Order Optimal Matching for Incomplete Matching Problems

Massimo Cannas ✉
massimo.cannas@unica.it
Department of Business and Economic Sciences
University of Cagliari, Sardinia (Italy)

Optimal Matching Algorithms

Optimal matching algorithms are commonly used for balancing covariates in observational studies (Rosenbaum, 2002; Hansen et al., 2006). Given a set of treated units T and a set of control units C an optimal matching solution assigns each $t \in T$ to a subset of C in such a way that the total cost of matching is minimized. If the subset of C has unit size we talk of optimal pair matching. If all t_i are matchable we talk about a complete matching.
A complete pair matching may not possible because of *caliper* or because the number of treated units is *less than* the number of controls.

Order Optimal matching

Suppose matching priority is available. Can we pair the "most important" units of T? It is not obvious that this problem has a solution since for example the matchable subset t_1, t_2 cannot be compared straightforwardly with t_2, t_3, t_4: the former matches the two most important units but the latter matches more units. However, it is always possible to obtain a maximum size matching $t_1, \cdots t_m$ which is optimal in the following sense (Gale, 1968):

- $m \leq n$ for any other matchable subset $u_1, \cdots u_m$;
- $t_i \leq u_i$ for all $i \leq m$.

Order Optimal *versus* Cost Optimal

Order and cost optimal matchings are maximum size matching but they do not necessarily coincide.

`cost:`

	A	B	C	D
t1	0	0	0	∞
t2	∞	0	0	∞
t3	∞	0	∞	∞
t4	0	∞	0	∞
t5	0	∞	∞	0

`matlist(cost):`

t1	A B C
t2	B C
t3	B
t4	A C
t5	A D

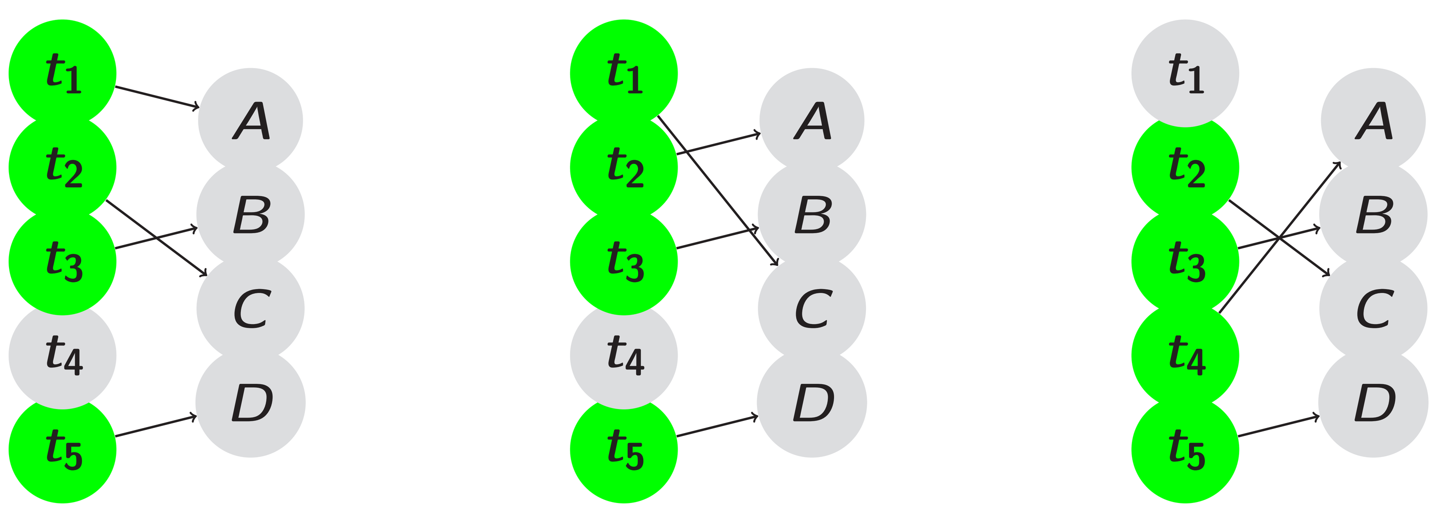


Figure: Several maximum size matching are possible from the `cost` matrix, both order optimal (down-left and down-center) and not (down-right) but also `BCOAD` (not shown). They are all minimum cost (**0**) matchings.

R implementation

Order optimal matching can be implemented via the R-package `OSDR`.

- The package main function `OSDR` gives an order optimal matching from the ordered list of suitable controls (all controls *within-caliper*).
- Utility `matlist` converts a cost matrix to a list of suitable controls (all controls with finite cost).
- The algorithm is based upon a constructive proof of Hall's theorem communicated by logician D.J. Shoesmith to mathematician Ian Anderson.
- The `OSDR` routine executes an algorithm for finding the optimum which runs in $\mathcal{O}(n^2)$.

Case study on gender gap

A research project required comparison of several work outcomes across men and women employees[a]. Matching women and men having the same job characteristics i) has a strong intuitive appeal ii) allows separation of the design and analysis stage.

Firm ID	N of men	N of women	Sample size
1	16	10	26
2	6	10	16
3	7	16	23
4	9	3	12
5	12	15	27
6	13	13	26
7	7	4	11
8	10	12	22
9	1	0	1
10	1	0	1
11	34	27	61
12	1	3	4
13	1	6	7
14	9	11	20
15	11	16	27
16	18	44	62

Salary gap increases with position in the organizational hierarchy, with highest differences at the highest levels where the variable part of remuneration is more variable (Lyness et al.,1997)
⇒ It is desirable to give matching priority to women executives in apical position. But in many firms classic optimal pair matching may drop some of the women in apical positions due to

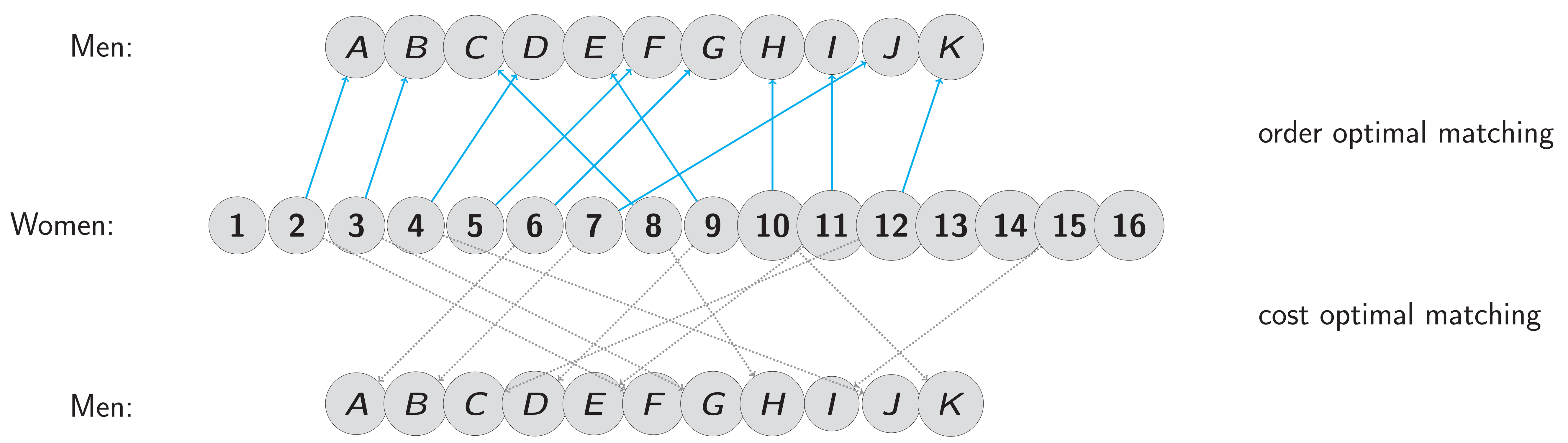
- caliper restriction;
- firms having more women than men executives.

Thus, we used the following "mixed" matching strategy

1. find the order optimal matchable subset;
2. find the minimum cost matching on this subset.

[a]The project involved several firms in different sectors and sample size based on the sustained cost. Sample sizes varied widely as based on sustained cost of research from 20 observations for smaller firms inserted to have a benchmark to a maximum of 100 observations for firms commissioning the research. The average non-response rate was about 30% so the final sample size range from 16 to 80. Archival data were collected via randomized stratified sampling on strata defined by `sex` (female=1 or male=0), age (years), `position` (top manager=4, middle manager/ executive=3, first line manager/supervisor=2), `education` (Post-graduate=5, Graduate=4, High school=3), `contract type` (fixed term=4,permanent=3)and `seniority` (years in the position). A questionnaire was distributed to retrieve information on compensation components.

Matching solution for company number 15



The cost and order optimal solution trade off higher priority matching units with matching cost. For this company woman 5 is matched in the order optimal — with an increase of $0.15\sigma_{cost}$ — instead of women 15 in the cost optimal solution.

Main References

Rosenbaum P.R. (2002): Observational Studies. *Springer Series in Statistics.*
Hansen B. and Olsen K. (2006) Optimal full matching and related designs via network flows *Journal of Computational and Graphical Statistics*, *15*, 609–627.
Lyness K and Thompson E. (1997) Above the glass ceiling? A comparison of matched samples of female and male executives *Journal of Applied Psycology*, *82*, 359–375.
Gale D. (1968): Optimal matching in an ordered set: an application of matroid theory *Journal of Combinatorial Theory*, *4*, 176–180.