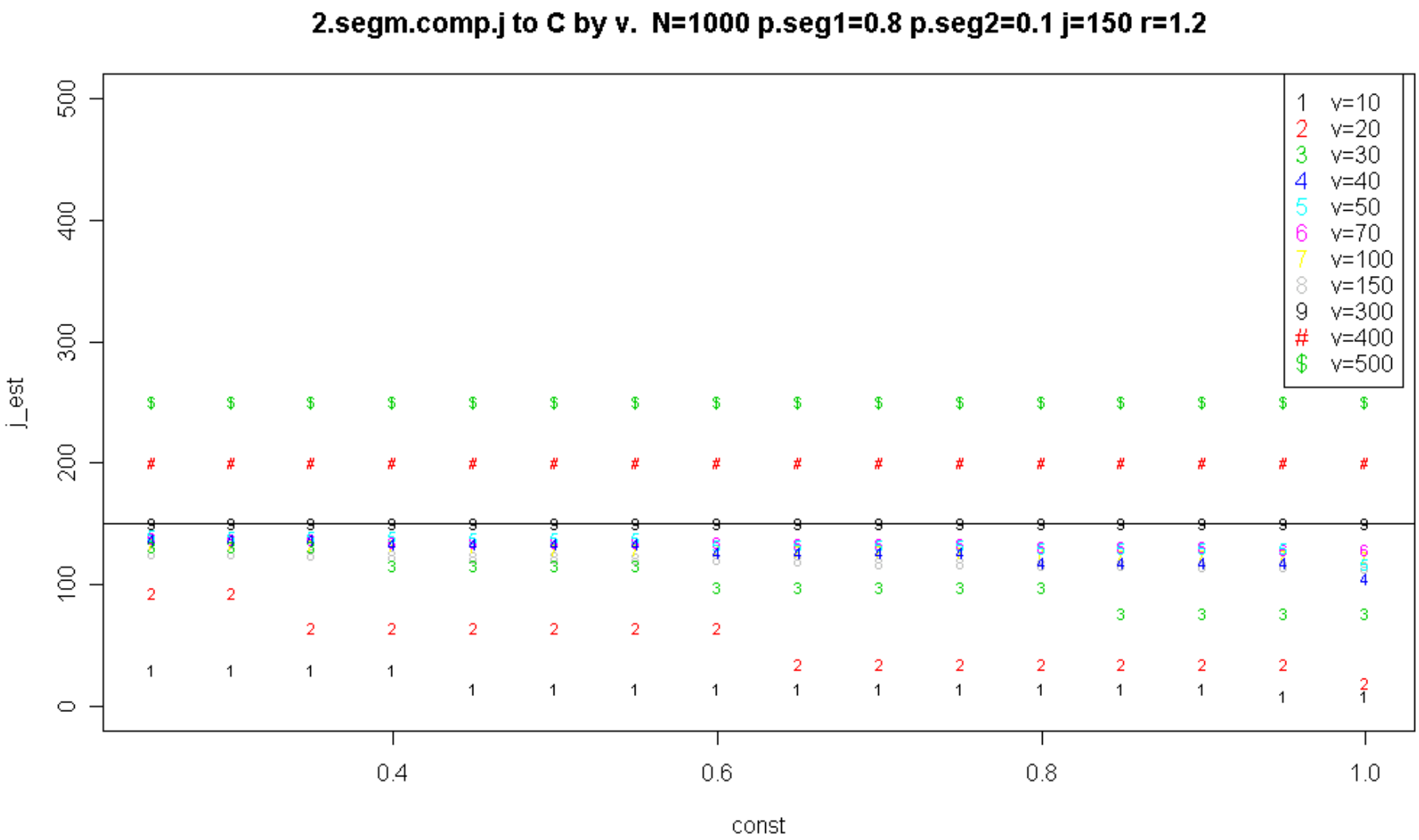
- Length of paired rank list $N = 1000$
- Two segments variable in length, separated by j_0, point of degeneration into noise
- $p.seg1 \in [0.6, 0.7, 0.8, 0.9, 1]$, $p.seg2 \in [0.1, 0.2, 0.3]$
- $C \in [0.251, 0.3, 0.35, 0.4, 0.45, 0.50, \ldots, 1]$
- $j \in [10, 20, 30, 40, 50, 100, 150, 200, 250, \ldots, 500]$
- $\nu \in [10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 350, 400]$

Main results

- C compensates for poor segment separability: $0.25 < C \leq 0.4$ is best choice
- The pilot sample size ν should be approximately $2 * j_0$
- Technical constant r can be fixed to, say, $r = 1.2$
- Choice of ν much more critical than of C

Medizinische Universität Graz

Typical simulation result: influence of ν and C on $\hat{j}_0$



Microarray data and results

Breast cancer data due to Sørlie et al. (2001, PNAS 98, 10869-10874)

- $N = 500$ genes selected by SAM
- 7 samples hybridized on different microarray platforms, we selected 2 of them (30 resp. 40 arrays)
- Input are 2 rank lists of differentially expressed genes

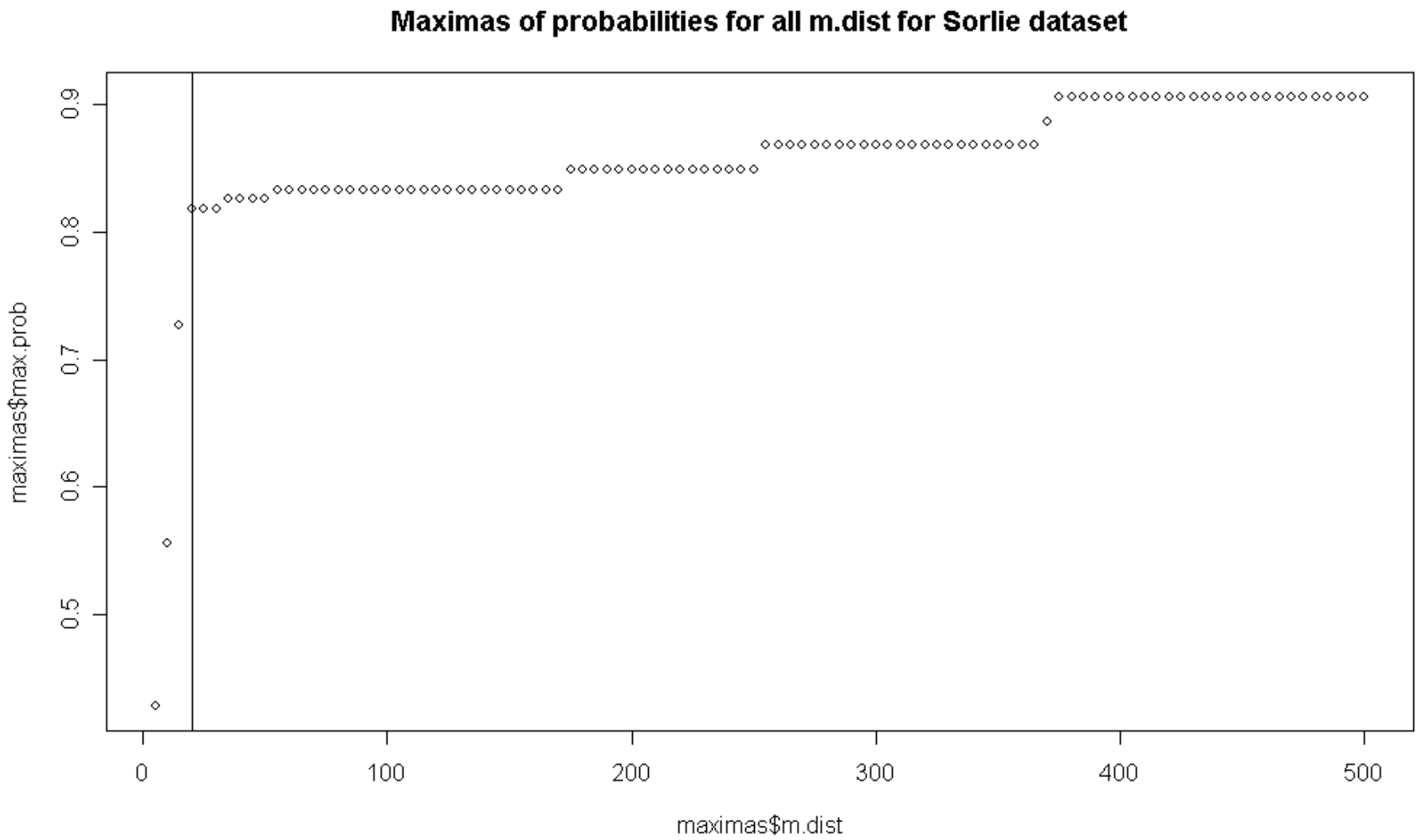
Goal is to estimate the point of degeneration into noise and a set of genes that is supported by both platforms

User has to specify distance parameter m (exploratory plot of the maximal proportions of ones for the range 5 to 500)

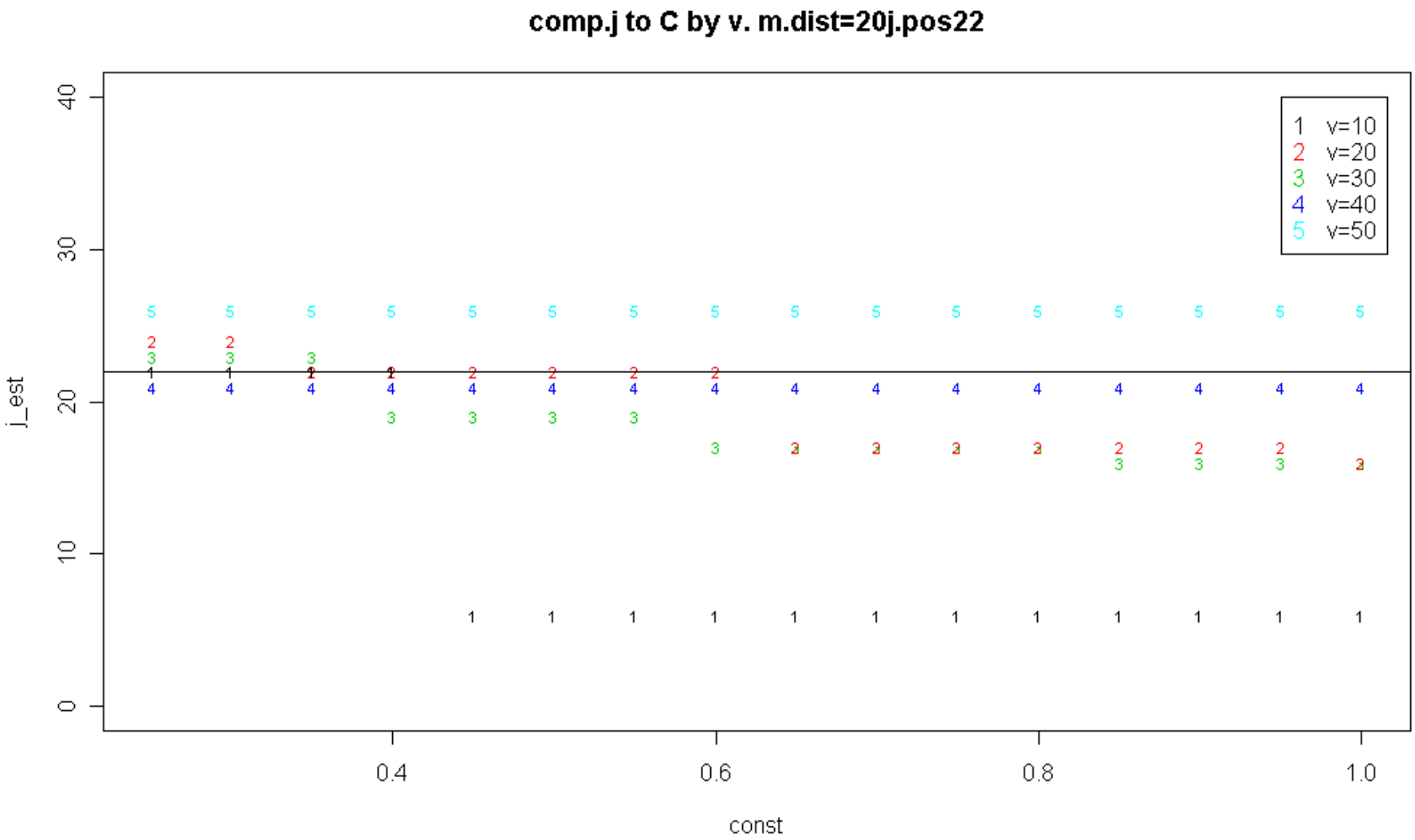
Results

- Exploratory plot hints at $m = 20$
- Pilot sample size $\nu = 10$ and $C \in [0.251, 0.3, 0.35, 0.4]$ resulted in a **point of degeneration** $\hat{j}_0 = 22$
- **The set size is** $\hat{j}_0 - 1 + m = 41$ **objects (genes)**

Microarray data: exploratory plot for *m* selection



Microarray data: influence of ν and C on $\hat{j}_0$



Summary

- **Our iterative algorithm**
  - **works perfectly well for regular rankings**
  - **works well even for irregular rankings**
  - **selects top ranking sets well** even when they are small compared to the overall number of objects
  - **is computationally highly efficient**
  - is not too sensitive to the choice of tuning parameters (apart from ν and m for real data)
  - is promising for real rank list data of unknown structure
- **Future research needs to address**
  - the problem of **objective choice of the smoothing parameter** ν (much more critical than choice of C)
  - the situation of **more than two assessors** (a generalization of our algorithm is feasible)
  - the problem of **missing rankings**