Exact Methods for Hierarchical Clustering

Rick Willemsen, Carlo Cavicchia, Wilco van den Heuvel, Michel van de Velden

willemsen@ese.eur.nl

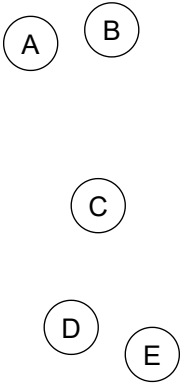
Econometric Institute
Erasmus University Rotterdam

May 17, 2023

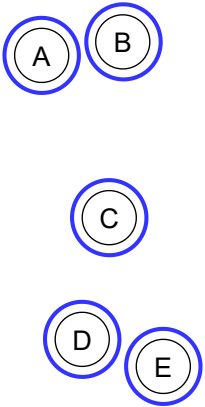
Table of Contents

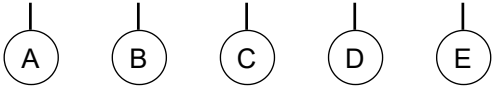
Table of Contents

What is clustering?

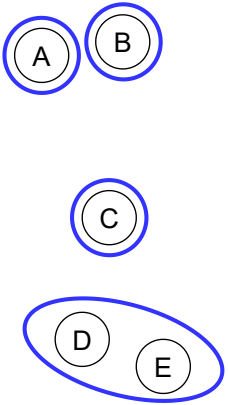


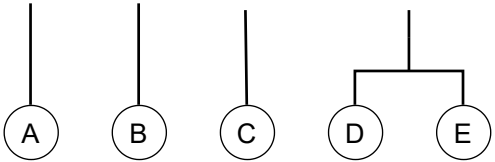
What is clustering?



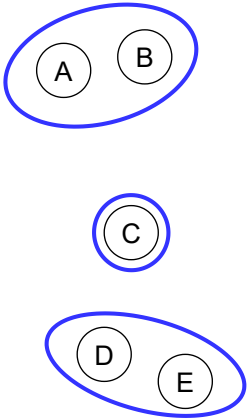


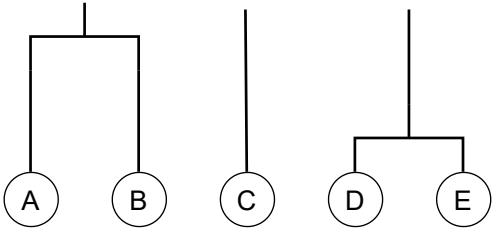
What is clustering?



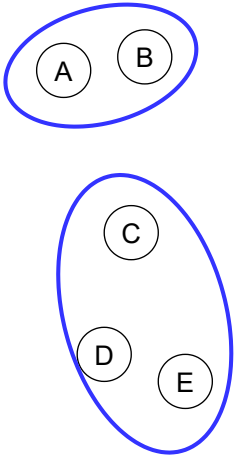


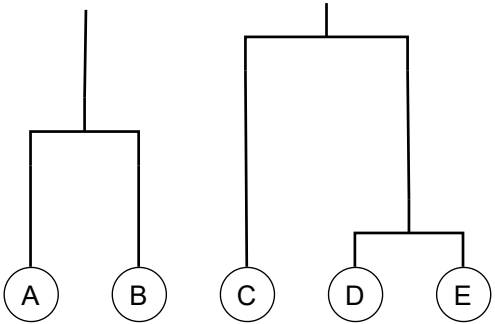
What is clustering?



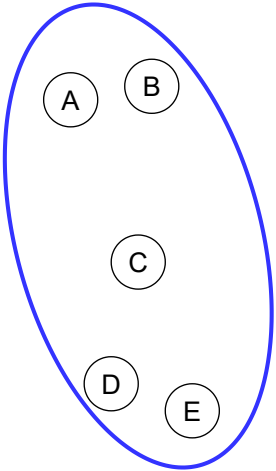


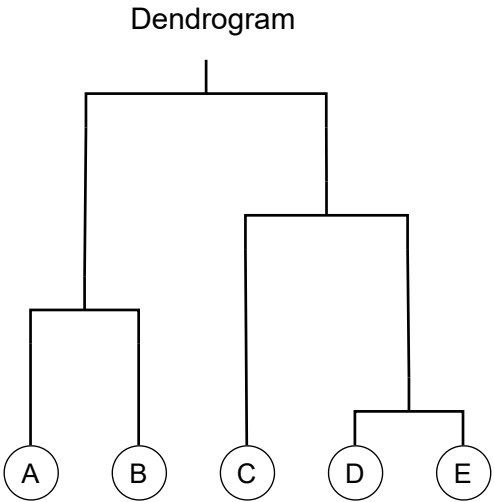
What is clustering?





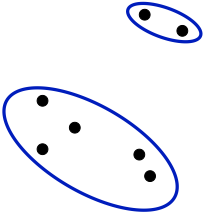
What is clustering?



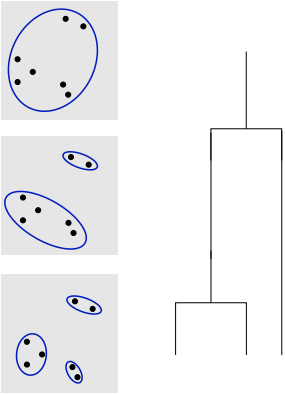


Clustering models

Partitional clustering



Hierarchical clustering

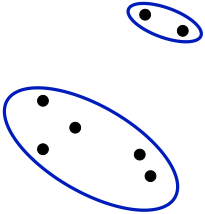


Clustering models

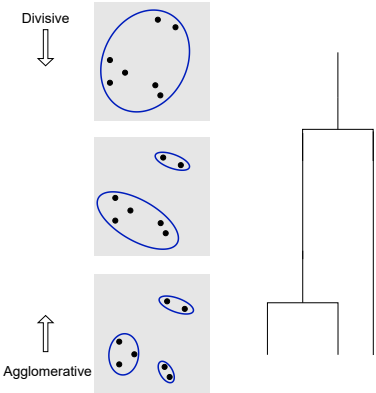
Partitional clustering

Minimise

obj_2



Hierarchical clustering



$\Rightarrow$ Heuristics are used

Disadvantages of hierarchical clustering heuristics

Recursive and greedy $\Rightarrow$  cannot undo a previous step

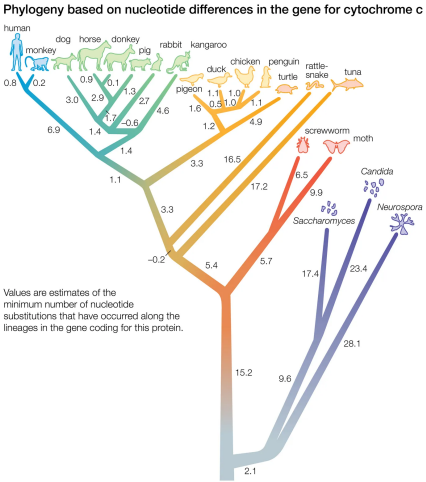
No statistical model or objective function $\Rightarrow$  cannot find best hierarchy

Presence of outliers $\Rightarrow$  mask clusters

All n partitions are created $\Rightarrow$  difficult to interpret

Why do we want to find the optimal hierarchy?

- We know there exists a ground truth (evolutionary trees, fMRI, languages)
- It takes a long time to collect data
- Measure performance of heuristics and approximation algorithms



[1] https://www.britannica.com/science/phylogeny/Taxonomic-systems

Table of Contents

Literature

Exact methods
partitional clustering

- Vinod (1969); Rao (1971)
- Jensen (1969)
- Koontz et al. (1975)
- Du Merle et al. (1999)
- Hansen and Mladenović (2001)
- Van Os and Meulman (2004)
- Xia and Peng (2005)
- Peng and Xia (2005)
- Brusco (2006)
- Aloise et al. (2012)
  $n = 2310$ in ~ 24 hours
- Ágoston and E.-Nagy (2021)
- Piccialli et al. (2022)
- Burgard et al. (2023)

Exact methods
hierarchical clustering

- Greenberg et al. (2021)
  $n \approx 15$ in ~ 24 hours

Objectives & approximations
hierarchical clustering

- Gilpin et al. (2013)
- Dasgupta (2016)
- Roy and Pokutta (2016)
- Gilpin and Davidson (2017)
- Charikar and Chatziafratis (2017)
- Moseley and Wang (2017)
- Cohen-Addad et al. (2019)
- Charikar et al. (2019)
- Wang and Moseley (2020)
- Chami et al. (2020)
- Naumov et al. (2021)
- Rajagopalan et al. (2021)
- Vichi et al. (2022)

Literature

Exact methods
partitional clustering

- Vinod (1969); Rao (1971)
- Jensen (1969)
- Koontz et al. (1975)
- Du Merle et al. (1999)
- Hansen and Mladenović (2001)
- Van Os and Meulman (2004)
- Xia and Peng (2005)
- Peng and Xia (2005)
- Brusco (2006)
- Aloise et al. (2012)
  $n = 2310$ in ~ 24 hours
- Ágoston and E.-Nagy (2021)
- Piccialli et al. (2022)
- Burgard et al. (2023)

Exact methods
hierarchical clustering

- Greenberg et al. (2021)
  $n \approx 15$ in ~ 24 hours

Objectives & approximations
hierarchical clustering

- Gilpin et al. (2013)
- Dasgupta (2016)
- Roy and Pokutta (2016)
- Gilpin and Davidson (2017)
- Charikar and Chatziafratis (2017)
- Moseley and Wang (2017)
- Cohen-Addad et al. (2019)
- Charikar et al. (2019)
- Wang and Moseley (2020)
- Chami et al. (2020)
- Naumov et al. (2021)
- Rajagopalan et al. (2021)
- Vichi et al. (2022)

Literature

Exact methods
partitional clustering

- Vinod (1969); Rao (1971)
- Jensen (1969)
- Koontz et al. (1975)
- Du Merle et al. (1999)
- Hansen and Mladenović (2001)
- Van Os and Meulman (2004)
- Xia and Peng (2005)
- Peng and Xia (2005)
- Brusco (2006)
- Aloise et al. (2012)
  $n = 2310$ in ~ 24 hours
- Ágoston and E.-Nagy (2021)
- Piccialli et al. (2022)
- Burgard et al. (2023)

Exact methods
hierarchical clustering

- Greenberg et al. (2021)
  $n \approx 15$ in ~ 24 hours

Objectives & approximations
hierarchical clustering

- Gilpin et al. (2013)
- Dasgupta (2016)
- Roy and Pokutta (2016)
- Gilpin and Davidson (2017)
- Charikar and Chatziafratis (2017)
- Moseley and Wang (2017)
- Cohen-Addad et al. (2019)
- Charikar et al. (2019)
- Wang and Moseley (2020)
- Chami et al. (2020)
- Naumov et al. (2021)
- Rajagopalan et al. (2021)
- Vichi et al. (2022)

Table of Contents

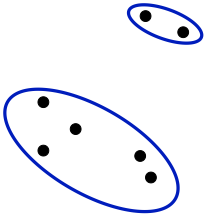
Problem description details

Partitional clustering

Divide n objects into k clusters

Minimise

obj_2

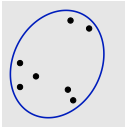


Hierarchical clustering (Vichi et al., 2022)

Construct hierarchy of at most K levels

Minimise

obj_1



+

obj_2

+

obj_3

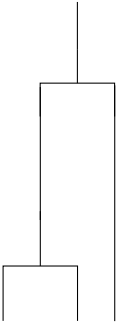


Table of Contents

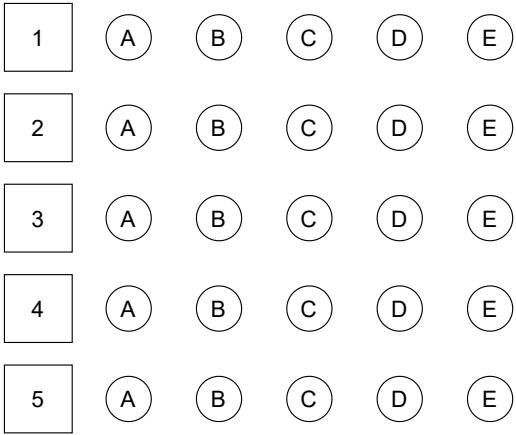
Methodology

Partitional clustering				Hierarchical clustering
MILP (Ágoston and E.-Nagy, 2021)	+	objective (Vichi et al., 2022)	$\Rightarrow$	MILP
B&P (Aloise et al., 2012)	+	objective (Vichi et al., 2022)	$\Rightarrow$	B&P

Our MILP can solve $n = 15$ objects and $K = 15$ levels in 5 seconds

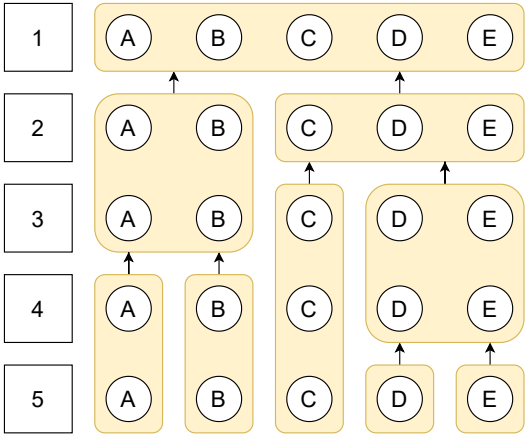
$\Rightarrow$ We focus on branch-and-price

Formulation - example



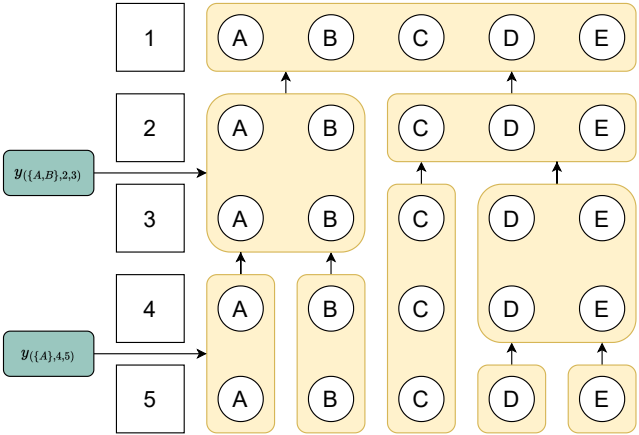
We solve K nested partitional clustering problems

Formulation - example



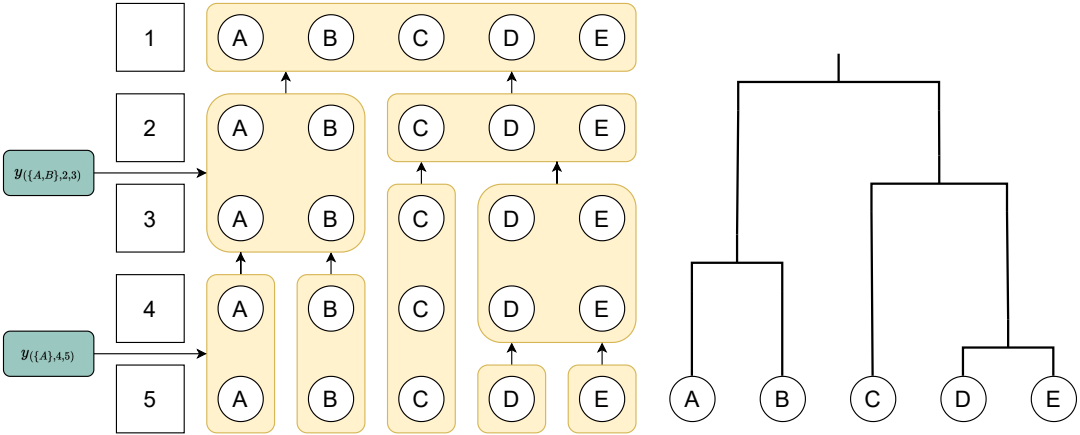
We solve K nested partitional clustering problems

Formulation - example



We solve K nested partitional clustering problems

Formulation - example



We solve K nested partitional clustering problems

Formulation details

$$\min \sum_{k=1}^{K} \sum_{h=1}^{k} \sum_{g \in G_{hk}} c_{ghk} y_{ghk}$$ ⇒ Select clusters based on distances

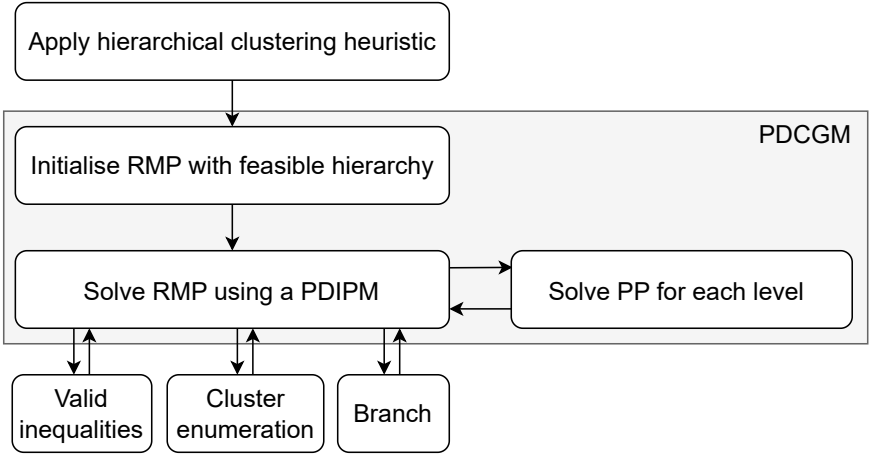
$$\text{s.t.} \sum_{k=h}^{K} \sum_{j=1}^{h} \sum_{g \in G_{jk}} a_{igjk} y_{gjk} = 1, \forall i = 1, \ldots, n, \forall h = 1, \ldots, K$$ ⇒ Each object is covered once

$$\sum_{k=h}^{K} \sum_{g \in G_{hk}} y_{ghk} = 2, \forall h = 2, \ldots, K$$ ⇒ Merge 2 clusters

$$\sum_{j=1}^{h} \sum_{g \in G_{jh}} y_{gjh} = \begin{cases} 1, \forall h = 2, \ldots, K-1 \\ h, h = K \end{cases}$$ ⇒ Begin a new cluster

$$y_{ghk} \in \mathbb{B}, \forall g \in G_{hk}, \forall h = 1, \ldots, k, \forall k = 1, \ldots, K$$ ⇒ Select cluster from level h up to k

Branch-and-price overview heuristic RMP

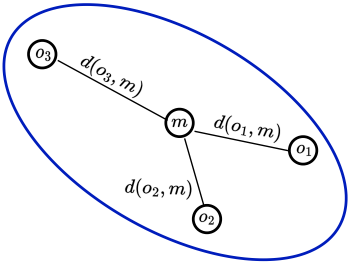


Pricing problem

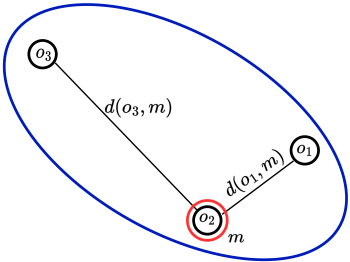
For each level pair (h, k) we solve a pricing problem

$$\pi^* = \min_w \mu + \sum_{j=1}^{n} (d(o_j, m) - \lambda_j) w_j$$

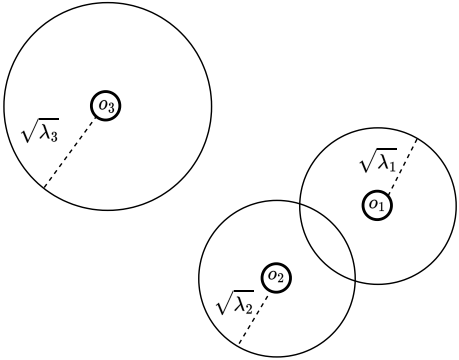
Centroids



Medoids

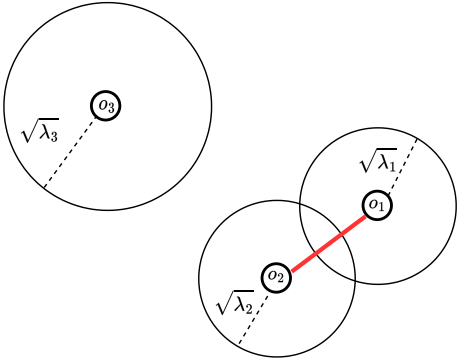


Pricing problem - centroids details



Optimal solution to PP is a clique in G (Aloise et al., 2012)

Pricing problem - centroids details



Optimal solution to PP is a clique in G (Aloise et al., 2012)

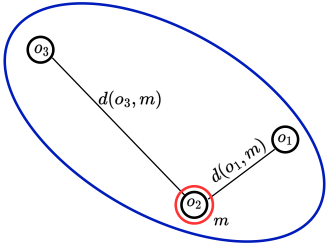
Pricing problem - medoids

Lemma

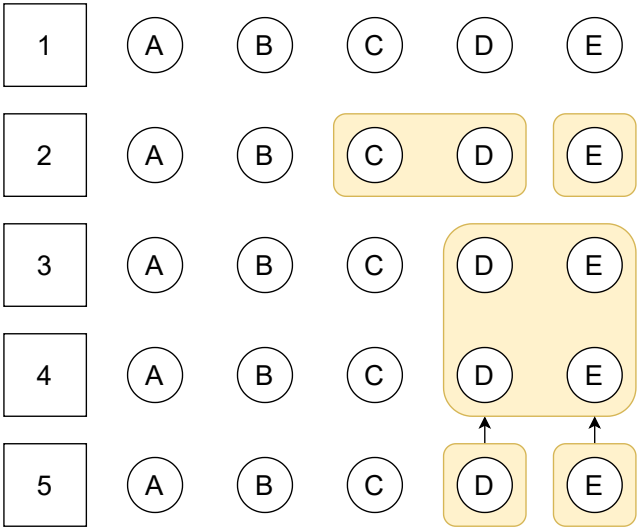
Optimal solution to PP contains all objects o_j satisfying $d(o_j, m) \leq \lambda_j$

Greedy algorithm

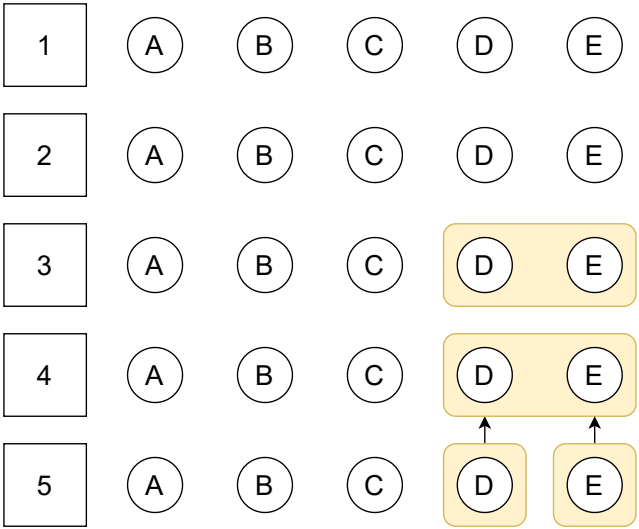
1. Consider each object o_i as medoid
2. Add all objects o_j satisfying $d(o_j, m) \leq \lambda_j$



Fractional solution - conflict 1



Fractional solution - conflict 2



Valid inequalities

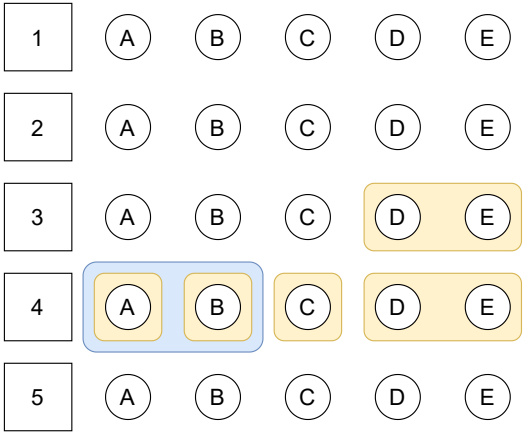
- Type 1 (conflict 1)
- Type 2 (conflict 1 & 2)
  - Consider a conflict graph $G_{conflict}$
  - Finding a maximum weight clique is NP hard (Schrijver, 2003)
  - Simple separation heuristic (Hoffman and Padberg, 1993)

Preliminary results show that we find the optimal solution when using a cutting plane heuristic

Valid inequality - Type 2

Simple separation heuristic (Hoffman and Padberg, 1993)

- Constraint: each object is covered exactly once (even in fractional solution)
- We look for $y_{ghk} > 0$ which conflicts with all selected variables in constraint



Cluster enumeration

Results indicate a small duality gap $\alpha \rightarrow$ cluster enumeration (Baldacci et al., 2008; Pecin et al., 2017)

Theorem

All solutions to PP with $RC \in [0, \alpha]$ form a clique in $G(\alpha)$

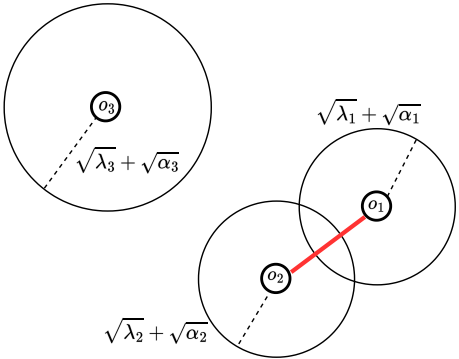
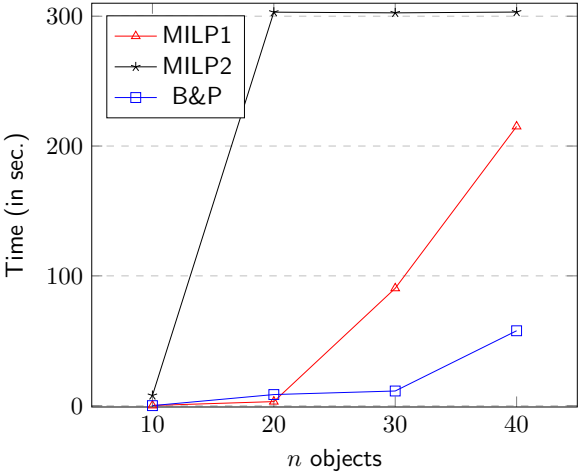


Table of Contents

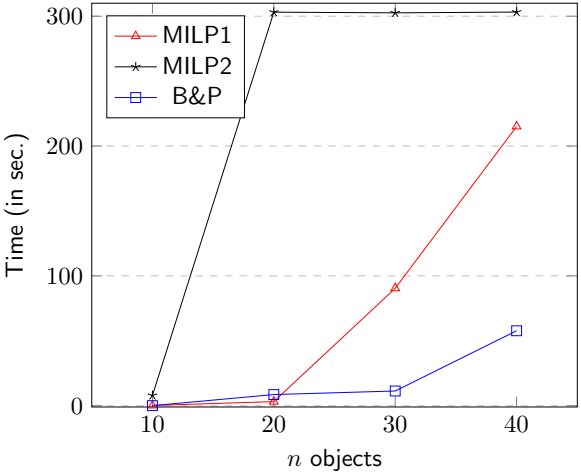
Results

Data	Purpose
Randomly generated data	MILP vs basic B&P
Labelled instances from Thrun and Ultsch (2020)	validate objective function
Real-world instances	scalability of B&P

Randomly generated data - MILP vs basic B&P details

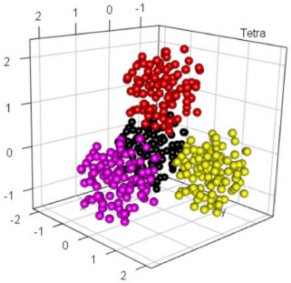


Randomly generated data - MILP vs basic B&P details

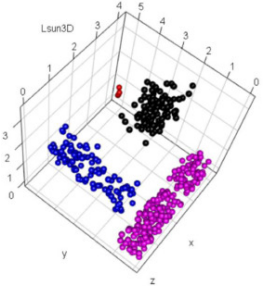


The heuristic of Vichi et al. (2022) did **not** find the optimal solution in 10 out of 80 instances

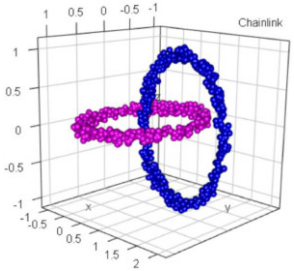
Empirical validation of the objective function



(a) Tetra



(b) Lsun3D



(c) Chainlink

Instance (sample $n = 50$)	ARI	
	Centroid	Medoid
Tetra	1.00	1.00
Lsun3D	0.65	0.70
Chainlink	0.02	0.26

Real-world instances

Instance	n	Time (min)	
		$K = 3$	$K = 6$
Ruspini	75	0.69	1.03
Iris	150	8.80	16.91
Seeds	210	109.27	-

We can solve more realistic instances to optimality

Table of Contents

Conclusion

Solve hierarchical clustering problems to optimality by solving K nested partitional clustering problems

- Mixed-integer linear programming
- Branch-and-price
- General objective function

Computational results

- We empirically validated correctness of the objective function
- Small instances $\Rightarrow$ heuristics do not always find optimal solution

Further research

- Under which conditions do greedy heuristics (e.g. agglomerative) obtain optimal solutions
- Investigate the performance of commonly used heuristics

Thank you for listening!

References I

Aloise, D., Hansen, P., and Liberti, L. (2012). An improved column generation algorithm for minimum sum-of-squares clustering. *Mathematical Programming*, 131:195–220.

Andersen, E. D., Gondzio, J., Mészáros, C., and Xu, X. (1996). Implementation of interior point methods for large scale linear programming. Technical report.

Baldacci, R., Christofides, N., and Mingozzi, A. (2008). An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts. *Mathematical Programming*, 115:351–385.

Brusco, M. J. (2006). A repetitive branch-and-bound procedure for minimum within-cluster sums of squares partitioning. *Psychometrika*, 71:347–363.

Burgard, J. P., Costa, C. M., Hojny, C., Thomas, K., and Schmidt, M. (2023). Mixed-integer programming techniques for the minimum sum-of-squares clustering problem. *Journal of Global Optimization*, ?(?):?

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. (2020). From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33.

Charikar, M. and Chatziafratis, V. (2017). Approximate hierarchical clustering via sparsest cut and spreading metrics. *Proceedings of the 2017 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 841–854.

Charikar, M., Chatziafratis, V., and Niazadeh, R. (2019). Hierarchical clustering better than average-linkage. *Proceedings of the 2019 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2291–2304.

Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., and Mathieu, C. (2019). Hierarchical clustering: Objective functions and algorithms. *Journal of the Association for Computing Machinery*, 66:1–42.

References II

Dasgupta, S. (2016). A cost function for similarity-based hierarchical clustering. *STOC '16: Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, page 118–127.

Du Merle, O., Hansen, P., Jaumard, B., and Mladenovic, N. (1999). An interior point algorithm for minimum sum-of-squares clustering. *SIAM Journal on Scientific Computing*, 21(4):1485–1505.

Gilpin, S. and Davidson, I. (2017). A flexible ilp formulation for hierarchical clustering. *Artificial Intelligence*, 244:95–109.

Gilpin, S., Nijssen, S., and Davidson, I. (2013). Formalizing hierarchical clustering as integer linear programming. *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 27.

Gondzio, J., González-Brevis, P., and Munari, P. (2016). Large-scale optimization with the primal-dual column generation method. *Mathematical Programming Computation*, 8:47–82.

Gondzio, J. and Sarkissian, R. (1996). Column generation with a primal-dual method. Technical report.

Greenberg, C. S., Macaluso, S., Monath, N., Dubey, A., Flaherty, P., Zaheer, M., Ahmed, A., Cranmer, K., and McCallum, A. (2021). Exact and approximate hierarchical clustering using a^*. *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, PMLR*, 161:2061–2071.

Hansen, P. and Mladenović, N. (2001). J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34(2):405–413.

Hoffman, K. L. and Padberg, M. (1993). Solving airline crew scheduling problems by branch-and-cut. *Management Science*, 39:657–682.

Jensen, R. E. (1969). A dynamic programming algorithm for cluster analysis. *Operations Research*, 17(6):927–1092.

References III

Koontz, W., Narendra, P. M., and Fukunaga, K. (1975). A branch and bound clustering algorithm. *IEEE Transactions on Computers*, 24(9):908–915.

Mehrotra, S. (1992). On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4):575–601.

Moseley, B. and Wang, J. R. (2017). Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. *Advances in Neural Information Processing Systems*, 30:3097–3106.

Naumov, S., Yaroslatsev, G., and Avdiukhin, D. (2021). Objective-based hierarchical clustering of deep embedding vectors.

Pecin, D., Pessoa, A., Poggi, M., and Uchoa, E. (2017). Improved branch-cut-and-price for capacitated vehicle routing. *Mathematical Programming Computation*, 9:61–100.

Peng, J. and Xia, Y. (2005). A new theoretical framework for k-means-type clustering. *Foundations and Advances in Data Mining*, 180:79–96.

Piccialli, V., Sudoso, A. M., and Wiegele, A. (2022). Sos-sdp: An exact solver for minimum sum-of-squares clustering. *INFORMS Journal on Computing*, 34(4):2144–2162.

Rajagopalan, A., Vitale, F., Vainstein, D., Citovsky, G., Procopiuc, C. M., and Gentile, C. (2021). Hierarchical clustering of data streams: Scalable algorithms and approximation guarantees. *Proceedings of the 38th International Conference on Machine Learning*, 139:8799–8809.

Rao, M. R. (1971). Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, 66:622–626.

References IV

Roy, A. and Pokutta, S. (2016). Hierarchical clustering via spreading metrics. *Advances in Neural Information Processing Systems*, 29.

Schrijver, A. (2003). *Combinatorial Optimization*. Springer.

Thrun, M. C. and Ultsch, A. (2020). Clustering benchmark datasets exploiting the fundamental clustering problems. *Data in Brief*, 30.

Van Os, B. and Meulman, J. (2004). Improving dynamic programming strategies for partitioning. *Journal of Classification*, 21:207–230.

Vichi, M., Cavicchia, C., and Groenen, P. J. F. (2022). Hierarchical means clustering. *Journal of Classification*, 39:553–577.

Vinod, H. D. (1969). Integer programming and the theory of grouping. *Journal of the American Statistical association*, 65:506–519.

Wang, Y. and Moseley, B. (2020). An objective for hierarchical clustering in euclidean space and its connection to bisecting k-means. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:6307–6314.

Xia, Y. and Peng, J. (2005). A cutting algorithm for the minimum sum-of-squared error clustering. *Proceedings of the SIAM International Data Mining Conference*, pages 150–160.

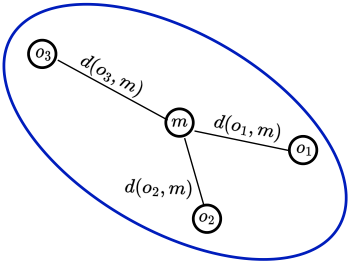
Ágoston, K. C. and E.-Nagy, M. (2021). Mixed integer linear programming formulation for k-means cluster problem.

Problem description return

Partitional clustering

Divide n objects into k clusters

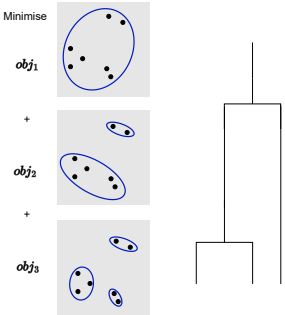
$$\min \sum_{g=1}^{k} \sum_{i=1}^{n} d(o_i, m_g) x_{ig}$$



Hierarchical clustering

Construct hierarchy of at most K levels

$$\min \sum_{k=1}^{K} \sum_{g=1}^{k} \sum_{i=1}^{n} d(o_i, m_g) x_{igk}$$



Objective function return

$$\min \sum_{k=1}^{K} \sum_{h=1}^{k} \sum_{g \in G_{hk}} w_{ghk} c_{ghk} y_{ghk}$$

Options for weight

- $w_{ghk} = \frac{1}{k-h+1}$, we no longer have "double-counting"
- $w_{ghk} = 1$, same as Vichi et al. (2022)

Initialisation return

Apply Q times the hierarchical clustering heuristic proposed by Vichi et al. (2022)

For $2 \leq k \leq K$, perform following steps

1. Solve k-means++
2. Perform agglomerative heuristic (Ward's method) from k to 2
3. Perform divisive heuristic (bisecting k-means) from k to K

This results in

- $Q(K-1)$ start solutions for our RMP
- A valid upper bound on the objective

Restricted master problem return

We have an exponential number of clusters

- In RMP we relax the binary variables and consider a subset of clusters

Large tailing-off effect

- In partitional clustering IPMs are successful (Du Merle et al., 1999; Aloise et al., 2012)
- Gondzio et al. (2016) show that PDCGM outperforms other IPMs
- We implement PDCGM described in Gondzio and Sarkissian (1996)
- We apply standard PDIPM (Andersen et al., 1996), with Mehrotra correction (Mehrotra, 1992)

Pricing problem - centroids return

$$\min \mu + \sum_{i=1}^{n} (||(o_i - m)||^2 - \lambda_i) w_i$$

Replace explicit expression for centroid (Edwards and Cavalli-Sforza, 1965)

$$\min \mu + \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (||o_i - o_j||^2 - \lambda_i - \lambda_j) w_i w_j - \sum_{i=1}^{n} \lambda_i w_i}{\sum_{i=1}^{n} w_i}$$

Find a clique in a graph G (Aloise et al., 2012)

$$\min \mu + \frac{\sum_{i=1}^{n'} \sum_{j=i+1}^{n'} (||o_i - o_j||^2 - \lambda_i - \lambda_j) w_i w_j - \sum_{i=1}^{n'} \lambda_i w_i}{\sum_{i=1}^{n'} w_i}$$

Apply Dinkelbach's algorithm (Du Merle et al., 1999)

$$\min \mu + \sum_{i=1}^{n'} \sum_{j=i+1}^{n'} (||o_i - o_j||^2 - \lambda_i - \lambda_j) w_i w_j - \sum_{i=1}^{n'} (\lambda_i + f_k) w_i$$

Unconstrained binary quadratic problem

Randomly generated data return

Data generating process

- n objects
- $p = 2$-dimensional space
- $k = 5$ clusters
- add noise

We obtain a hierarchy with $K = 5$ levels

Randomly generated data - basic B&P

n	Time (in sec.)				Avg. #iter	Avg. #columns	Avg. #QP
	Total	Heuristic	RMP	PP			
10	0.05	0.08	0.030	0.01	9	100	2,400
20	8.58	0.17	1.59	0.26	29	550	16,200
30	37.39	0.28	34.17	2.64	58	1,390	48,900
40	49.57	0.35	40.92	7.79	51	1,200	59,500