Minimum-Width Confidence Bands via Constraint Optimization

Jeremias Berg[1] Emilia Oikarinen[2]
Matti Järvisalo[1] Kai Puolamäki[2]

[1]HIIT, Dept. Computer Science, University of Helsinki, Finland

[2]Finnish Institute of Occupational Health, Helsinki, Finland

CP 2017
Melbourne
August 29, 2017

Motivation: Confidence Intervals

- Confidence Interval, important tool in data analysis.
  - Summarize data, estimate accuracy, posterior distribution etc.
  - Advantages to p-values.
  - Significance AND effect size.

Gardner and Altman [1986]

Nuzzo [2014]

Trafimow and Marks [2015]

Woolston [2015]

Korpela et al. [2014]

Hyndman and Fan [1996]

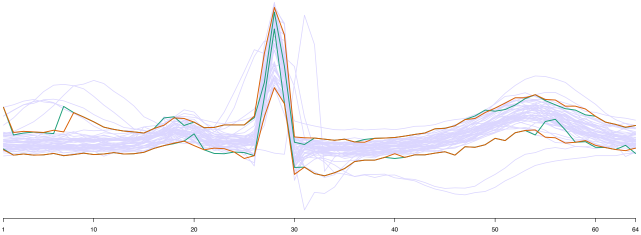
Kolsrud [2007]

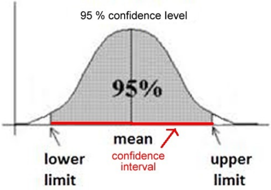
Schüssler and Trede [2016]

Motivation: Confidence Intervals

- Confidence Interval, important tool in data analysis.
  - Summarize data, estimate accuracy, posterior distribution etc.
  - Advantages to p-values.
  - Significance AND effect size.
- Easily interpretable together with the data

Gardner and Altman [1986]
Nuzzo [2014]
Trafimow and Marks [2015]
Woolston [2015]





Related work

- Many approaches to single univariate distributions Hyndman and Fan [1996], extensions to multivariate non-trivial
  - Extensions of p-values received more interest (multiple hypothesis testing).
- Computing *confidence bands* (multivariate confidence intervals) can be formulated as combinatorial optimization → Minimum-Width Confidence Band problem (MWCB)
  - Schüssler and Trede [2016]: MIP model for Min Width Envelope (MWCB(k))
  - Korpela et al. [2017]: More general definition. (MWCB(k, s)), Greedy Approximation

Related work

- Many approaches to single univariate distributions Hyndman and Fan [1996], extensions to multivariate non-trivial
  - Extensions of p-values received more interest (multiple hypothesis testing).
- Computing *confidence bands* (multivariate confidence intervals) can be formulated as combinatorial optimization $\rightarrow$ Minimum-Width Confidence Band problem (MWCB)
  - Schüssler and Trede [2016]: MIP model for Min Width Envelope (MWCB(k))
  - Korpela et al. [2017]: More general definition. (MWCB(k, s)), Greedy Approximation

Related work

- Many approaches to single univariate distributions Hyndman and Fan [1996], extensions to multivariate non-trivial
  - Extensions of p-values received more interest (multiple hypothesis testing).
- Computing *confidence bands* (multivariate confidence intervals) can be formulated as combinatorial optimization $\rightarrow$ Minimum-Width Confidence Band problem (MWCB)
  - Schüssler and Trede [2016]: MIP model for Min Width Envelope (MWCB(k))
  - Korpela et al. [2017]: More general definition. (MWCB(k, s)), Greedy Approximation

Our Contributions

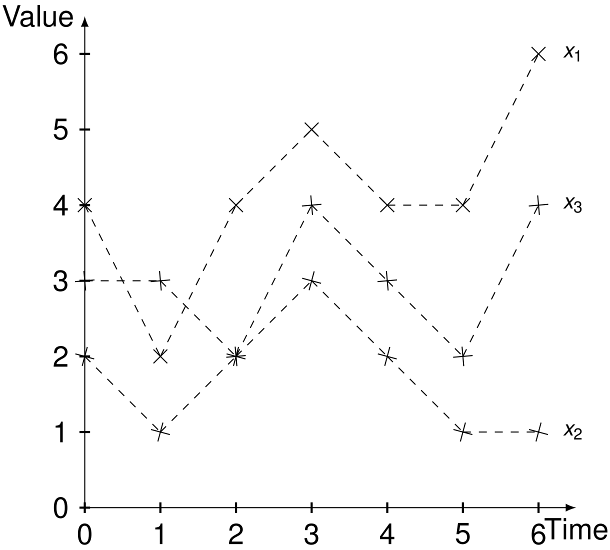
- Highlight and address drawbacks of previous formulations of MWCB with a new formulation.
- Propose 2 constraint optimization models and an approximative greedy procedure.
- Experimental Evaluation (MIP and MaxSAT).

Outline

- Minimum-Width Confidence Band Problem (MWCB)
- Constraint Models for MWCB
- A Greedy Algorithm for MWCB
- Experimental Evaluation
- Summary

Minimum-Width Confidence Band Problem

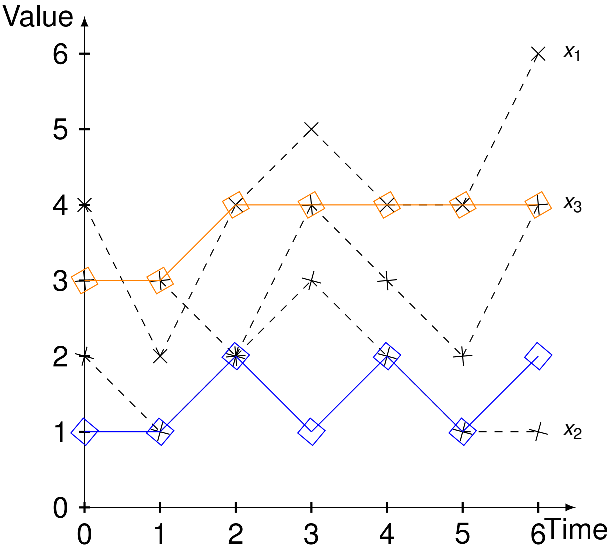
MWCB as Combinatorial Optimization



Setting:

Set of data vectors x_i each with m components (timepoints)

MWCB as Combinatorial Optimization



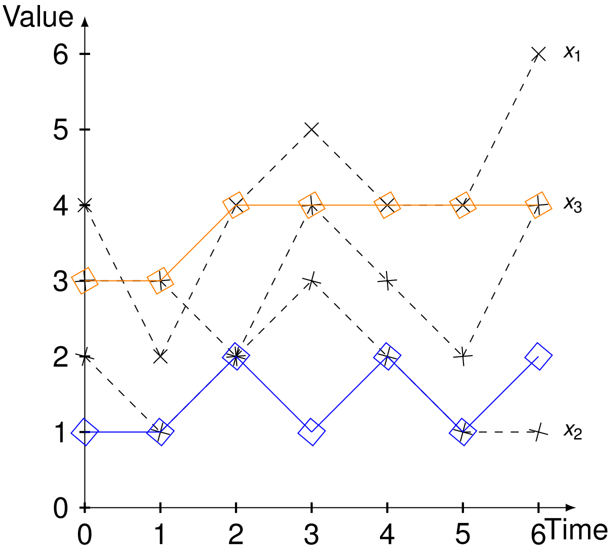
Setting:

Set of data vectors x_i each with m components (timepoints)

Confidence Band:

A pair of vectors (l, u) s.t. $l_i < u_j$ for all i

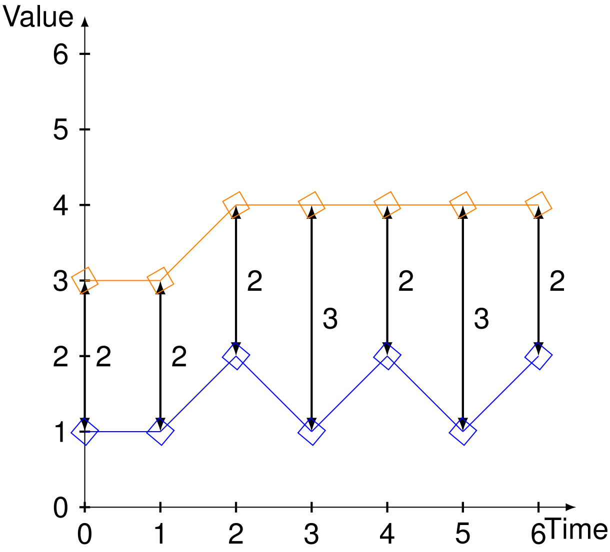
MWCB as Combinatorial Optimization



Task:

Compute minimum size CB with at most k outliers

MWCB as Combinatorial Optimization

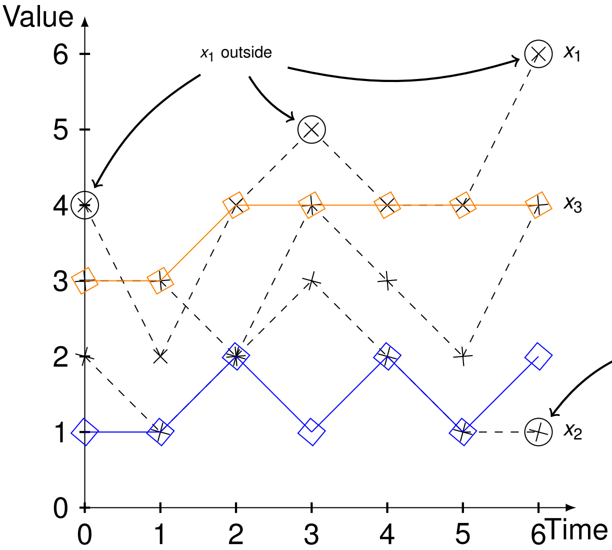


Task:

Compute minimum size CB with at most k outliers

Size of CB:

$5 \cdot 2 + 2 \cdot 3 = 16$

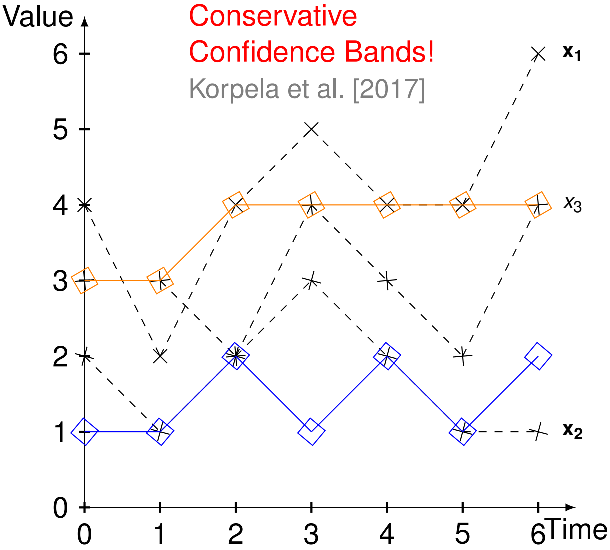


Task:

Compute minimum size CB with at most k outliers

MWCB(k)

A point x_i is an outlier if it lies outside the CB at any time

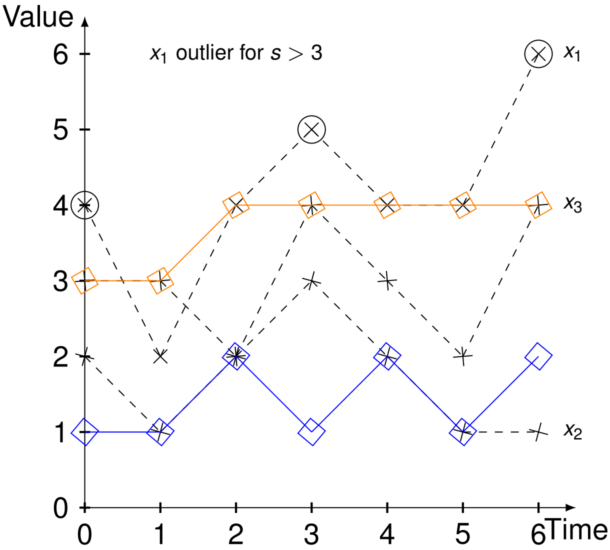


Task:

Compute minimum size CB with at most k outliers

MWCB(k)

A point x_i is an outlier if it lies outside the CB at any time



Task:

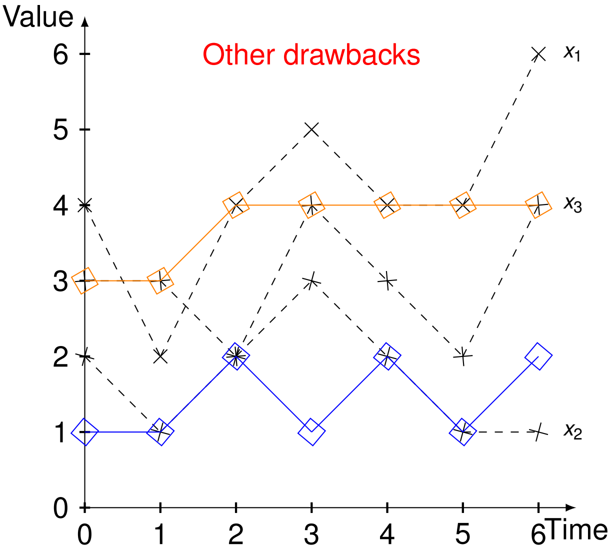
Compute minimum size CB with at most k outliers

MWCB(k)

A point x_i is an outlier if it lies outside the CB at any time

MWCB(k, s)

A point x_i is an outlier if it lies outside the CB more than s times



Task:

Compute minimum size CB with at most k outliers

MWCB(k)

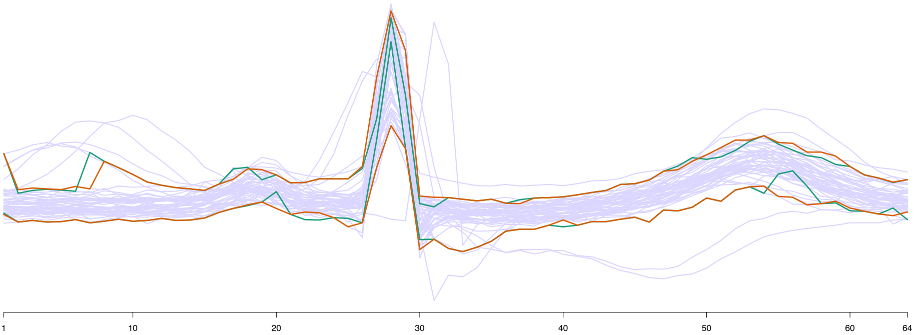
A point x_i is an outlier if it lies outside the CB at any time

MWCB(k, s)

A point x_i is an outlier if it lies outside the CB more than s times

Drawback of MWCB(k, s), real-world data

Korpela et al. [2014]



Drawback of MWCB(k, s), real-world data

Korpela et al. [2014]

Our problem formulation

MWCB(k, s, t)

- Compute minimum size $CB = (l, u)$ with at most k outliers AND at most t datapoints outside CB at any time.
- A point x_i is an outlier if it lies outside the CB more than s times

Relationships between formulations

$$\text{MWCB}(k) = \text{MWCB}(k, 0) = \text{MWCB}(k, 0, n)$$

$$\text{MWCB}(k, s) = \text{MWCB}(k, s, n)$$

Our problem formulation

MWCB(k, s, t)

- Compute minimum size $CB = (l, u)$ with at most k outliers AND at most t datapoints outside CB at any time.
- A point x_i is an outlier if it lies outside the CB more than s times

Relationships between formulations

$$\text{MWCB}(k) = \text{MWCB}(k, 0) = \text{MWCB}(k, 0, n)$$

$$\text{MWCB}(k, s) = \text{MWCB}(k, s, n)$$

Our problem formulation

MWCB(k, s, t)

- Compute minimum size $CB = (l, u)$ with at most k outliers AND at most t datapoints outside CB at any time.
- A point x_i is an outlier if it lies outside the CB more than s times

Relationships between formulations

$$\text{MWCB}(k) = \text{MWCB}(k, 0) = \text{MWCB}(k, 0, n)$$

$$\text{MWCB}(k, s) = \text{MWCB}(k, s, n)$$

NOTE: All problems are NP-hard!

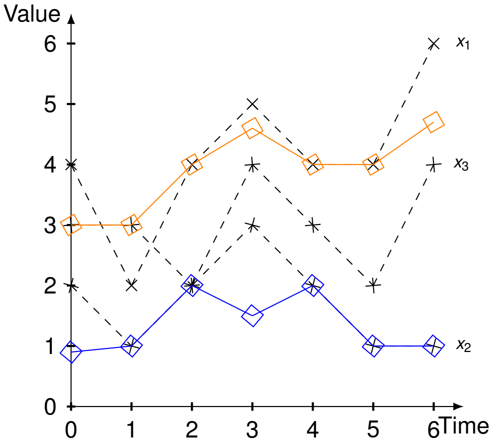
Constraint models and Greedy Algorithm for MWCB(k, s, t)

Variables in constraint models

- Binary Variables: d_{ij}, y_i
  - $d_{ij} = 1$ iff x_i outside CB at time j
  - $y_i = 1$ iff point x_i is an outlier

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0,1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).

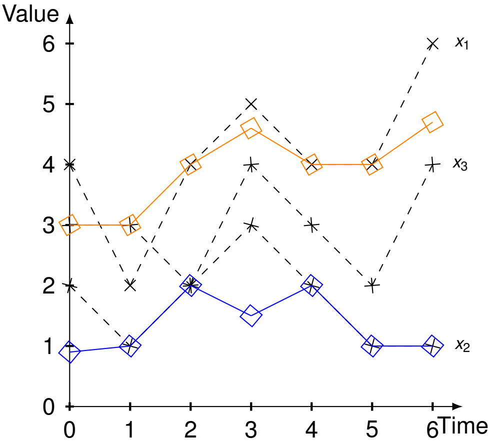


Real variables:

$u_0 = 3$ $l_0 = 0.9$
$u_1 = 3$ $l_1 = 1$
$u_2 = 4$ $l_2 = 2$
$u_3 = 4.6$ $l_3 = 1.5$
$u_4 = 4$ $l_4 = 2$
$u_5 = 4$ $l_5 = 1$
$u_6 = 4.7$ $l_6 = 1$

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0, 1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).

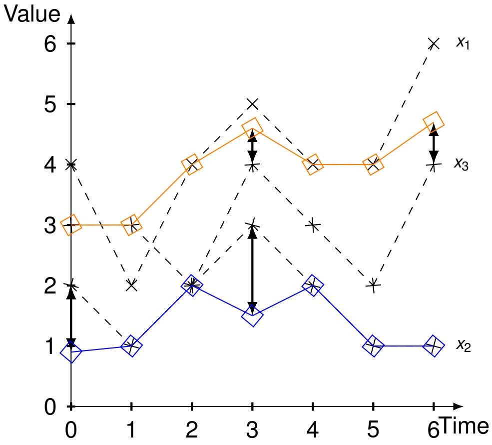


Real variables:

Mixed Integer Programming

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0, 1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).

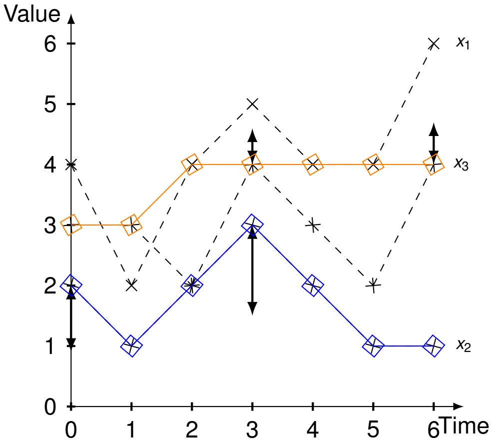


Improving CB without breaking constraints

Size: 17.9

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0, 1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).

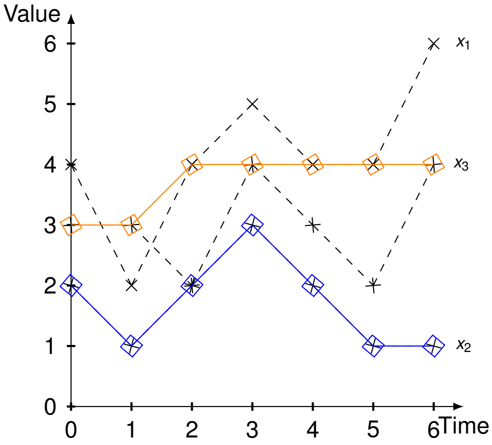


Improving CB without breaking constraints

Size: 14

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0, 1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).

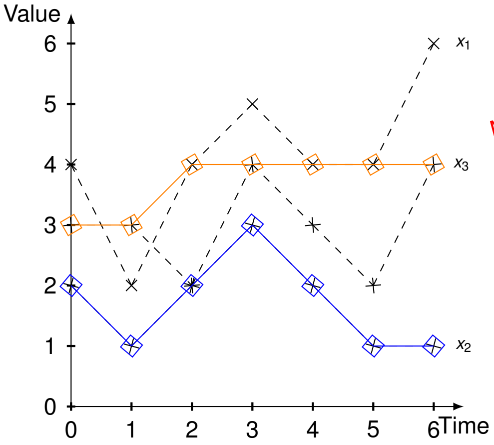


Binary variables:

$u_0^2 = 1$	$l_0^1 = 1$
$u_1^3 = 1$	$l_1^1 = 1$
$u_2^3 = 1$	$l_2^1 = 1$
$u_3^2 = 1$	$l_3^1 = 1$
$u_4^3 = 1$	$l_4^1 = 1$
$u_5^3 = 1$	$l_5^1 = 1$
$u_6^2 = 1$	$l_6^1 = 1$

Variables in constraint models

- Two choices for modeling CB
  1. $u_j, l_j \in \mathbb{R}$
  2. $u_j^k \in \{0, 1\}$, $u_j^k = 1$ iff $u_j = x_{(k)j}$ (k:th smallest value).



Binary variables:

Maximum Satisfiability

Full Model(s) for MWCB(k, s, t)

MaxSAT **Hard Constraints** **MIP**

Full Model(s) for MWCB(k, s, t)

MaxSAT	**Hard Constraints**	**MIP**
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$

Full Model(s) for MWCB(k, s, t)

MaxSAT	**Hard Constraints**	**MIP**
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	At most t points outside CB at each time	$\sum_i d_{ij} \leq t$

Full Model(s) for MWCB(k, s, t)

MaxSAT	**Hard Constraints**	**MIP**
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	T parameter restricts search space	$\sum_i d_{ij} \leq t$

Full Model(s) for MWCB(k, s, t)

MaxSAT	Hard Constraints	MIP
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	At most t points outside CB at each time	$\sum_i d_{ij} \leq t$
$(\sum_j d_{ij} > s) \rightarrow y_i$	Semantics of y_i	$\sum_j d_{ij} - (m - s)y_i \leq s$

Full Model(s) for MWCB(k, s, t)

MaxSAT	**Hard Constraints**	**MIP**
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	At most t points outside CB at each time	$\sum_i d_{ij} \leq t$
$(\sum_j d_{ij} > s) \rightarrow y_i$	Semantics of y_i	$\sum_j d_{ij} - (m - s)y_i \leq s$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=(r+1)}^{t+1} \neg l_j^h$	l_j less than than all points inside CB at j	$l_j - M \cdot d_{ij} \leq x_{ij}$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=r_j^{max}-t}^{(r-1)} \neg u_j^h$	u_j higher than all points inside CB at j	$u_j + M \cdot d_{ij} \geq x_{ij}$

Full Model(s) for MWCB(k, s, t)

MaxSAT	Hard Constraints	MIP
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	At most t points outside CB at each time	$\sum_i d_{ij} \leq t$
$(\sum_j d_{ij} > s) \rightarrow y_i$	Semantics of y_i	$\sum_j d_{ij} - (m - s)y_i \leq s$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=(r+1)}^{t+1} \neg l_j^h$	l_j less than than all points inside CB at j	$l_j - M \cdot d_{ij} \leq x_{ij}$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=r_j^{max}-t}^{(r-1)} \neg u_j^h$	u_j higher than all points inside CB at j	$u_j + M \cdot d_{ij} \geq x_{ij}$
$\sum_r l_j^r = 1$	l_j uniquely defined at j	
$\sum_r u_j^r = 1$	u_j uniquely defined at j	

Full Model(s) for MWCB(k, s, t)

MaxSAT	**Hard Constraints**	**MIP**
$\sum_i y_i \leq k$	At most k outliers	$\sum_i y_i \leq k$
$\sum_i d_{ij} \leq t$	At most t points outside CB at each time	$\sum_i d_{ij} \leq t$
$(\sum_j d_{ij} > s) \rightarrow y_i$	Semantics of y_i	$\sum_j d_{ij} - (m - s)y_i \leq s$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=(r+1)}^{t+1} \neg l_j^h$	l_j less than than all points inside CB at j	$l_j - M \cdot d_{ij} \leq x_{ij}$
$\neg d_{(r)j} \rightarrow \bigwedge_{h=r_j^{max}-t}^{(r-1)} \neg u_j^h$	u_j higher than all points inside CB at j	$u_j + M \cdot d_{ij} \geq x_{ij}$
$\sum_r l_j^r = 1$	l_j uniquely defined at j	
$\sum_r u_j^r = 1$	u_j uniquely defined at j	
Soft: $(\neg l_j^r \vee \neg u_j^h)$	**Minimize size of CB**	MINIMIZE
$weight : x_{(h)j} - x_{(r)j}$		$\sum_j (u_j - l_j)$

Greedy Procedure

input : dataset $X \in \mathbb{R}^{n \times m}$, integers k,s,t
output: $CB = (l, u)$
while CanImprove(X) **do**
    $x_{ij} \leftarrow$getBestRemovalCandidate(X)
    if *canRemove*(x_{ij}) **then**
        $X \leftarrow$remove(x_{ij})
return ConstructCB(X)

Algorithm 1: Simplified Pseudocode of Greedy Procedure

- Similar to Korpela et al. [2014]
- Greedily remove *single data points* while keeping the constraints.
- $\mathcal{O}(mn\log(mn))$ time and $\mathcal{O}(mn)$ memory.

Experimental Evaluation

Set-Up of Experiments

- Compare the scalability of our constraint programming models
- Compare quality of exact solutions compared to greedy procedure.
- Greedy procedure implemented in R
- MIP models solved with CPLEX IBM ILOG [2017]
- MaxSAT models with: QMaxSAT Koshimura et al. [2012], MSCG Morgado et al. [2014], and MaxHS Davies and Bacchus [2013].

Set-Up of Experiments

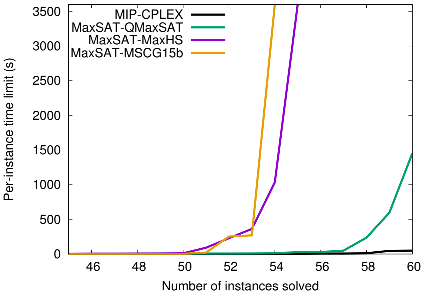
- Compare the scalability of our constraint programming models
- Compare quality of exact solutions compared to greedy procedure.
- Greedy procedure implemented in R
- MIP models solved with CPLEX IBM ILOG [2017]
- MaxSAT models with: QMaxSAT Koshimura et al. [2012], MSCG Morgado et al. [2014], and MaxHS Davies and Bacchus [2013].

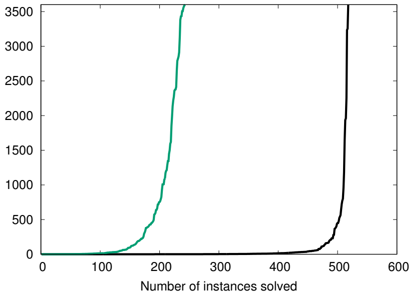
Datasets

- 3 different real-world datasets,
- Milan: Temperature data for a station located in Milan Korpela et al. [2014]; Menne et al. [2012b,a]
- Heartbeat data (MITDB): Heartbeat data from MIT-BIH arrhythmia database Goldberger et al. [2000]
- UCI-Power data (POWER): Individual household electric power consumption, UCI machine learning repository Lichman [2013]

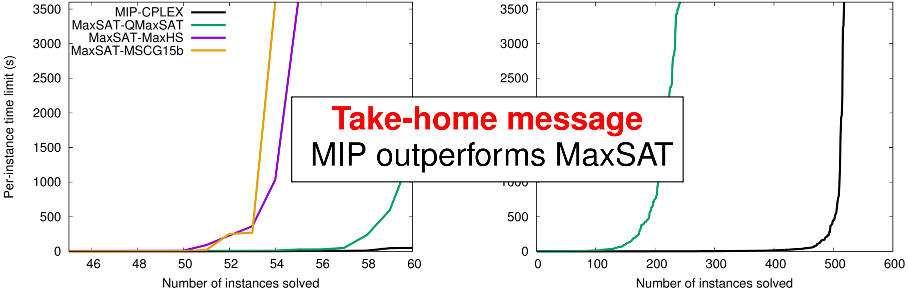
Dataset	sample n	sample m
MILAN (n=245, m=12)	$50, \ldots, 245$	12
POWER (n=1417, m=24)	$200, \ldots, 1000$	24
MITDB (n=2027, m=253)	$100, \ldots, 300$	26, 32, 43, 64, 127, 253

Experimental Evaluation: Solver Scalability





Experimental Evaluation: Solver Scalability



Take-home message

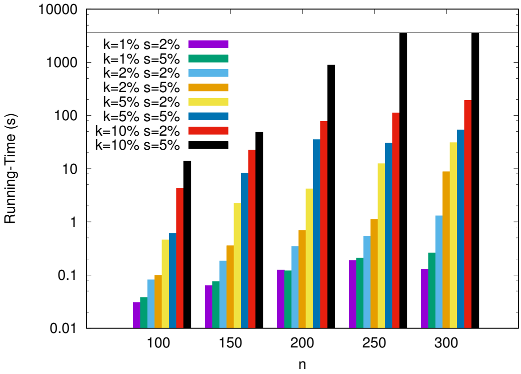
MIP outperforms MaxSAT

Experimental Evaluation: MIP scalability, MITDB m=43

K: max number of outliers (% of n)

S: max times a point can lie outside (% of m)

T: max points outside at any time (=k here)

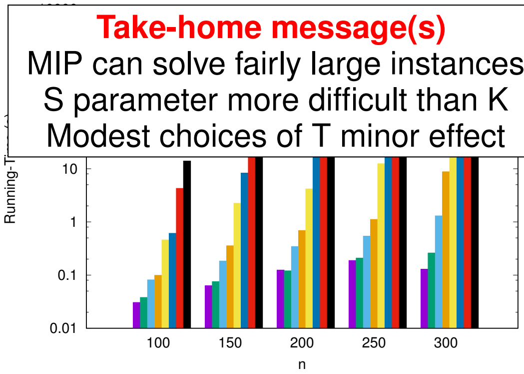


Experimental Evaluation: MIP scalability, MITDB m=43

K: max number of outliers (% of n)

S: max times a point can lie outside (% of m)

T: max points outside at any time (=k here)



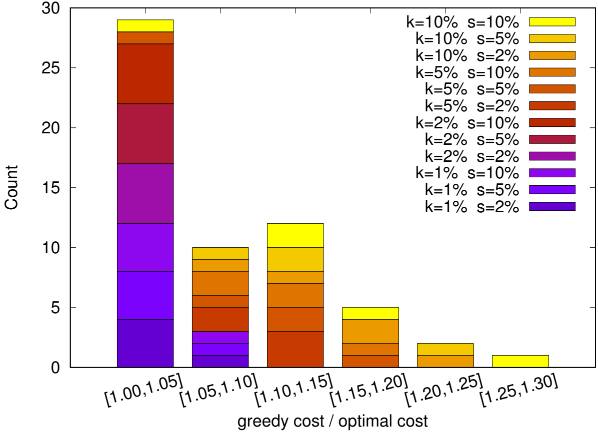
Take-home message(s)

MIP can solve fairly large instances

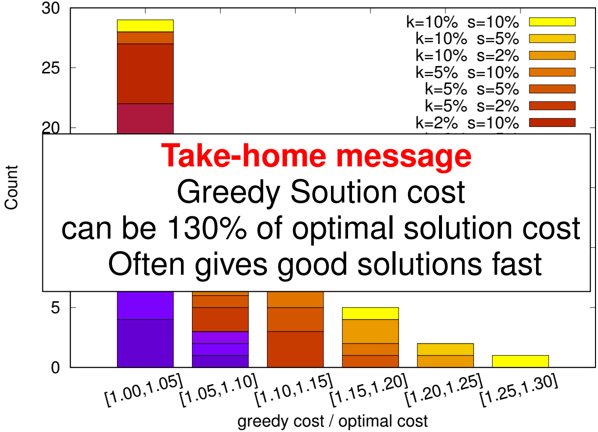
S parameter more difficult than K

Modest choices of T minor effect

Experimental Evaluation: Solution Quality



Experimental Evaluation: Solution Quality



Take-home message

Greedy Soution cost

can be 130% of optimal solution cost

Often gives good solutions fast

Summary

- Minimum-width multivariate confidence intervals, an important problem in Data Analysis
- Proposed MWCB(k, s, t), a generalization addressing limitations of earlier formulations.
- Proposed MIP and MaxSAT models and greedy approximation for solving MWCB(k, s, t).
- Experimental Evaluation:
  - MIP model scales to reasonably large datasets.
  - Greedy Algorithm can be used to get fairly good solutions quickly
- Future work: other problem formulations and constraint optimization paradigms.

Bibliography I

John O. R. Aoga, Tias Guns, and Pierre Schaus. An efficient algorithm for mining frequent sequence with constraint programming. In *Proc. ECML-PKDD*, volume 9852 of *Lecture Notes in Computer Science*, pages 315–330. Springer, 2016.

Roberto Asín, Robert Nieuwenhuis, Albert Oliveras, and Enric Rodríguez-Carbonell. Cardinality networks: a theoretical and empirical study. *Constraints*, 16(2):195–221, 2011.

Behrouz Babaki, Tias Guns, and Siegfried Nijssen. Constrained clustering using column generation. In *Proc. CPAIOR*, volume 8451 of *Lecture Notes in Computer Science*, pages 438–454. Springer, 2014.

Jeremias Berg and Matti Järvisalo. Cost-optimal constrained correlation clustering via weighted partial maximum satisfiability. *Artificial Intelligence*, 244:110–143, 2017.

Christian Bessiere, Frédéric Koriche, Nadjib Lazaar, and Barry O'Sullivan. Constraint acquisition. *Artificial Intelligence*, 244: 315–342, 2017.

A. Biere, M. Heule, H. van Maaren, and T. Walsh. *Handbook of Satisfiability*, volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009.

Abdelhamid Boudane, Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. A SAT-based approach for mining association rules. In *Proc. IJCAI*, pages 2472–2478. IJCAI/AAAI Press, 2016.

J. Davies and F. Bacchus. Exploiting the power of MIP solvers in MaxSAT. In *Proc. SAT*, volume 7962 of *Lecture Notes in Computer Science*, pages 166–181. Springer, 2013.

Martin J Gardner and Douglas G Altman. Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Edition)*, 292(6522):746–750, 1986.

Martin Gebser, Thomas Guyet, René Quiniou, Javier Romero, and Torsten Schaub. Knowledge-based sequence mining with ASP. In *Proc. IJCAI*, pages 1497–1504, 2016.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Tias Guns, Siegfried Nijssen, and Luc De Raedt. Itemset mining: A constraint programming perspective. *Artificial Intelligence*, 175(12-13):1951–1983, 2011.

Bibliography II

Tias Guns, Siegfried Nijssen, and Luc De Raedt. k-pattern set mining under constraints. *IEEE Trans. Knowl. Data Eng.*, 25(2): 402–418, 2013.

Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

Antti Hyttinen, Patrik Hoyer, Frederick Eberhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proc. UAI*, pages 301–310. AUAI Press, 2013.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proc. UAI*, pages 340–349. AUAI Press, 2014.

IBM ILOG. CPLEX optimizer, 2017. `http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/`.

Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. Mining top-k motifs with a SAT-based framework. *Artificial Intelligence*, 244: 30–47, 2017.

Mehdi Khiari, Patrice Boizumault, and Bruno Crémilleux. Constraint programming for mining n-ary patterns. In *Proc. CP*, volume 6308 of *Lecture Notes in Computer Science*, pages 552–567. Springer, 2010.

Dag Kolsrud. Time-simultaneous prediction band for a time series. *Journal of Forecasting*, 26(3):171–188, 2007.

J. Korpela, E. Oikarinen, K. Puolamäki, and A. Ukkonen. Multivariate confidence intervals. In *Proc. SDM*, pages 696–704. SIAM, 2017.

Jussi Korpela, Kai Puolamäki, and Aristides Gionis. Confidence bands for time series data. *Data Mining and Knowledge Discovery*, 28(5-6):1530–1553, 2014.

Miyuki Koshimura, Tong Zhang, Hiroshi Fujita, and Ryuzo Hasegawa. QMaxSAT: A partial Max-SAT solver. *Journal of Satisfiability, Boolean Modeling and Computation*, 8(1/2):95–100, 2012.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

M.J. Menne, I. Durre, B. Korzeniewski, S. McNeal, K. Thomas, X. Yin, S. Anthony, R. Ray, R.S. Vose, B.E. Gleason, and T.G. Houston. Global Historical Climatology Network — Daily (GHCN-Daily), version 3.11, 2012a.

M.J. Menne, I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston. An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910, 2012b. doi: 10.1175/JTECH-D-11-00103.1.

A. Morgado, C. Dodaro, and J. Marques-Silva. Core-guided MaxSAT with soft cardinality constraints. In *Proc. CP*, volume 8656 of *Lecture Notes in Computer Science*, pages 564–573. Springer, 2014.

Benjamin Négrevergne and Tias Guns. Constraint-based sequence mining using constraint programming. In *Proc. CPAIOR*, volume 9075 of *Lecture Notes in Computer Science*, pages 288–305. Springer, 2015.

Regina Nuzzo. Scientific method: Statistical errors. *Nature*, 506:150–152, 2014.

R. Schüssler and M. Trede. Constructing minimum-width confidence bands. *Economics Letters*, 145:182–185, 2016.

Carsten Sinz. *Towards an Optimal CNF Encoding of Boolean Cardinality Constraints*, pages 827–831. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

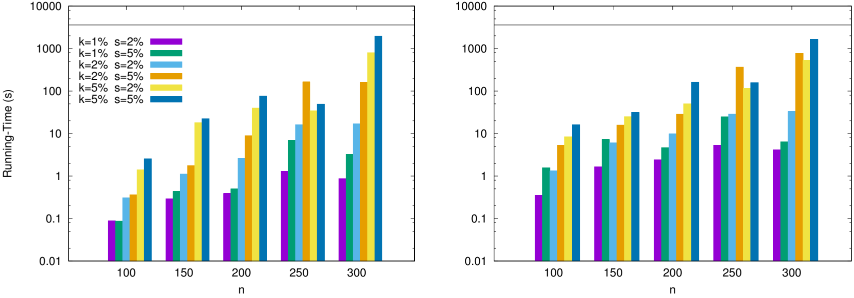
David Trafimow and Michael Marks. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015.

Willy Ugarte, Patrice Boizumault, Samir Loudni, Bruno Crémilleux, and Alban Lepailleur. Soft constraints for pattern mining. *J. Intell. Inf. Syst.*, 44(2):193–221, 2015.

Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems. *Artificial Intelligence*, 244:48–69, 2017.

Chris Woolston. Psychology journal bans P values. *Nature*, 519:9, 2015.

MIP scalability, MITDB m=127



Left: $t = k$, right: $t = k + 2$