Computation of Nonparametric Survival Functions in the Presence of Interval Censoring

Stephen M. Taylor

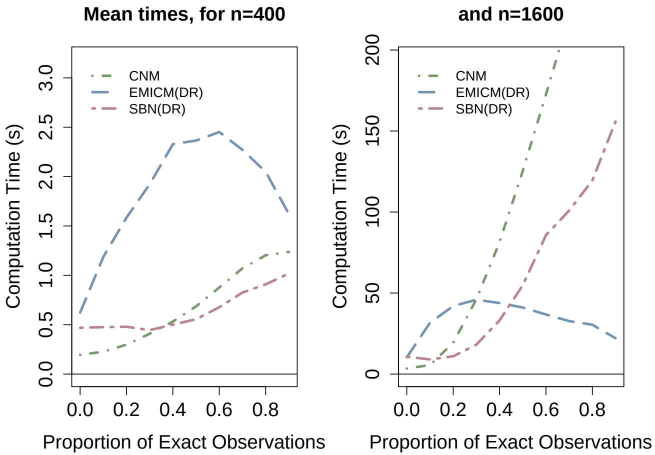
MSc Candidate
Department of Statistics, The University of Auckland
stay020@aucklanduni.ac.nz

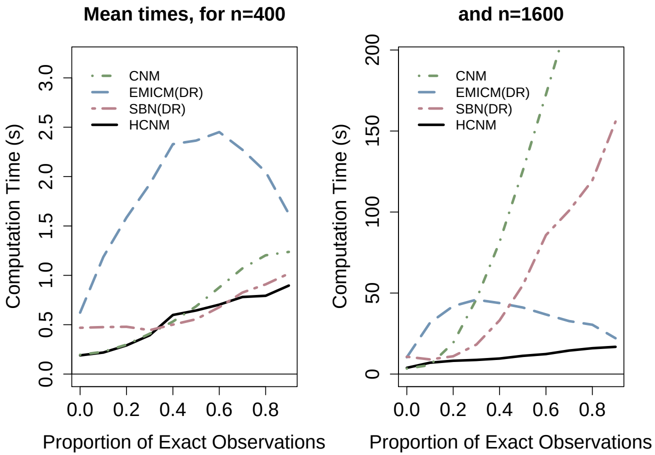
NZ Mathematics and Statistics Postgraduate Conference 2008
20 November 2008

Research Aims

- Create a robust algorithm for solving the NPMLE problem
- One that is fastest in all circumstances

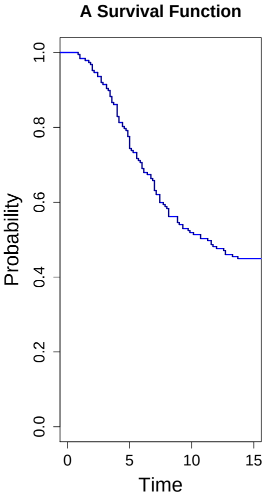
Hierarchical Constrained Newton Method (HCNM)


Mean times, for n=400
and n=1600
CNM
EMICM(DR)
SBN(DR)
Computation Time (s)
Proportion of Exact Observations
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 50 100 150 200
0.0 0.2 0.4 0.6 0.8


Mean times, for n=400
and n=1600
CNM
EMICM(DR)
SBN(DR)
HCNM
Computation Time (s)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
Proportion of Exact Observations

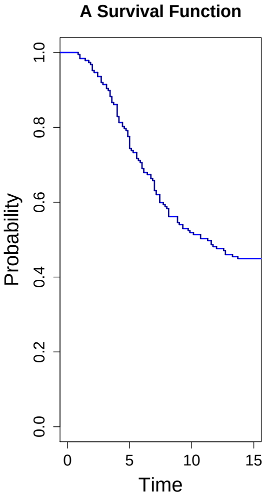
Survival Analysis

- Time to event data



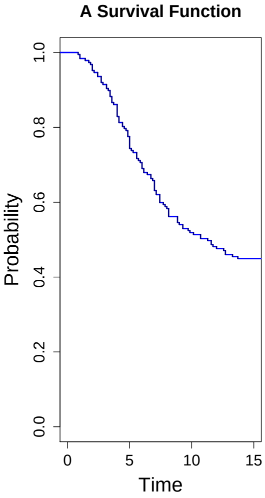
Survival Analysis

- Time to event data
- Want to model the distribution of times to 'failure'



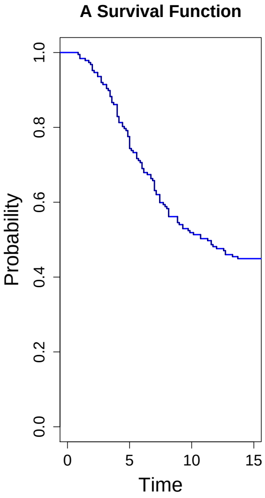
Survival Analysis

- Time to event data
- Want to model the distribution of times to ‘failure’
- Interested in the survival function, $S(t) = P(T > t)$



Survival Analysis

- Time to event data
- Want to model the distribution of times to 'failure'
- Interested in the survival function, $S(t) = P(T > t)$
- Example: Time to healing



Censoring

- Time of event may not be directly measurable

Censoring

- Time of event may not be directly measurable
- Check periodically to see if it has occurred

Censoring

- Time of event may not be directly measurable
- Check periodically to see if it has occurred
- Example: healing occurred some time between doctor visits

Censoring

- Time of event may not be directly measurable
- Check periodically to see if it has occurred
- Example: healing occurred some time between doctor visits
- The event may never occur for some subjects

Censoring

- Time of event may not be directly measurable
- Check periodically to see if it has occurred
- Example: healing occurred some time between doctor visits
- The event may never occur for some subjects
- Example: end of study or “lost to followup”

Interval Censoring

- Event times are not known exactly, only within intervals

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly
- Interval censored: event occurred somewhere in $(t_L, t_R]$

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly
- Interval censored: event occurred somewhere in $(t_L, t_R]$
- Right censored: (t_L, ∞)

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly
- Interval censored: event occurred somewhere in $(t_L, t_R]$
- Right censored: (t_L, ∞)
- Left censored: $(0, t_R]$

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly
- Interval censored: event occurred somewhere in $(t_L, t_R]$
- Right censored: (t_L, ∞)
- Left censored: $(0, t_R]$
- Exact observation: event occurred at time t

Interval Censoring

- Event times are not known exactly, only within intervals
- Perhaps no event time is observed exactly
- Interval censored: event occurred somewhere in $(t_L, t_R]$
- Right censored: (t_L, ∞)
- Left censored: $(0, t_R]$
- Exact observation: event occurred at time t
- Call these intervals O_i for $i = 1, \ldots, n$

Why Nonparametric?

- Let the data speak for itself

Why Nonparametric?

- Let the data speak for itself
- Don't make assumptions about the distribution

Why Nonparametric?

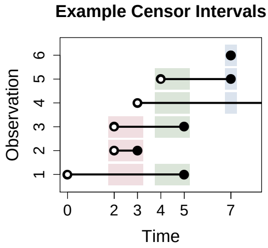
- Let the data speak for itself
- Don't make assumptions about the distribution
- Maximise the likelihood

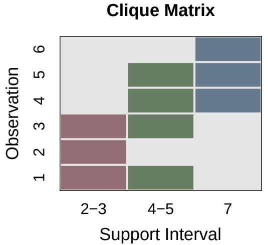
Why Nonparametric?

- Let the data speak for itself
- Don't make assumptions about the distribution
- Maximise the likelihood
- Explore the data before choosing a parametric model

The NPMLE Survival Function with Interval Censored Data

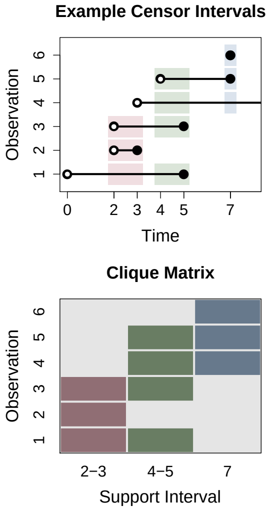
- Partition the positive real line





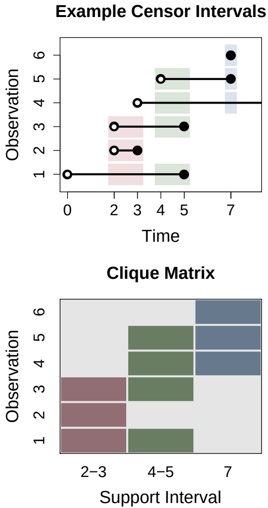
The NPMLE Survival Function with Interval Censored Data

- Partition the positive real line
- All unique values of t_L and t_R



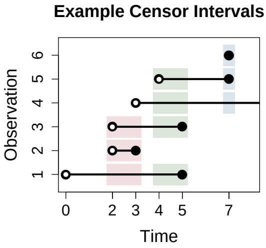
The NPMLE Survival Function with Interval Censored Data

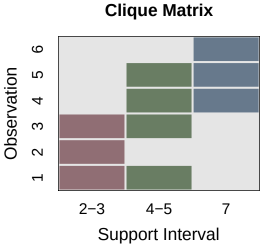
- Partition the positive real line
- All unique values of t_L and t_R
- Potential support intervals



The NPMLE Survival Function with Interval Censored Data

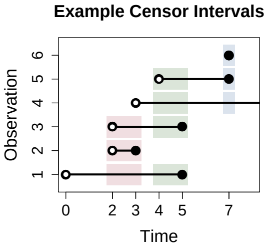
- Partition the positive real line
- All unique values of t_L and t_R
- Potential support intervals
- Only use maximal cliques

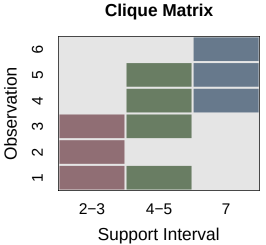




The NPMLE Survival Function with Interval Censored Data

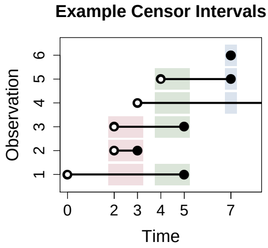
- Partition the positive real line
- All unique values of t_L and t_R
- Potential support intervals
- Only use maximal cliques
- Support set: I_j for $j = 1, \ldots, m$

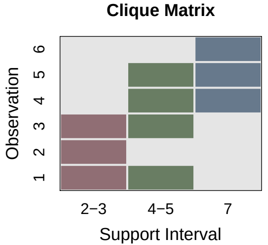




The NPMLE Survival Function with Interval Censored Data

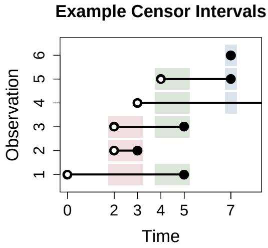
- Partition the positive real line
- All unique values of t_L and t_R
- Potential support intervals
- Only use maximal cliques
- Support set: I_j for $j = 1, \ldots, m$
- The clique matrix $A_{n \times m}$ gives δ_{ij} membership of each O_i in each I_j

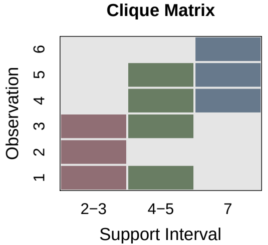




The NPMLE Survival Function with Interval Censored Data

- Partition the positive real line
- All unique values of t_L and t_R
- Potential support intervals
- Only use maximal cliques
- Support set: I_j for $j = 1, \ldots, m$
- The clique matrix $A_{n \times m}$ gives δ_{ij} membership of each O_i in each I_j
- NPMLE assigns probability mass to each support interval





Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$

Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$
- Assign probability p_j to support interval I_j

Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$
- Assign probability p_j to support interval I_j
- Probability of observation O_i using A and $\mathbf{p}$

Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$
- Assign probability p_j to support interval I_j
- Probability of observation O_i using A and $\mathbf{p}$
- Take logs and add them up

Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$
- Assign probability p_j to support interval I_j
- Probability of observation O_i using A and $\mathbf{p}$
- Take logs and add them up
- Goal: find $\hat{\mathbf{p}} \in \mathbb{R}^m$ to maximise $\ell(\hat{\mathbf{p}})$

Likelihood Function for the NPMLE

- Likelihood of an interval $(t_1, t_2]$ is $S(t_1) - S(t_2)$
- Assign probability p_j to support interval I_j
- Probability of observation O_i using A and $\mathbf{p}$
- Take logs and add them up
- Goal: find $\hat{\mathbf{p}} \in \mathbb{R}^m$ to maximise $\ell(\hat{\mathbf{p}})$
- Subject to: $\hat{\mathbf{p}} \geq \mathbf{0}$ and $\hat{\mathbf{p}}^T \mathbf{1} = 1$

Honey as Adjuvant Leg Ulcer Therapy (HALT)

- Randomised Clinical Trial, 368 participants

Honey as Adjuvant Leg Ulcer Therapy (HALT)

- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland

Honey as Adjuvant Leg Ulcer Therapy (HALT)

- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland
- Effect of Manuka Honey dressings for treatment of leg ulcers

Honey as Adjuvant Leg Ulcer Therapy (HALT)

- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland
- Effect of Manuka Honey dressings for treatment of leg ulcers
- Participants assessed weekly and also at a 12-week follow-up

Honey as Adjuvant Leg Ulcer Therapy (HALT)

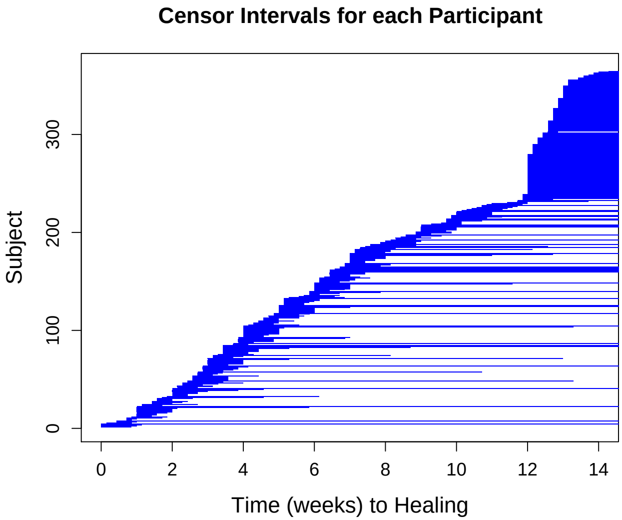
- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland
- Effect of Manuka Honey dressings for treatment of leg ulcers
- Participants assessed weekly and also at a 12-week follow-up
- Nurse changes dressing and assesses healing status

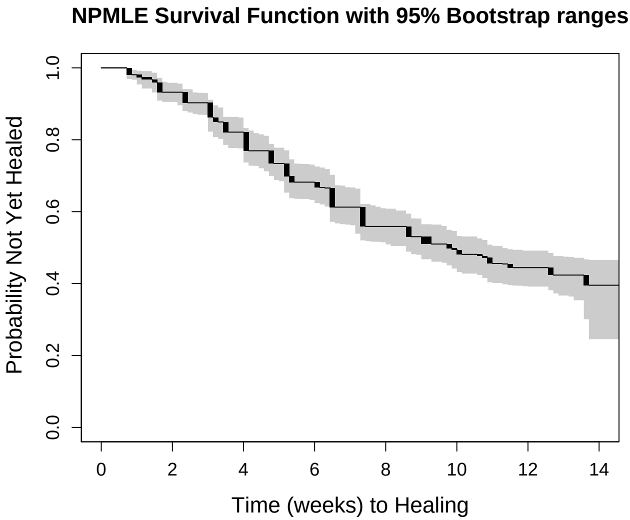
Honey as Adjuvant Leg Ulcer Therapy (HALT)

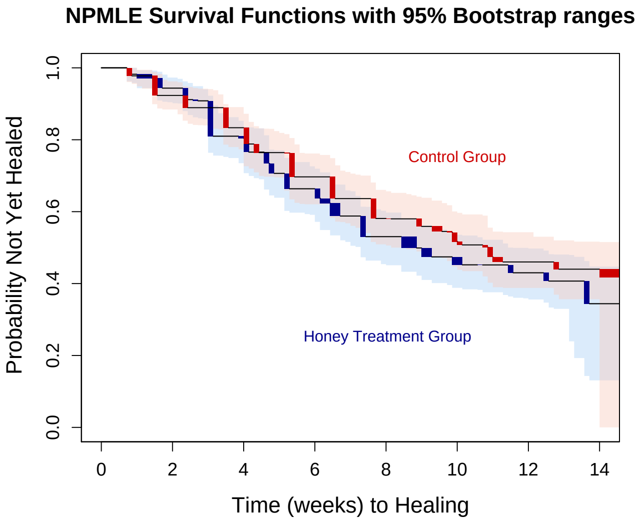
- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland
- Effect of Manuka Honey dressings for treatment of leg ulcers
- Participants assessed weekly and also at a 12-week follow-up
- Nurse changes dressing and assesses healing status
- Event times cannot be observed exactly

Honey as Adjuvant Leg Ulcer Therapy (HALT)

- Randomised Clinical Trial, 368 participants
- Clinical Trials Research Unit in Auckland
- Effect of Manuka Honey dressings for treatment of leg ulcers
- Participants assessed weekly and also at a 12-week follow-up
- Nurse changes dressing and assesses healing status
- Event times cannot be observed exactly
- Thanks to Andrew Jull and Varsha Parag of CTRU for providing the data


Censor Intervals for each Participant
Subject
0
100
200
300
0
2
4
6
8
10
12
14
Time (weeks) to Healing


NPMLE Survival Function with 95% Bootstrap ranges
Probability Not Yet Healed
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
Time (weeks) to Healing


NPMLE Survival Functions with 95% Bootstrap ranges
Probability Not Yet Healed
0.0
0.2
0.4
0.6
0.8
1.0
Control Group
Honey Treatment Group
0
2
4
6
8
10
12
14
Time (weeks) to Healing

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:
    - EM, ISDM, EMICM, VEM and PGM

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:
  - EM, ISDM, EMICM, VEM and PGM
- Subspace-based Newton method (Dümbgen et al. 2006)

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:
  - EM, ISDM, EMICM, VEM and PGM
- Subspace-based Newton method (Dümbgen et al. 2006)
- Wang (2008) introduced:

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:
  - EM, ISDM, EMICM, VEM and PGM
- Subspace-based Newton method (Dümbgen et al. 2006)
- Wang (2008) introduced:
  - Constrained Newton Method

Existing Algorithms for finding the NPMLE

- The Icens package in R provides five algorithms:
  - EM, ISDM, EMICM, VEM and PGM
- Subspace-based Newton method (Dümbgen et al. 2006)
- Wang (2008) introduced:
  - Constrained Newton Method
  - Dimension-reduced approach to improve any algorithm

Times to compute the NPMLE survival function for 100 Bootstrap samples of the HALT data using:

- EMICM, PGM and VEM from the Icens package
- Methods SBN(DR) and EMICM(DR) from Wang (2008)
- The new HCNM algorithm (and CNM)

	Time (s)
EMICM	113.03
PGM	791.00
VEM	610.42
SBN(DR)	14.34
EMICM(DR)	26.93
HCNM	9.41

Problems with Existing Algorithms

- Some are very slow and may fail to converge

Problems with Existing Algorithms

- Some are very slow and may fail to converge
- No algorithm outperforms the others in all situations

Problems with Existing Algorithms

- Some are very slow and may fail to converge
- No algorithm outperforms the others in all situations
- Inefficent use of Hessian matrix or gradient

Problems with Existing Algorithms

- Some are very slow and may fail to converge
- No algorithm outperforms the others in all situations
- Ineffcient use of Hessian matrix or gradient
- Best choice depends on size of dataset and proportion of exact observations

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$
- Makes use of mixture structure of solution

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$
- Makes use of mixture structure of solution
- Uses NNLS to find new estimate of $\mathbf{p}$

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$
- Makes use of mixture structure of solution
- Uses NNLS to find new estimate of $\mathbf{p}$
- Computation time of NNLS is of order $O(nm^2)$

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$
- Makes use of mixture structure of solution
- Uses NNLS to find new estimate of $\mathbf{p}$
- Computation time of NNLS is of order $O(nm^2)$
- Very fast for fully censored datasets

Constrained Newton Method

- Calculates gradient S of $\ell(\mathbf{p})$ at current estimate $\mathbf{p}$
- Makes use of mixture structure of solution
- Uses NNLS to find new estimate of $\mathbf{p}$
- Computation time of NNLS is of order $O(nm^2)$
- Very fast for fully censored datasets
- Can be slow in cases with many exact observations

Hierarchical CNM

- Uses a divide and conquer approach

Hierarchical CNM

- Uses a divide and conquer approach
- Breaks the support set up into blocks

Hierarchical CNM

- Uses a divide and conquer approach
- Breaks the support set up into blocks
- Adapts to the data to make efficient use of Hessian

Hierarchical CNM

- Uses a divide and conquer approach
- Breaks the support set up into blocks
- Adapts to the data to make efficient use of Hessian
- Examines data to choose number/size of blocks

Hierarchical CNM

- Uses a divide and conquer approach
- Breaks the support set up into blocks
- Adapts to the data to make efficient use of Hessian
- Examines data to choose number/size of blocks
- Solves each block using NNLS

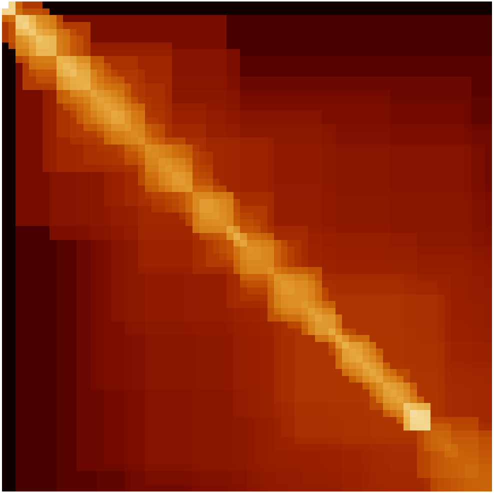
Hierarchical CNM

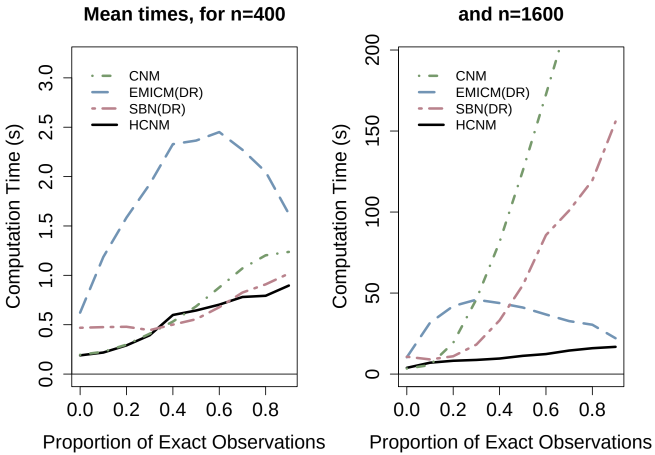
- Uses a divide and conquer approach
- Breaks the support set up into blocks
- Adapts to the data to make efficient use of Hessian
- Examines data to choose number/size of blocks
- Solves each block using NNLS
- Globally reallocates probability among blocks, calling itself recursively

Hierarchical CNM

- Uses a divide and conquer approach
- Breaks the support set up into blocks
- Adapts to the data to make efficient use of Hessian
- Examines data to choose number/size of blocks
- Solves each block using NNLS
- Globally reallocates probability among blocks, calling itself recursively
- Guaranteed convergence to the solution

Heatmap of HALT Hessian




Mean times, for n=400
and n=1600
CNM
EMICM(DR)
SBN(DR)
HCNM
Computation Time (s)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
Proportion of Exact Observations

Conclusions

- Where Interval Censoring is present in survival data, it can be allowed for in the analysis.
- The NPMLE Survival Function combined with Bootstrap methods can create an informative picture of survival progression in such cases.
- The HCNM algorithm provides a fast and robust solution to this problem.

Thanks to:

- My supervisor, Dr Yong Wang
- Andrew Jull and Varsha Parag of CTRU for providing the HALT data