A non-parametric Bayesian change-point method for assessing the risk of novice teenage drivers

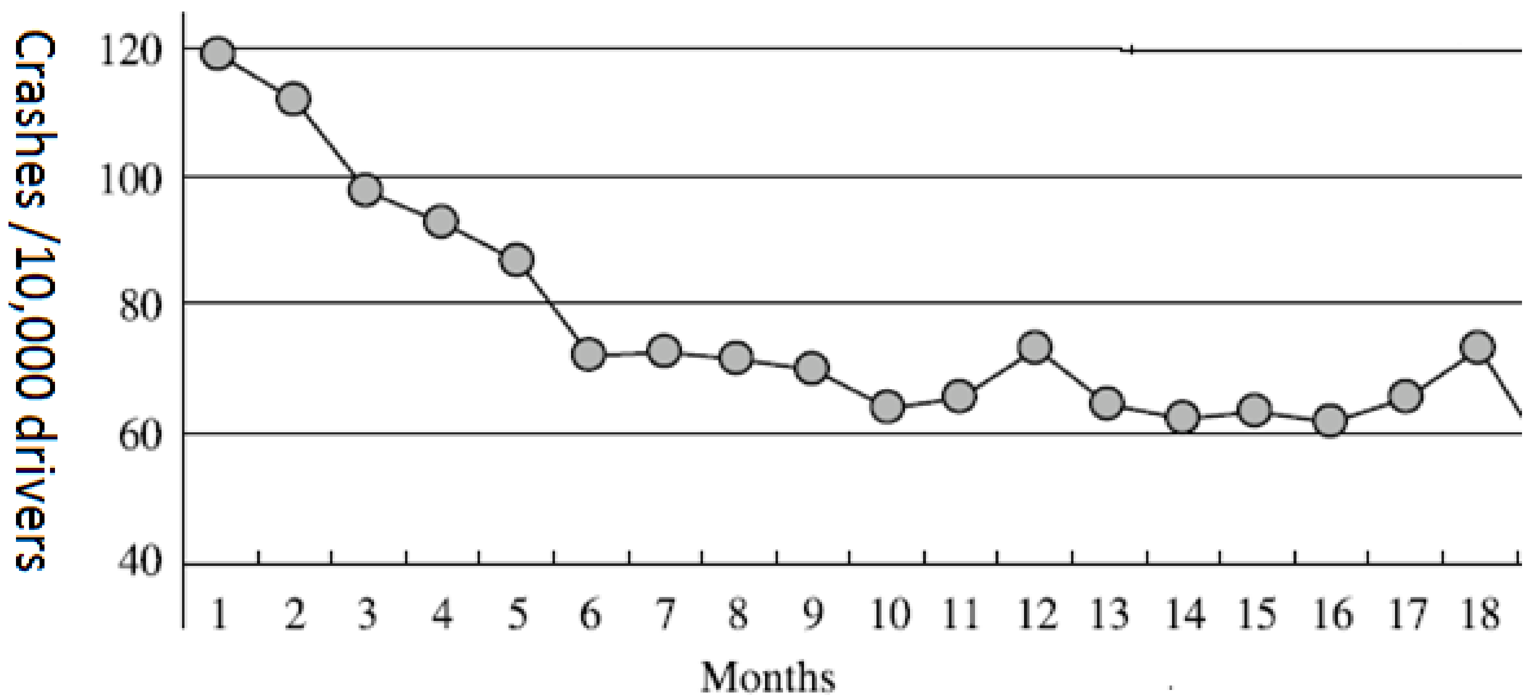
Qing Li
Dept. of IMSE
Iowa State University
Joint work with Feng Guo and Inyoung Kim

Background

- Motor vehicle crashes are the leading cause of mortality for teenagers and young adults.
- Teenage drivers’ risk has been a focus of traffic safety research.
- National Center for Statistics and Analysis: young drivers between 15 and 20 years old had more fatal crashes than other age groups.

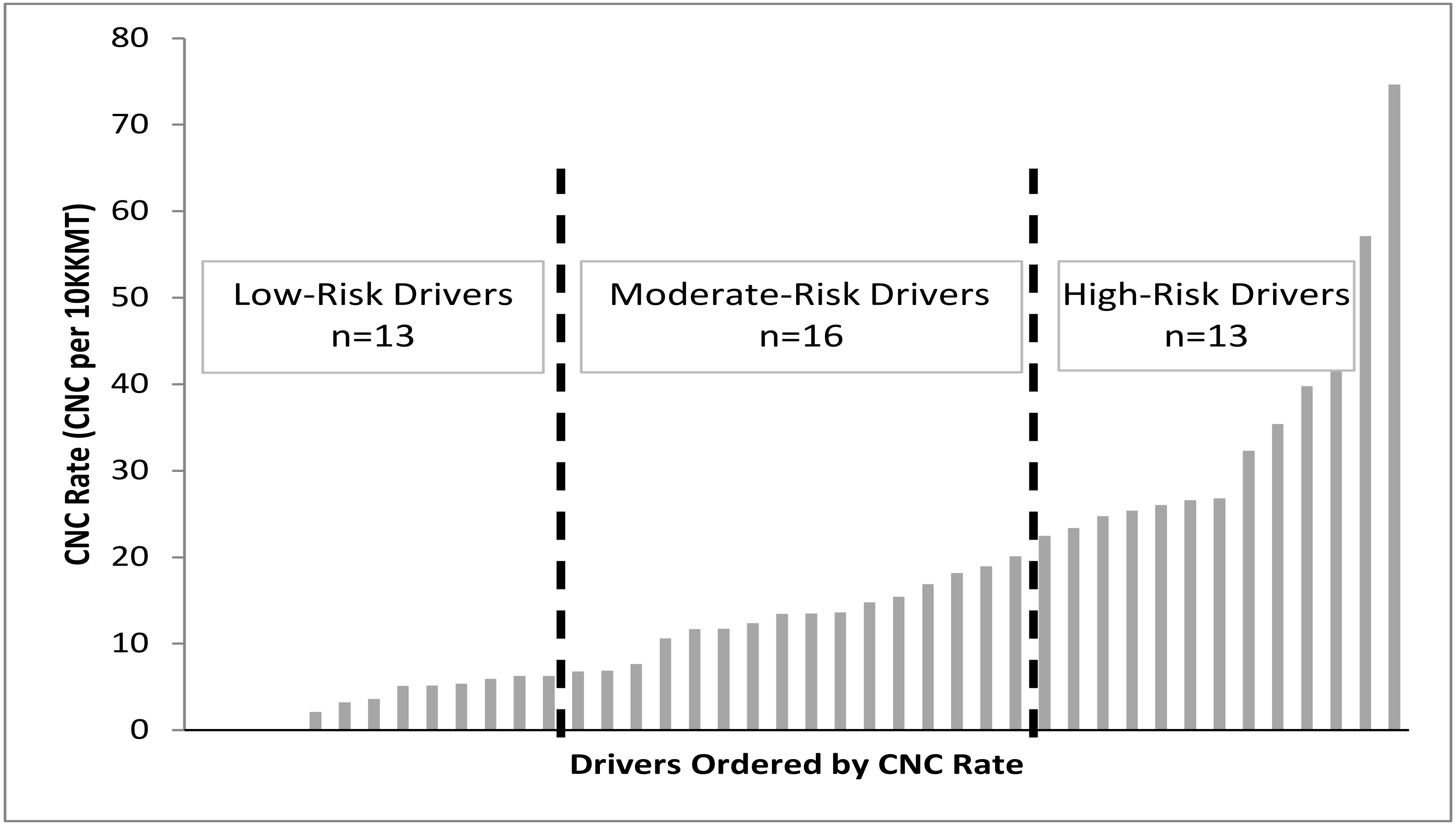
Driving risk of the teenagers

- The initial period after licensure was dangerous.
- Typically followed by a quick decrease in driving risk.
- Teenagers became safer after change.



(Williams 2003)

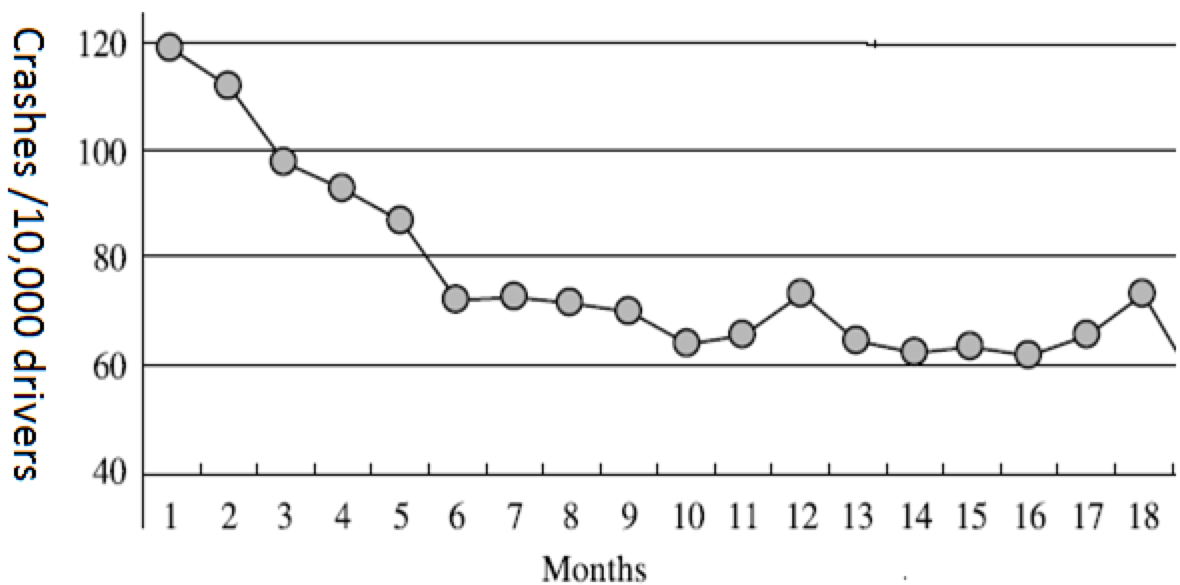
Change-points might be subject-specific: clusters of subjects with different change-points exist.



(Guo et al. 2013)

Motivation

- Detect the change-points of driving risk in terms of cumulative driving hours.
- Maintain high temporal resolution from raw data.
- Allow different change-points and intensity rates among subjects.
- Cluster the drivers.



Why the change-points in driving time matters

- Driving experience is more directly related to the actual driving time than calendar time.
- Novice teenage drivers' safety education program.
- Parent management programs.
- The graduated driver licensing (GDL) regulations.

Data

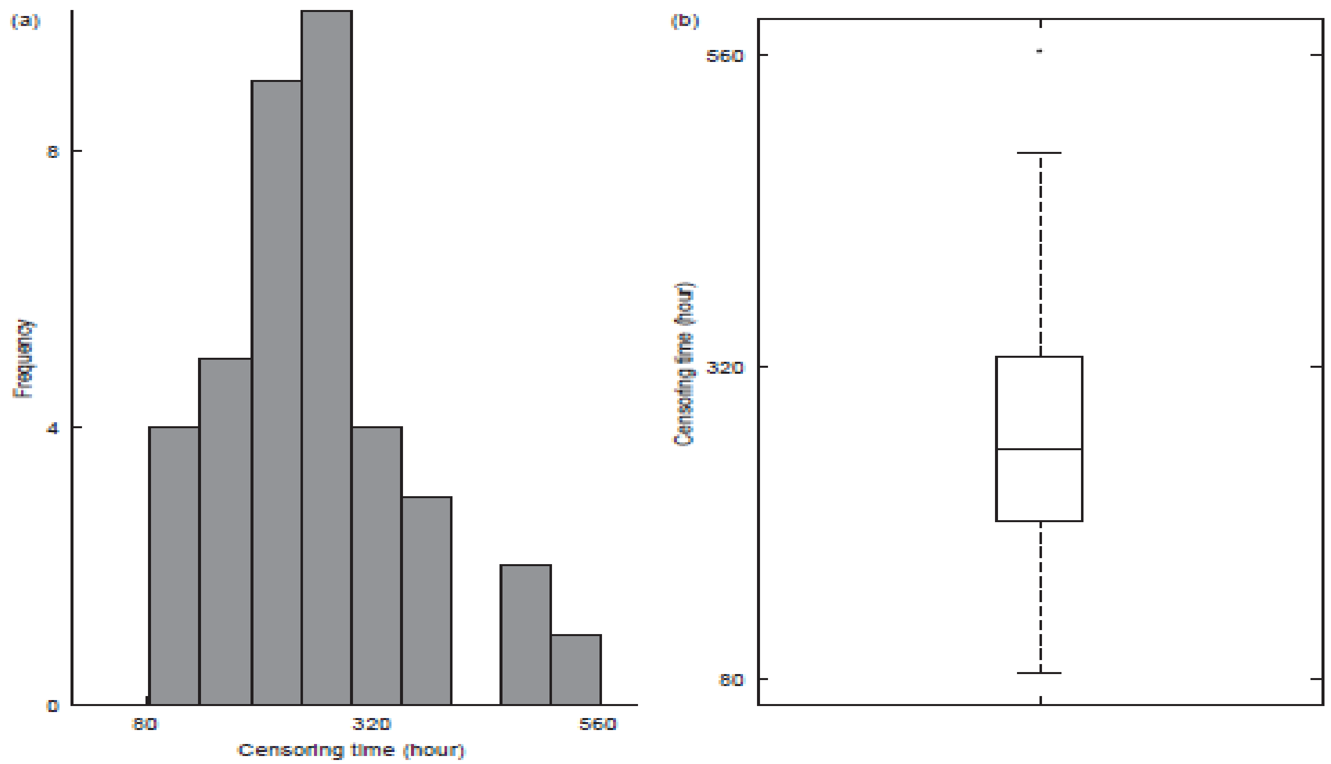
- Naturalistic Teenage Driving Study (NTDS).
- 42 teenagers (22 females, 20 males) just obtained drivers' license from Virginia, 18 months (2006-2009).
- Vehicles equipped with devices to record the driving data continuously. The participants drive as in everyday life without special instructions or the presence of experimenters.
- 279 crash and near-crash (CNC) events (37 crashes).

(a) Quad-image of four continuous video feeds. (b) Quad image with two continuous video feeds (top) and two still frames (bottom).

(Lee et al. 2011)

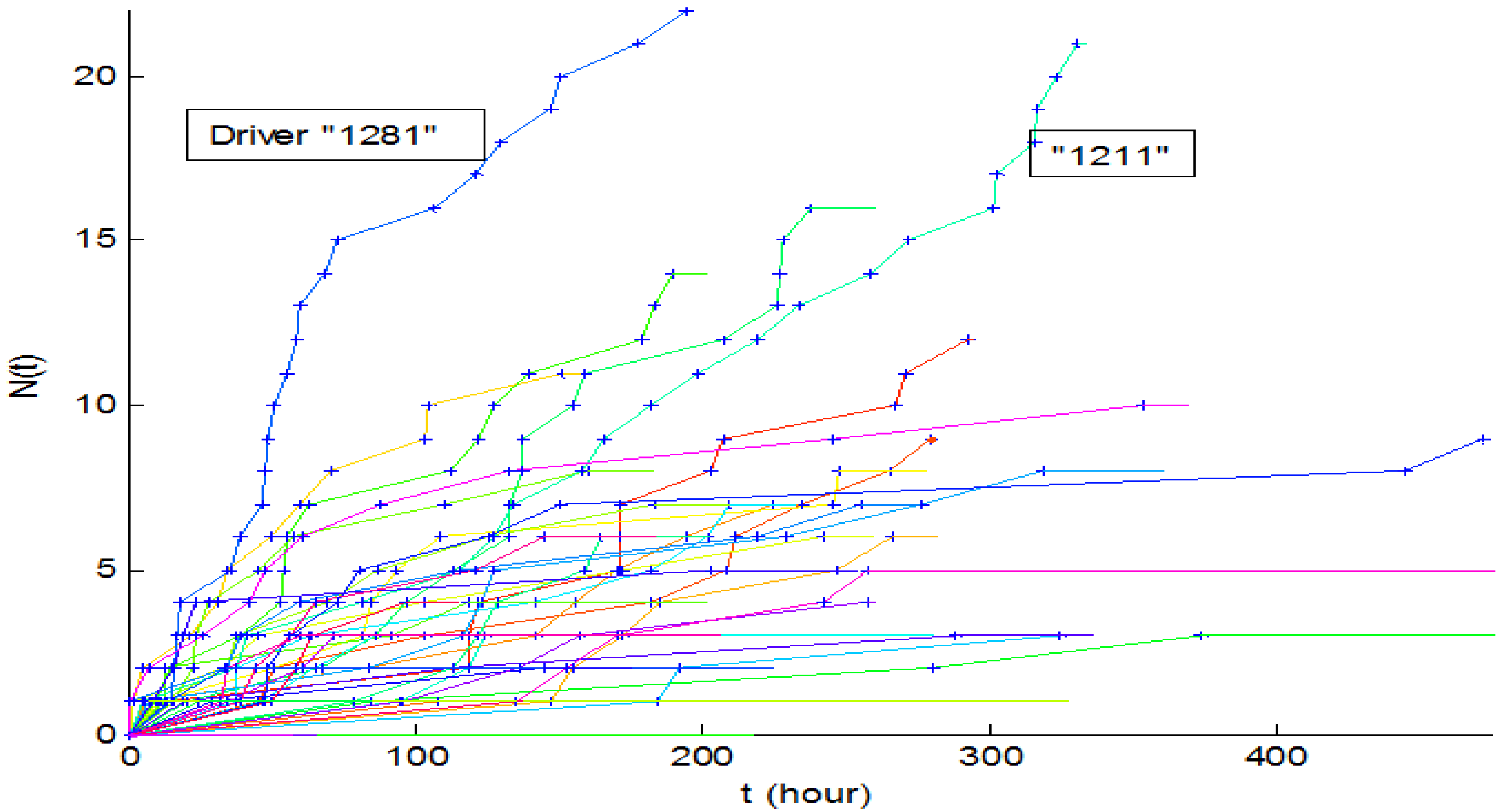
NTDS (Cont.)

Average age: 16.4 years.
Average driving time: 263.07 hours.
SD=102.91 hours.



NTDS cumulative event plot

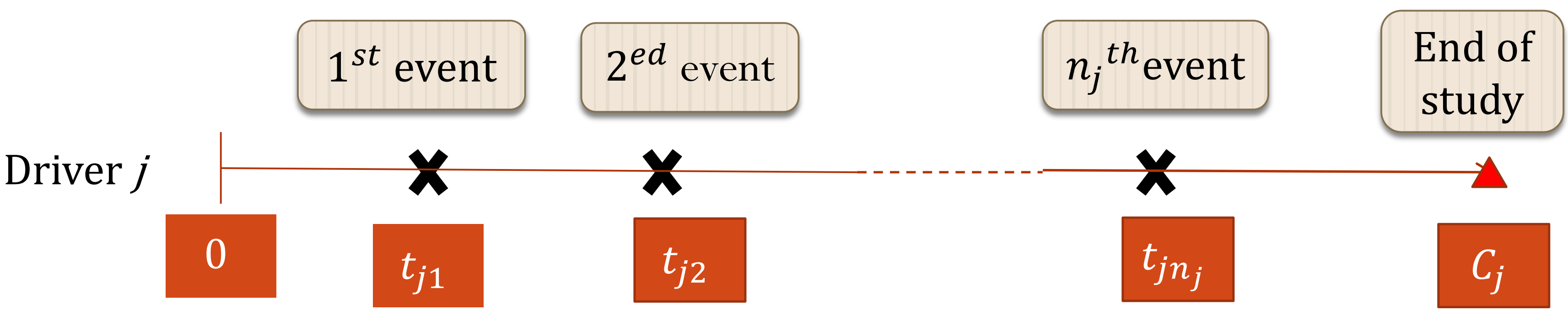
- Event rates vary among drivers.
- Change-points may vary among drivers.



Data setting

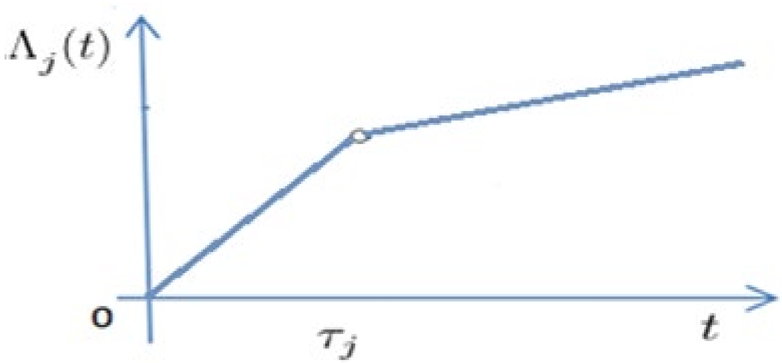
A driver may have multiple crashes and near crashes (CNC): ***Recurrent events***.

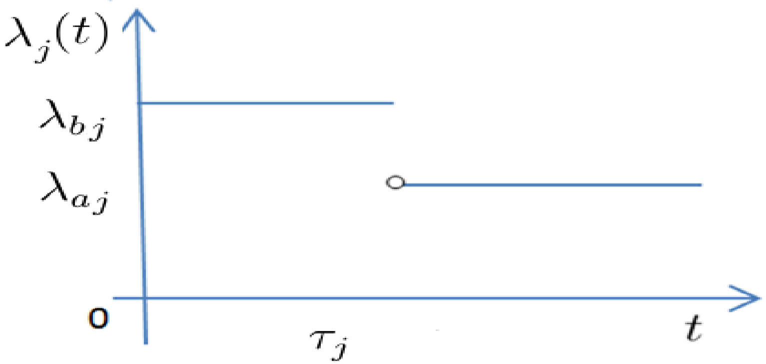
- Driver j = 1, 2, …, m.
- Event i = 1, 2, …, n_j for driver j.
- t_{ji}: time to i^{th} event for the j^{th} driver .
- C_j: end of study for the j^{th} driver .



The Dirichlet Process Mixture Model (DPMM)

- $N_j(t)$ ~ Poisson($\Lambda_j(t)$): number of events over (0,t].
- $N_j(t) = \Lambda_j(t) + M_j(t)$: $M_j(t)$ is a martingale.
- Intensity function: $\lambda_j(t) = \frac{d\Lambda_j(t)}{dt}$.
  - Instantaneous probability of an event occurring at t, conditional on the process history.
  - Nonhomogeneous Poisson Process (NHPP).
- Change-point τ_j: the time point of shift in $\lambda_j(t)$.
- $\tau_1, \tau_2, \ldots, \tau_m \in (0, C_j)$; $\tau_j | G \overset{iid}{\sim} G, G \sim DP(\alpha_0, G_0(\theta))$





DPMM (Cont.)

- Traditional latent class modeling, e.g. the Bayesian finite mixture model (BFMM): model selection to choose the best number of clusters.
- Model selection is full of difficulties.
- An alternative is the Bayesian non-parametric approach: the prior and posterior are stochastic processes.
- Dirichlet Process (DP) is frequently used when the number of clusters is unknown or the number of clusters grows without upper bound as the amount of data increases (Neal 2000).

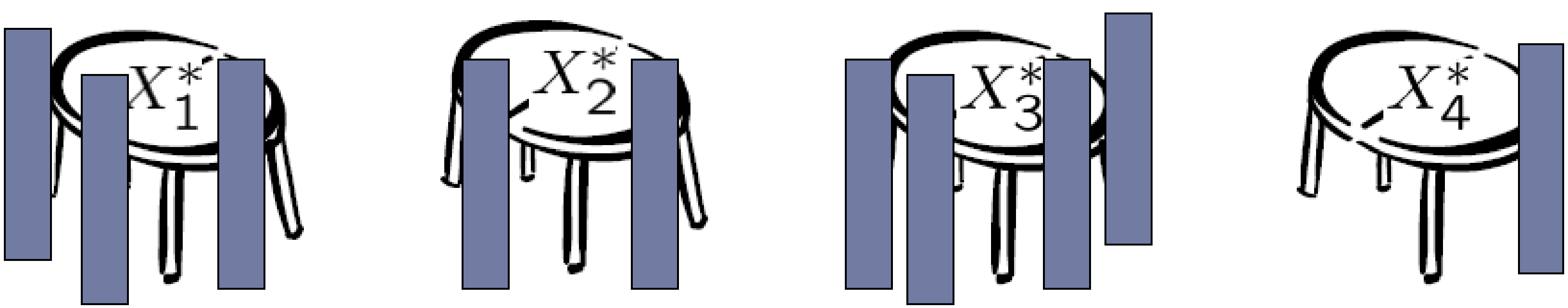
DPMM (Cont.)

- DPMM fits a single model and adapts the model complexity according to the data.
- Automatic clustering is achieved without specifying the number of latent clusters.
- If $G \sim \mathrm{DP}\,(\alpha, G_0)$, G is a discrete distribution with a countably infinite number of point masses.

Chinese Restaurant Process

- Aldous (1985)
- P(customer m sat at table k | allocation of previous $m-1$ custome

$$\begin{cases} \dfrac{\alpha}{\alpha + m - 1}, & \text{if new table,} \\ \dfrac{n_k}{\alpha + m - 1}, & \text{otherwise.} \end{cases}$$



DPMM (Cont.)

- The likelihood for a driver:

$$L_j(\lambda_{bj}, \lambda_{aj}, \tau_j|\boldsymbol{X}_j) = exp[-\Lambda(C_j)]\prod_{i=1}^{n_j}\lambda_j(t_{ji})$$

$$= exp\left\{-(\lambda_{bj} - \lambda_{aj})\tau_j - \lambda_{aj}C_j\right\}\lambda_{bj}^{N_j^{(1)}}\lambda_{aj}^{N_j^{(2)}}$$

- The joint posterior distribution:

$$f(\lambda_b, \lambda_a, \boldsymbol{\tau}|\boldsymbol{X}) \propto L(\lambda_b, \lambda_a, \boldsymbol{\tau}|\boldsymbol{X})f(\lambda_b)f(\lambda_a)f(\boldsymbol{\tau})$$

Full conditional distributions for MCMC

$$f(\tau_j|\boldsymbol{\tau}_{-j},\lambda_{bj},\lambda_{aj},\alpha_0,\theta,\boldsymbol{X}_j) = b\alpha_0 q_0 H(\tau_j|\lambda_{bj},\lambda_{aj},\boldsymbol{X}_j) + b\sum_{p\neq j} L_j(\tau_p,\lambda_{bj},\lambda_{aj},\boldsymbol{X})\delta(\tau_j,\tau_p)$$

$$f(\lambda_{bj}|\tau,\lambda_{aj},\boldsymbol{X}_j) \sim Gamma(a_1 + N_j^{(1)}, b_1 + \tau_j)$$

$$f(\lambda_{aj}|\tau_j,\lambda_{bj},\boldsymbol{X}_j) \sim Gamma(a_2 + N_j^{(2)}, b_2 + (C_j - \tau_j))$$

$$(\eta|\alpha_0,k) \sim Beta(\alpha_0 + 1, m)$$

$$(\alpha_0|\eta,k) \sim \pi_\eta Gamma(a_0+k, b_0-log(\eta))+(1-\pi_\eta)Gamma(a_0+k-1, b_0-log(\eta))$$

Inference

- The number of clusters: the posterior mode of the number of unique values after 'Burn-in'.
- The similarity measure for clustering: the estimates of the posterior pairwise probabilities for two drivers to be in the same cluster,

$$P_{jp} = \frac{\text{No. of iterations after `burn-in' for which } \tau_j = \tau_p}{B_t - B}$$

- Complete-linkage clustering is then conducted using $D_{jp} = 1 - P_{jp}$ as the distance measure (Medvedovic et al. 2002).

Simulation

- Check the model performance under different simulation settings.
- Compare the DPMM with the BFMM.
- Twelve settings with different: change-points, intensity rates, mixture proportions, sample sizes, number of clusters, censoring time.

Data generation

- Data generation from a NHPP with piecewise-constant intensity functions is based on the inter-event times' distribution.
- Given the previous event times, the i^{th} inter-event time X_i for each subject has the cumulative density function (CDF):

$$
\begin{aligned}
F_i(x) &= Pr\left[X_i \le x | Y_p = y_p,\ p = 1, 2, \cdots, i-1\right] \\
&= 1 - exp\left[\Lambda(y_{i-1}) - \Lambda(y_{i-1} + x)\right],
\end{aligned}
$$

Simulation(Cont.)

- P_1 (%): the percentage of correctly estimated number of clusters out of 200 data sets.
- P_2 (%): the average percentage of correctly grouped subjects given the correct number of clusters.

Setting	1	2	3	4	5	6
DPMM P_1	95.5	94.0	95.0	89.0	95.5	92.5
BFMM P_1	69.5	71.0	67.0	48.0	66.0	69.5
DPMM P_2	92.21	92.35	93.90	56.26	89.86	92.33
BFMM P_2	91.20	90.34	88.51	75.56	86.23	91.30
Setting	7	8	9	10	11	12
DPMM P_1	93.5	99.0	90.0	84.0	91.5	99.5
BFMM P_1	65.0	64.5	47.5	50.0	-	60.0
DPMM P_2	93.10	85.17	66.17	36.06	92.18	95.70
BFMM P_2	89.37	85.66	57.88	38.72	-	98.50

Simulation (DPMM) given the correct number of clusters: m = 40, B = 200

Setting	Parameter	True value	Average of estimates	RMSE	|Bias (%)|	Coverage probability (%)
1	μ_1	150	163.57	14.57	9.04	82.5
	μ_2	300	288.45	12.89	3.85	89.5
	λ_b	250	248.87	9.51	0.45	100.0
	λ_a	100	100.56	5.90	0.56	100.0
12	μ_1	150	160.38	11.51	6.92	84.0
	μ_2	300	290.86	10.18	3.05	88.0
	λ_{1b}	250	242.64	14.88	2.95	100.0
	λ_{1a}	100	102.77	6.51	2.77	100.0
	λ_{2b}	100	102.84	6.78	2.84	100.0
	λ_{2a}	250	241.46	13.72	3.42	100.0

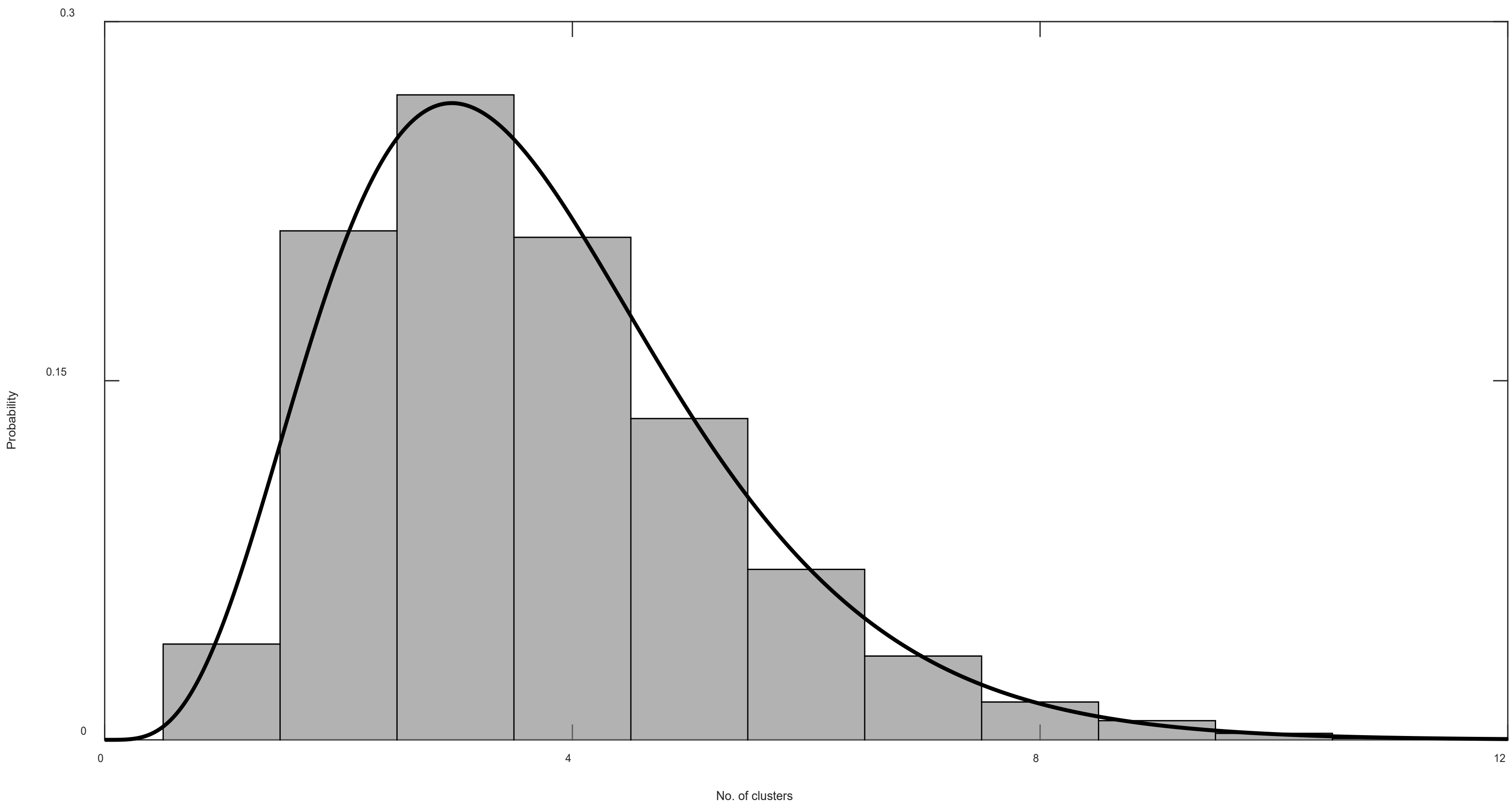
$$RMSE = \sqrt{(1/B)\sum_{k=1}^{B}(\hat{\tau}-\tau)^2},\ |Bias(\%)| = (1/B)|\sum_{k=1}^{B}(\hat{\tau}-\tau)|/\tau \times 100\%$$

Simulation Summary

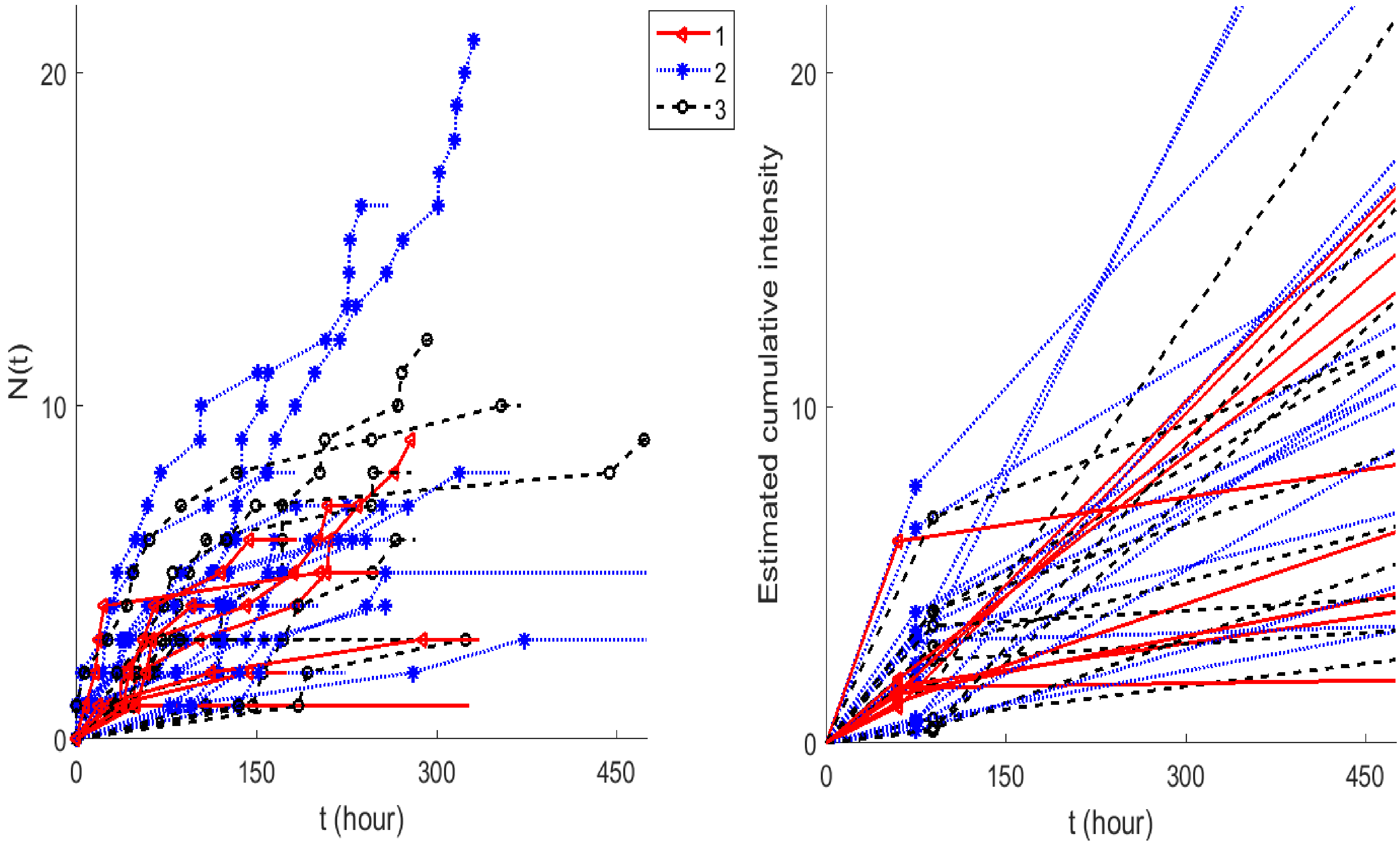
- When the change-points and intensity rates are dispersed, the estimates will be more accurate.
- The estimates will be less stable for the clusters with smaller mixture proportion.
- When the sample size is larger, the estimation will be more accurate.
- Both models are robust to different parameter settings in estimation.
- Both are not sensitive to initial values and priors.
- The DPMM outperforms BFMM in detecting the number of clusters and grouping the subjects given the correct number of clusters.
- The automatic clustering property of DPMM yields higher computational efficiency than BFMM.

Application to the NTDS (DPMM)

- The model with three clusters is chosen.



Application to the NTDS (DPMM)



The posterior summaries of the change-points

Change-points	Mean	SD	25%	Median	75%	Size
μ_1	60.42	33.38	35.25	44.26	76.67	9
μ_2	74.94	34.83	37.52	73.14	106.53	16
μ_3	89.23	32.70	69.76	90.72	115.01	11

The average cumulative driving time at the end of several months per teenager for NTDS.

Month	4	5	6	7
Average driving time (hour)	55.07	68.11	82.29	95.12

The intensity rates

Parameter	Average	SD	25%	Median	75%	NTDS average
λ_{1b}	33.28	38.92	7.33	21.24	44.60	31.26
λ_{1a}	18.11	21.15	1.79	10.16	29.58	14.52
λ_{2b}	36.78	38.50	7.57	26.11	52.97	30.02
λ_{2a}	27.10	24.78	8.82	19.39	38.58	25.73
λ_{3b}	27.65	29.78	3.20	18.95	42.74	26.49
λ_{3a}	18.36	26.34	2.34	10.83	24.79	16.17
λ_{b}	33.11	36.38	5.97	22.59	47.63	29.00
λ_{a}	22.18	24.82	4.60	14.60	32.05	20.38

Summary

- Implement clustering and change-point detection in the recurrent-event context.
- Detect driving risk change-points in terms of cumulative driving hours.
- Bayesian hierarchical modeling is flexible.
- The DPMM outperforms the BFMM mainly in detecting the number of clusters automatically, grouping the individuals, and in efficiency.
- The DPMM is accurate, robust, and flexible.

Future Work

- Multiple change-points
- Other forms of intensity functions
- Incorporate covariates (gender, personality, cortisol level)
- Testing methods
- In big-data
- In natural disasters or global warming with spatial effects

Reference

- Aldous, D. J. (1985). Exchangeability and Related Topics. Springer.
- Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via pólya urn schemes. The Annals of Statistics, pages 353-355.
- Guo, F., Simons-Morton, B. G., Klauer, S. E., Ouimet, M. C., Dingus, T. A., and Lee, S. E. (2013). Variability in crash and near-crash risk among novice teenage drivers: a naturalistic study. The Journal of Pediatrics, 163:1670-6.
- Lee, S., Simons-Morton, B., Klauer, S., Ouimet, M., and Dingus, T. (2011). Naturalistic assessment of novice teenage crash experience. Accident Analysis & Prevention, 43:1472-9.
- Li, Q., Guo, F., Kim, I., Klauer, S., and Simons-Monton, B. (2017). A Bayesian finite mixture change-point model for assessing the risk of novice teenage drivers. Journal of Applied Statistics. 45:604-625.
- Li, Q., Guo, F., and Inyoung, K. (2018). A non-parametric Bayesian change-point detection method in the recurrent-event context. The Annals of Applied Statistics. Submitted.
- Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, 18:1194-206.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet Process mixture models. Journal of Computational and Graphical Statistics, 9:249-65.

Contact & Thanks

Qing Li

- qlijane@iastate.edu
- https://www.imse.iastate.edu/qing-li/