# Inference based on Mixture of Weibull Distributions for Competing Risks Data with Cure Rate

Ayon Ganguly
Department of Mathematics
Indian Institute of Technology Guwahati
Email: aganguly@iitg.ac.in

16th International OSD Conference on Order in Statistical Data: Order Statistics and Beyond
RWTH Aachen University

June, 2025

# Co-Authors

This is a joint work with

- Dr. Farha Sultana (IIIT Guwahati, India)
- Prof. Debasis Kundu (IIT Kanpur, India)
- Dr. Ayan Pal (The University of Burdwan, India)

# Main Sections

1. Motivating Data
2. Proposed Model
3. Likelihood Inference
4. Analysis of Melanoma Data

# Malignant Melanoma Cancer Data Set

- Melanoma : Skin cancer
- Collected at Odense University Hospital during 1962 to 1977
- Two hundred and five patients
- No of days after the operation
  - until they died
  - until they left the study
  - until the termination of the study in the year 1977

# Malignant Melanoma Cancer Data Set

| no | status | days | ulc | thick | sex |
|---|---|---|---|---|---|
| 789 | 3 | 10 | 1 | 676 | 1 |
| 13 | 3 | 30 | 0 | 65 | 1 |
| 97 | 2 | 35 | 0 | 134 | 1 |
| 16 | 3 | 99 | 0 | 290 | 0 |
| 21 | 1 | 185 | 1 | 1208 | 1 |
| 469 | 1 | 204 | 1 | 484 | 1 |

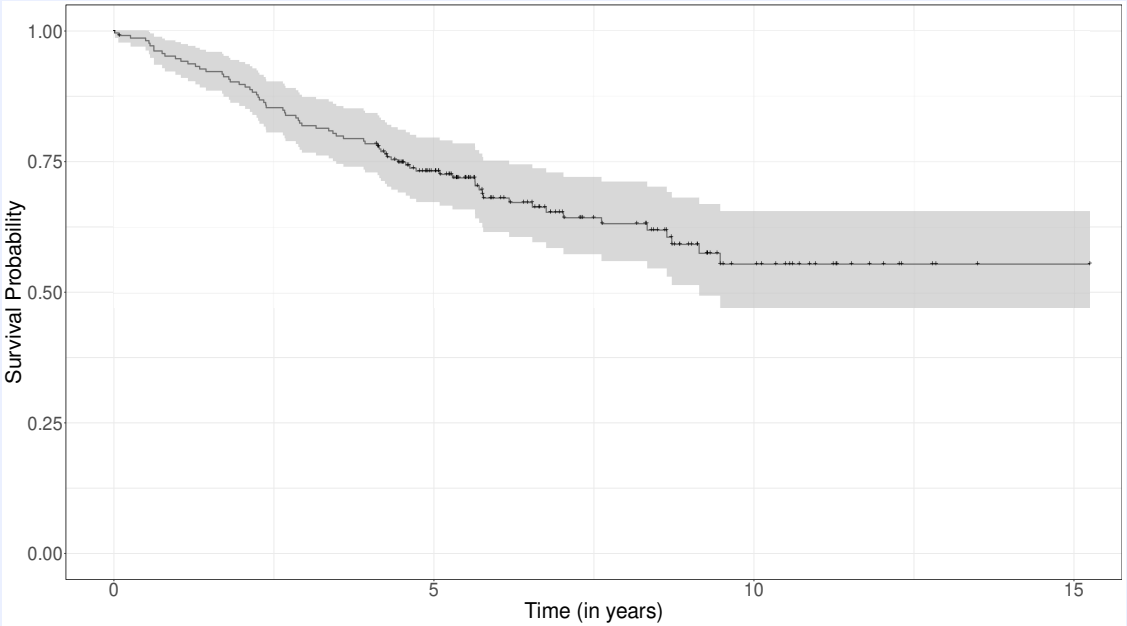
- no : Patient code
- status : Survival status (1: dead from melanoma, 2: alive, 3: dead from other cause)
- days : Survival time (in days)
- ulc : Ulceration (1: present, 0: absent)
- thick : Tumour thickness (in 1/100 mm)
- sex : Gender (0: female, 1: male)

# Malignant Melanoma Cancer Data Set

- Melanoma cancer: 28%
- Other causes: 7%
- Right censored: 65%
- Sex (Male 39%, Female 61%)
- Status of ulcer (Present 44%, Absent 56%)
- Thickness of tumor (Mean 2.92 mm, SD 2.92 mm)
- Available in `timereg` package

# Malignant Melanoma Cancer Data Set

- This data set was earlier analyzed by Rodrigues et al. 2011 and Pal and Balakrishnan 2016.
- Patients died due to other causes: Right censored
- The Kaplan-Meier estimate




- Levels off at a non-zero proportion
- Possibility of non-zero cure proportion

# Proposed Model

## Assumption

- The population consists of two groups
  - Susceptible
  - Cure
- There are $K$ mutually exclusive and exhaustive causes for the event

# Proposed Model

Mixture model for competing risks

- $\widetilde{T}$ : Time-to-event
- $\widetilde{I}$ : Indicator (0 : cured, 1 : susceptible)
- The conditional survival function (SF) of the time-to-event, given the subject belongs to susceptible group, is

$$S_{mix}(t) = P(\widetilde{T} > t | \widetilde{I} = 1) = \sum_{k=1}^{K} \pi_k S_k(t)$$

- $K$ : Number of competing risks
- $\widetilde{N}$ : Failure mode of a susceptible subject
- $\pi_k = P(\widetilde{N} = k | \widetilde{I} = 1)$ : Probability of failure due to $k$-th cause
- $S_k(t) = P(\widetilde{T} > t | \widetilde{N} = k, \widetilde{I} = 1)$ : Mode-specific conditional SF

# Proposed Model

## Mixture cure rate model

- The SF of $\widetilde{T}$ is

$$S_{pop}(t) = p_0 + (1 - p_0)S_{mix}(t) = p_0 + (1 - p_0)\sum_{k=1}^{K} \pi_k S_k(t)$$

- $p_0 = P(\widetilde{I} = 0)$ : Cure probability

# Proposed Model

## Modelling $S_k(\cdot)$

- $k-$th mode-specific time-to-event follows Weibull distribution
- $S_k(t) = \exp\{-\lambda_k t^{\alpha_k}\}$ for $t > 0$
- $\lambda_k > 0$ : Scale parameter
- $\alpha_k > 0$ : Shape parameter

# Proposed Model

## Model with covariates

- Probability of a subject being cured depends on covariates
- Binary regression models
- Logistic link function

$$p_0(\boldsymbol{x};\, \boldsymbol{\beta}) = \frac{1}{1 + \exp(\boldsymbol{x}^T \boldsymbol{\beta})}$$

- $\boldsymbol{x} = (1,\, x_1,\, \ldots,\, x_L)$ : Vector of covariates
- $\boldsymbol{\beta} = (\beta_0,\, \beta_1,\, \ldots,\, \beta_L)$ : Parameter vector

# Likelihood Inference

## Form of data

- Form of available data :

$$\{(t_\ell,\, \delta_\ell,\, \boldsymbol{x}_\ell) : \ell = 1,\, 2,\, \ldots,\, n\}$$

- $t_\ell$ : Observed value of $T_\ell = \min(\widetilde{T}_\ell, C_\ell)$
- $\delta_\ell$ : Observed value of $\Delta_\ell = \widetilde{N} \times I(\widetilde{T}_\ell \leq C_\ell)$
- $\widetilde{T}$ : Time-to-event
- $C_\ell$ : Censoring time
- $\boldsymbol{x}_\ell$ : Covariate vector

# Likelihood Inference

## Likelihood function

- Likelihood function (non-informative censoring)

$$L(\boldsymbol{\gamma}) \propto \prod_{\ell=1}^{n} \left[ \left\{ \prod_{k=1}^{K} \{(1 - p_0(\boldsymbol{x}_\ell; \boldsymbol{\beta}))\pi_k f_k(t_\ell; \boldsymbol{\theta}_k)\}^{I(\delta_\ell = k)} \right\} \{S_{pop}(t_\ell; \boldsymbol{\gamma})\}^{I(\delta_\ell = 0)} \right]$$

- $\boldsymbol{\gamma} = (\alpha_1, \ldots, \alpha_K, \lambda_1, \ldots, \lambda_K, \pi_1, \ldots, \pi_K, \beta_0, \beta_1, \ldots, \beta_L)$
- $\boldsymbol{\theta}_k = (\alpha_k, \beta_k)$

# Likelihood Inference

## Likelihood function

- MLEs can be found by maximizing likelihood function over

$$\boldsymbol{\Gamma} = \Bigg\{ \boldsymbol{\gamma} : \alpha_k > 0,\ \lambda_k > 0,\ k = 1,\ 2,\ \ldots,\ K,$$

$$0 < \pi_k < 1,\ k = 1,\ 2,\ \ldots,\ K,\ \sum_{k=1}^{K} \pi_k = 1,$$

$$\beta_\ell \in \mathbb{R},\ \ell = 0,\ 1,\ 2,\ \ldots,\ L \Bigg\}$$

- $(3K + L + 1)$ dimensional constrained optimization problem

# Likelihood Inference

## EM based method – Complete data

- $\widetilde{I}$ is latent if $\delta = 0$
- $\widetilde{I} = 1$ if $\delta \neq 0$
- EM algorithm is a natural choice to obtain the MLEs
- The complete data :

$$\{(t_\ell,\, \delta_\ell,\, \boldsymbol{x}_\ell,\, i_\ell) : \ell = 1,\, 2,\, \ldots,\, n\}$$

- $i_\ell$ : Realized value of $\widetilde{I}$ for $\ell$-th subject

# Likelihood Inference

## EM based method – Complete data likelihood

- Complete data likelihood function is

$$L_c(\boldsymbol{\theta}) \propto \left[ \prod_{\ell \in J_1} \prod_{k=1}^{K} \{(1 - p_0(x_\ell; \boldsymbol{\beta}))\pi_k f_k(t_\ell; \boldsymbol{\theta}_k)\}^{I(\delta_\ell = k)} \right] \times \left[ \prod_{\ell \in J_0} \{p_0(x_\ell; \boldsymbol{\beta})\}^{1-i_\ell} \{(1 - p_0(x_\ell; \boldsymbol{\beta}))S_{mix}(t_\ell; \boldsymbol{\psi})\}^{i_\ell} \right]$$

- $J_1 = \{l : \delta_\ell \neq 0\}$ and $J_0 = \{l : \delta_\ell = 0\}$
- $\boldsymbol{\psi} = (\alpha_1, \ldots, \alpha_K, \lambda_1, \ldots, \lambda_K, \pi_1, \ldots, \pi_K)$

# Likelihood Inference

## EM based method – Complete data log-likelihood

- Complete data log-likelihood function is

$$\begin{aligned}\ln L_c(\boldsymbol{\theta}) = & \sum_{\ell \in J_1} \boldsymbol{\beta}' \boldsymbol{x}_\ell + \sum_{\ell \in J_1} \sum_{k=1}^{K} I(\delta_\ell = k) \ln \pi_k \\ & + \sum_{\ell \in J_1} \sum_{k=1}^{K} I(\delta_\ell = k) \ln f_k(t_\ell; \boldsymbol{\theta}_k) \\ & + \sum_{\ell \in J_0} i_\ell \boldsymbol{\beta}' \boldsymbol{x}_\ell + \sum_{\ell \in J_0} i_\ell \ln S_{mix}(\tau_\ell; \boldsymbol{\psi}) \\ & - \sum_{\ell=1}^{n} \ln(1 + \exp\{\boldsymbol{\beta}' \boldsymbol{x}_\ell\})\end{aligned}$$

# Likelihood Inference

## EM based method – E and M step

- At the $(m+1)-$st step, the expectation of complete data log-likelihood function is

$$\widetilde{Q}^{(m+1)}(\boldsymbol{\gamma}) = \widetilde{Q}_1^{(m+1)}(\boldsymbol{\beta}) + \widetilde{Q}_2^{(m+1)}(\boldsymbol{\psi})$$

- $\widetilde{Q}_1^{(m+1)}(\cdot)$ is a (non-linear) function of $\boldsymbol{\beta}$ only
- $\widetilde{Q}_2^{(m+1)}(\cdot)$ is a (non-linear) function of $\boldsymbol{\psi}$ only
- Maximize $\widetilde{Q}_1^{(m+1)}(\cdot)$ and $\widetilde{Q}_2^{(m+1)}(\cdot)$ separately

# Likelihood Inference

## Maximization of $\widetilde{Q}_1^{(m+1)}(\cdot)$

- $\widetilde{Q}_1^{(m+1)}(\cdot)$ is given by

$$\widetilde{Q}_1^{(m+1)}(\boldsymbol{\beta}) = \sum_{\ell \in J_0} w_\ell^{(m+1)} \boldsymbol{x}_\ell' \boldsymbol{\beta} + \sum_{\ell \in J_0} w_\ell^{(m+1)} \boldsymbol{x}_\ell' \boldsymbol{\beta} - \sum_{\ell=1}^{n} \log(1 + \exp\{\boldsymbol{x}_\ell' \boldsymbol{\beta}\})$$

- $w_\ell^{(m+1)} = \left. \frac{\exp(\boldsymbol{x}_\ell^T \boldsymbol{\beta}) S_{mix}(t_\ell; \boldsymbol{\psi})}{1 + \exp(\boldsymbol{x}_\ell^T \boldsymbol{\beta}) S_{mix}(t_\ell; \boldsymbol{\psi})} \right|_{\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}^{(m)}}$
- A $(L+1)$ dimensional non-linear optimization problem
- Numerical method can be used to solve

# Likelihood Inference

## Maximization of $\widetilde{Q}_2^{(m+1)}(\cdot)$

- $\widetilde{Q}_2^{(m+1)}(\cdot)$ is given by

$$\widetilde{Q}_2^{(m+1)}(\boldsymbol{\psi}) = \sum_{\ell \in J_1} \sum_{k=1}^{K} I(\delta_\ell = k) \left\{ \ln \pi_k + \ln f_k(t_\ell; \boldsymbol{\theta}_k) \right\} + \sum_{\ell \in J_0} w_\ell^{(m+1)} \ln S_{mix}(t_\ell; \boldsymbol{\psi}) - \mu \left( \sum_{k=1}^{K} \pi_k - 1 \right)$$

- Maximization of $\widetilde{Q}_2^{(m+1)}(\boldsymbol{\psi})$ boils down to solving $K$ one-dimensional non-linear equations [Jiang and Kececioglu 1992]

# Likelihood Inference

## Maximization of $\widetilde{Q}_2^{(m+1)}(\cdot)$

- In $(m+1)$-th step, update $\pi_k$ as

$$\widehat{\pi}_k^{(m+1)} = \frac{1}{r + \sum_{l \in J_0} w_l^{(m+1)}} \left[ r_k + \sum_{l \in J_0} w_l^{(m+1)} P^{(m+1)}(k|\tau_l) \right]$$

- $P^{(m+1)}(k|\tau) = \dfrac{\widehat{\pi}_k^{(m)} S_k(\tau; \widehat{\boldsymbol{\theta}}_k^{(m)})}{S_{mix}(\tau; \widehat{\boldsymbol{\psi}}^{(m)})}$
- $r_k$ : Number of subjects experiencing the event of interest due to $k$–th mode
- $r$ : Number of subjects experienced the event of interest
- $\sum_{k=1}^{K} r_k = r$

# Likelihood Inference

## Maximization of $\widetilde{Q}_2^{(m+1)}(\cdot)$

- At $(m+1)$-th step, find updated $\alpha_k$, $\widehat{\alpha}_k^{(m+1)}$, by solving

$$\frac{r_k}{\alpha_k} + \sum_{\ell \in J_1} I(\delta_\ell = k) \ln t_l - \frac{r_k A_1(\alpha_k)}{A_2(\alpha_k)} = 0,$$

where

$$A_1(\alpha_k) = \sum_{\ell \in J_1} I(\delta_\ell = k) t_l^{\alpha_k} \ln t_\ell + \sum_{\ell \in J_0} w_\ell^{(m+1)} P^{(m+1)}(k|t_\ell) t_\ell^{\alpha_k} \ln t_\ell$$

$$A_2(\alpha_k) = \sum_{\ell \in J_1} I(\delta_\ell = k) t_l^{\alpha_k} + \sum_{\ell \in J_0} w_\ell^{(m+1)} P^{(m+1)}(k|t_\ell) t_\ell^{\alpha_k}$$

# Likelihood Inference

## Maximization of $\widetilde{Q}_2^{(m+1)}(\cdot)$

- In $(m+1)$-th step, update $\lambda_k$ as

$$\widehat{\lambda}_k^{(m+1)} = \frac{r_k}{A_2(\widehat{\alpha}_k^{(m+1)})}$$

# Likelihood Inference

## Confidence Interval

- Model parameters – Luis missing value principle
- Cure rate – $\delta$-method

# Analysis of Melanoma Data

## Recall data

| no | status | days | ulc | thick | sex |
|---|---|---|---|---|---|
| 789 | 3 | 10 | 1 | 676 | 1 |
| 13 | 3 | 30 | 0 | 65 | 1 |
| 97 | 2 | 35 | 0 | 134 | 1 |
| 16 | 3 | 99 | 0 | 290 | 0 |
| 21 | 1 | 185 | 1 | 1208 | 1 |
| 469 | 1 | 204 | 1 | 484 | 1 |

- no : Patient code
- status : Survival status (1: dead from melanoma, 2: alive, 3: dead from other cause)
- days : Survival time (in days)
- ulc : Ulceration (1: present, 0: absent)
- thick : Tumour thickness (in 1/100 mm)
- sex : Gender (0: female, 1: male)

# Analysis of Melanoma Data

## Modelling

- Mode 1 – death due to malignant melanoma
- Mode 2 – death due to any other causes
- 3 covariates (ulcer status, tumour thickness and sex)
- Scale survival time dividing by 365
- Weibull model for cause specific distribution
- Logistic link function

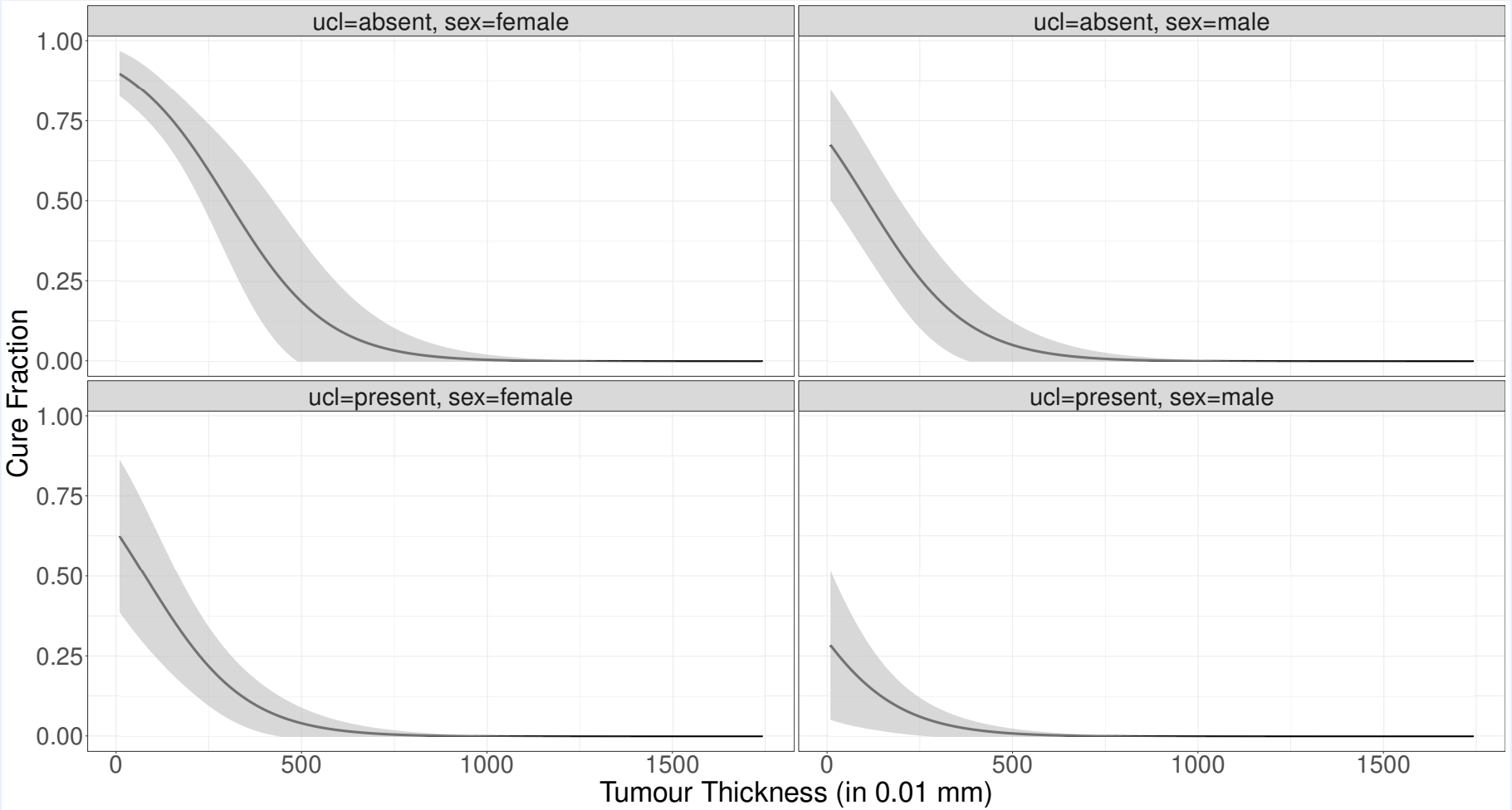
# Analysis of Melanoma Data

Model parameter estimates

| Parameters | MLE | 95% CI |
|---|---|---|
| $\alpha_1$ (shape, mode 1) | $1.603$ | $1.348 - 1.858$ |
| $\alpha_2$ (shape, mode 2) | $0.655$ | $0.321 - 0.988$ |
| $\lambda_1$ (scale, mode 1) | $0.085$ | $0.051 - 0.119$ |
| $\lambda_2$ (scale, mode 2) | $0.081$ | $0.039 - 0.123$ |
| $\pi_1$ (mixture coefficient, mode 1) | $0.564$ | $0.418 - 0.710$ |
| $\pi_2$ (mixture coefficient, mode 2) | $0.436$ | $0.206 - 0.666$ |
| $\beta_0$ (intercept) | $-2.238$ | $-3.014 - -1.461$ |
| $\beta_1$ (ulcer status) | $1.654$ | $0.796 - 2.512$ |
| $\beta_2$ (tumour thickness) | $0.007$ | $0.004 - 0.011$ |
| $\beta_3$ (sex) | $1.432$ | $0.608 - 2.256$ |

# Analysis of Melanoma Data

## Plots of cure fraction with respect to tumour thickness



# Bibliography


- Jiang, Siyuan and Dimitri Kececioglu (1992). "Maximum likelihood estimates, from censored data, for mixed-Weibull distributions". In: *IEEE Transactions on Reliability* 41.2, pp. 248–255.
- Pal, Suvra and N. Balakrishnan (2016). "Destructive negative binomial cure rate model and EM-based likelihood inference under Weibull lifetime". In: *Statistics & Probability Letters* 116, pp. 9–20.
- Rodrigues, J. et al. (2011). "Destructive weighted Poisson cure rate models.". In: *Lifetime data analysis* 17, pp. 333–346.

*Thank You*