A Flexible Finite Mixture Model Family for Analyzing Underdispersed Discrete Data, With Negative Weights

(Joint Statistical Meetings 2019)

Martial Luyts Geert Molenberghs Geert Verbeke Koen Matthijs

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Katholieke Universiteit Leuven, Belgium

martial.luyts@kuleuven.be

www.ibiostat.be

UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Denver, August 1, 2019

Contents

Part 1:

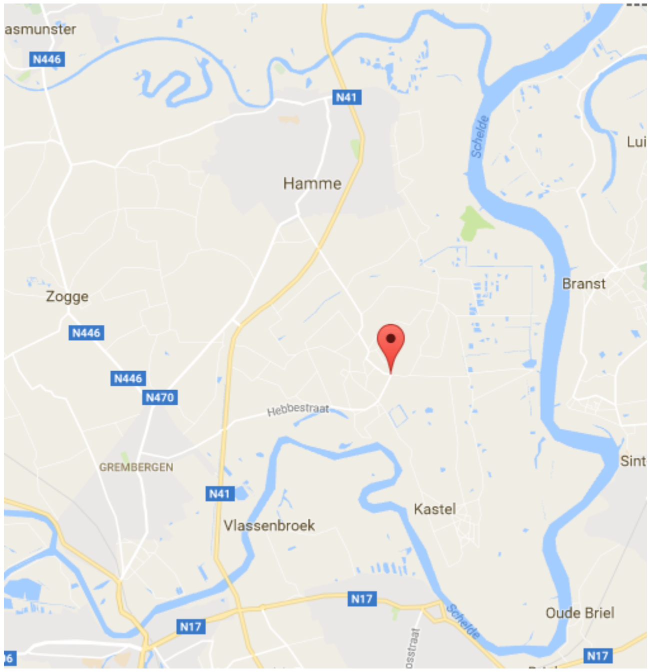
Introductory material

1.1 Demographic, historical data of Moerzeke

- **Moerzeke** is a small village in the center of Flanders (Belgium)

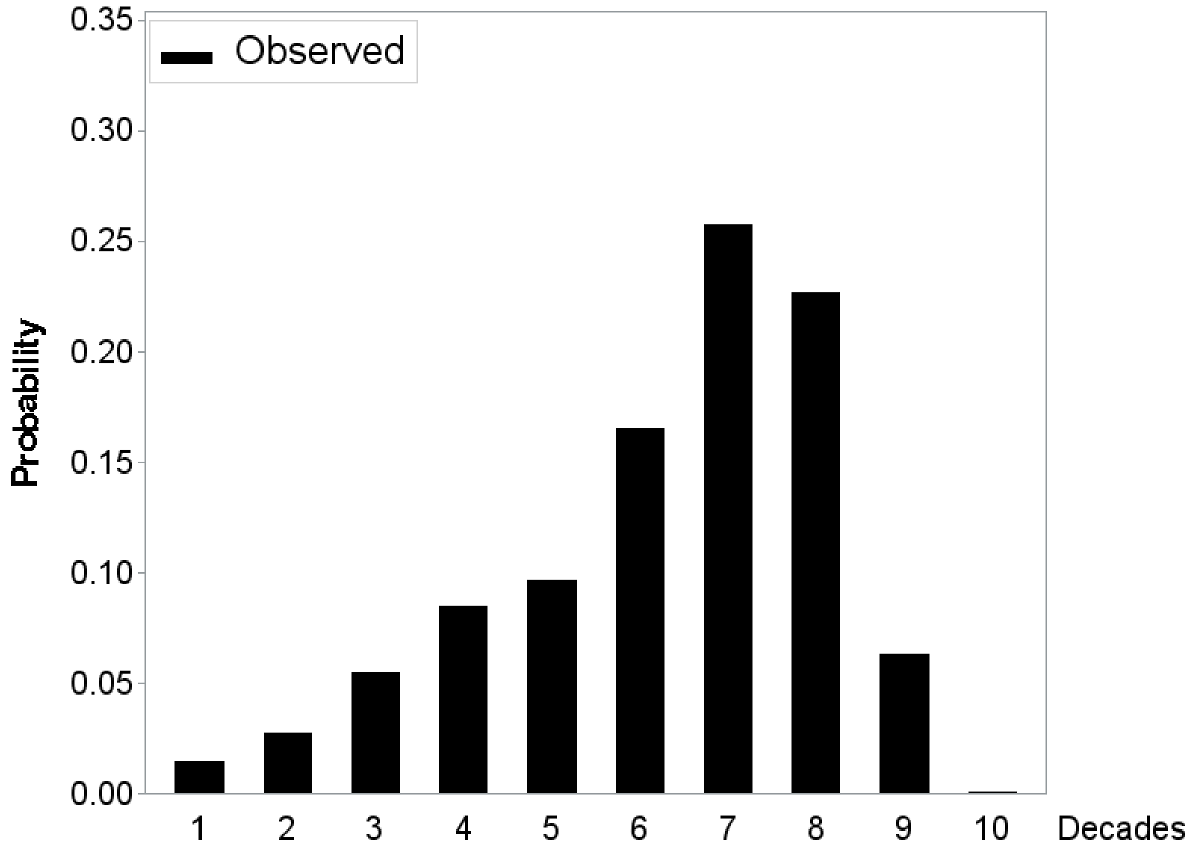

North Sea
Edinburgh
United Kingdom
Isle of Man
Copenhagen
Denmark
Dublin
Ireland
Manchester
Liverpool
Hamburg
Berlin
Amsterdam
Netherlands
London
Brussels
Cologne
Germany
Belgium
Frankfurt
Prague
Luxembourg
Czechia
Paris
Munich
Vienna
Austria
Switzerland
France
Slovenia
Zagreb
Milan
Croatia
Bay of Biscay
Monaco
Italy
Andorra

- It is a **geographical isolate**
- Mainly **populated by farmers** until well into the 20th century
- **Fertility** was traditionally **high** and dropped at the beginning of the 20th century



- The information in the database is drawn from church and civil registers
- The database contains information of individuals who were born, married or died in Moerzeke

- Focus is laid on the (discrete) longevity (measured per decades), i.e., a discretised time-to-event outcome



Part 2:

Methodology

2.1 Strategy

- Based on the given histogram, we give preference to a finite mixture model (FMM) approach:

$$p(Y = y \mid \boldsymbol{\theta}) = \sum_{j=1}^{k} \pi_j \cdot p_j(y \mid \boldsymbol{\theta}_j), \ \ \pi_j \geq 0 \ \text{ and } \ \sum_{j=1}^{k} \pi_j = 1$$

- We extend the traditional FMM approach to a more flexible framework, by

  1. Choosing flexible dispersed basic distributions $p_j(y \mid \theta_j)$

  2. Allowing for negative weights

$$p(Y = y \mid \boldsymbol{\theta}) = \sum_{j=1}^{k} \pi_j \cdot p_j(y \mid \boldsymbol{\theta}_j), \ \ \cancel{\pi_j \geq 0} \ \text{ and } \ \sum_{j=1}^{k} \pi_j = 1$$

Additional constraints:

$$\boxed{p(Y = y \mid \boldsymbol{\theta}) \geq 0, \forall y \ \text{ and } \ \text{Var}(Y) \geq 0}$$

2.1.1 Choosing flexible dispersed basic distributions

- Log-linear Poisson models are in standard use in count data
  - Main limitation: Restricted mean-variance relationship, i.e.,

$$\mathsf{E}_j(Y) = \lambda_j \text{ and } \mathsf{Var}_j(Y) = \lambda_j$$

$$(= \textbf{EQUIDISPERSION})$$

- Extended and alternative approaches have been developed that can flexibly handle over- and underdispersed situations

- Some examples:

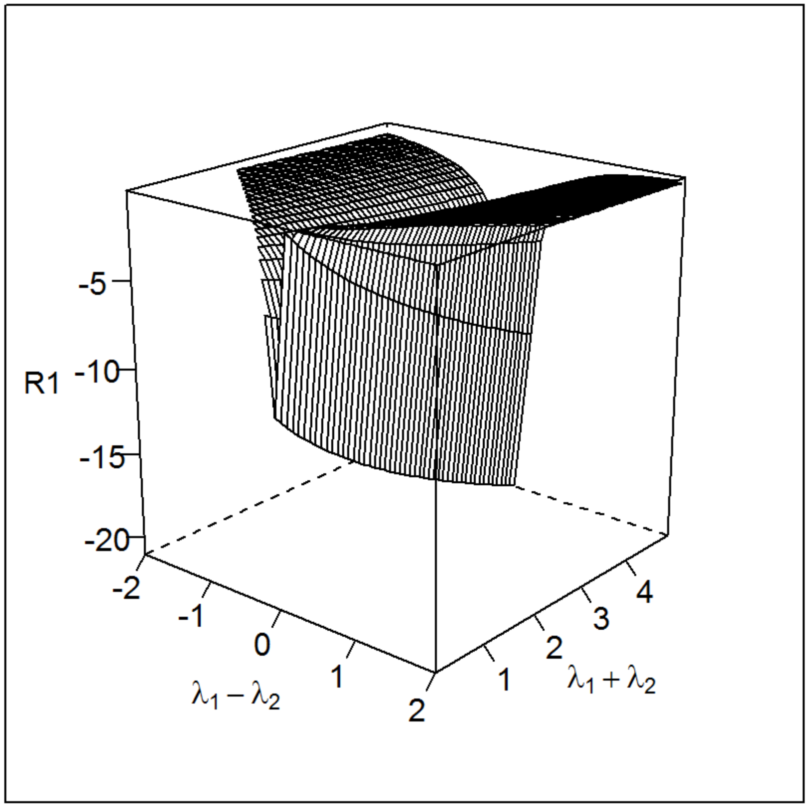
Element	Notation	Distribution	
Model		**Poisson**	**Discrete normal**
PMF	$p_j(y \mid \boldsymbol{\theta}_j)$	$\frac{e^{-\lambda_j}\lambda_j^y}{y!}$	$\Phi\left(\frac{y-\lambda_j+0.5}{\sigma_j}\right) - \Phi\left(\frac{y-\lambda_j-0.5}{\sigma_j}\right)$
Parameter(s)	$\boldsymbol{\theta}_j$	$\lambda_j > 0$	$(\lambda_j; \sigma_j) \in \mathbb{R}$
Mean	$\mathrm{E}_j(Y)$	λ_j	λ_j
Variance	$\mathrm{Var}_j(Y)$	λ_j	$\sigma_j^2 + 0.083333$
Dispersion		Only equi	Over/equi/under
Model		**Double Poisson**	**Discrete Weibull**
PMF	$p_j(y \mid \boldsymbol{\theta}_j)$	$K(\lambda_j, \phi_j)\phi_j^{1/2}\mathrm{e}^{-\phi_j\lambda_j}\frac{\mathrm{e}^{-y}y^y}{y!}\left(\frac{\mathrm{e}\lambda_j}{y}\right)^{\phi_j y}$	$\lambda_j^{y^{\rho_j}} - \lambda_j^{(y+1)^{\rho_j}}$
Constant		$\frac{1}{K(\lambda_j,\phi_j)} \approx 1 + \frac{1-\phi_j}{12\phi_j\lambda_j}\left(1 + \frac{1}{\phi_j\lambda_j}\right)$	
Parameter(s)	$\boldsymbol{\theta}_j$	$\lambda_j > 0; \phi_j \in \mathbb{R}$	$0 < \lambda_j < 1; \rho_j > 0$
Mean	$\mathrm{E}_j(Y)$	λ_j	$\sum_{n=1}^{+\infty} \lambda_j^{n^{\rho_j}}$
Variance	$\mathrm{Var}_j(Y)$	λ_j/ϕ_j	$2\sum_{n=1}^{+\infty} n\lambda_j^{n^{\rho_j}} - \mathrm{E}_j(Y) - [\mathrm{E}_j(Y)]^2$
Dispersion		Over/equi/under	Over/equi/under

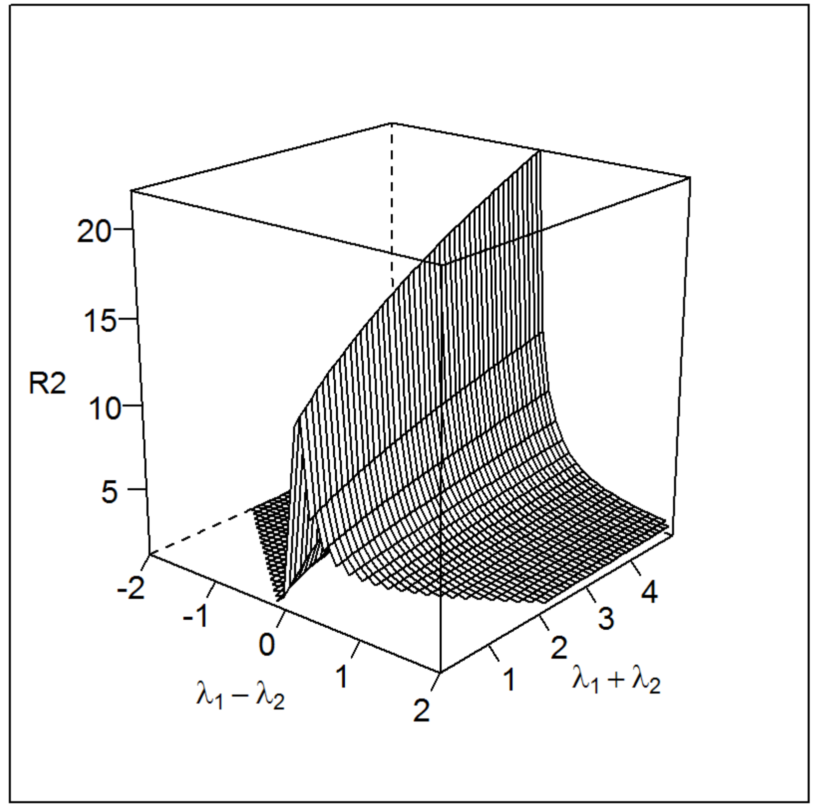
2.1.2 Allowing for negative weights

- Adds more flexible to the FMM framework
- But what added value does this creates?
- **Example:** 2-component mixture of Poisson models

$$
\begin{aligned}
p(Y = y \mid \lambda_1, \lambda_2) &= \pi_1 \frac{e^{-\lambda_1}\lambda_1^y}{y!} + (1 - \pi_1)\frac{e^{-\lambda_2}\lambda_2^y}{y!}, \\
\mathsf{E}(Y) &= \pi_1\lambda_1 + (1 - \pi_1)\lambda_2, \\
\mathsf{Var}(Y) &= \pi_1\lambda_1^2 + (1 - \pi_1)\lambda_2^2 - [\pi_1\lambda_1 + (1 - \pi_1)\lambda_2]^2 \\
&\quad + \pi_1\lambda_1 + (1 - \pi_1)\lambda_2,
\end{aligned}
$$

- The new constraints extends the boundary of weight π_1 from $[0, 1]$ to $[R_1, R_2]$





- Remark: $[0, 1] \subset [R_1, R_2]$

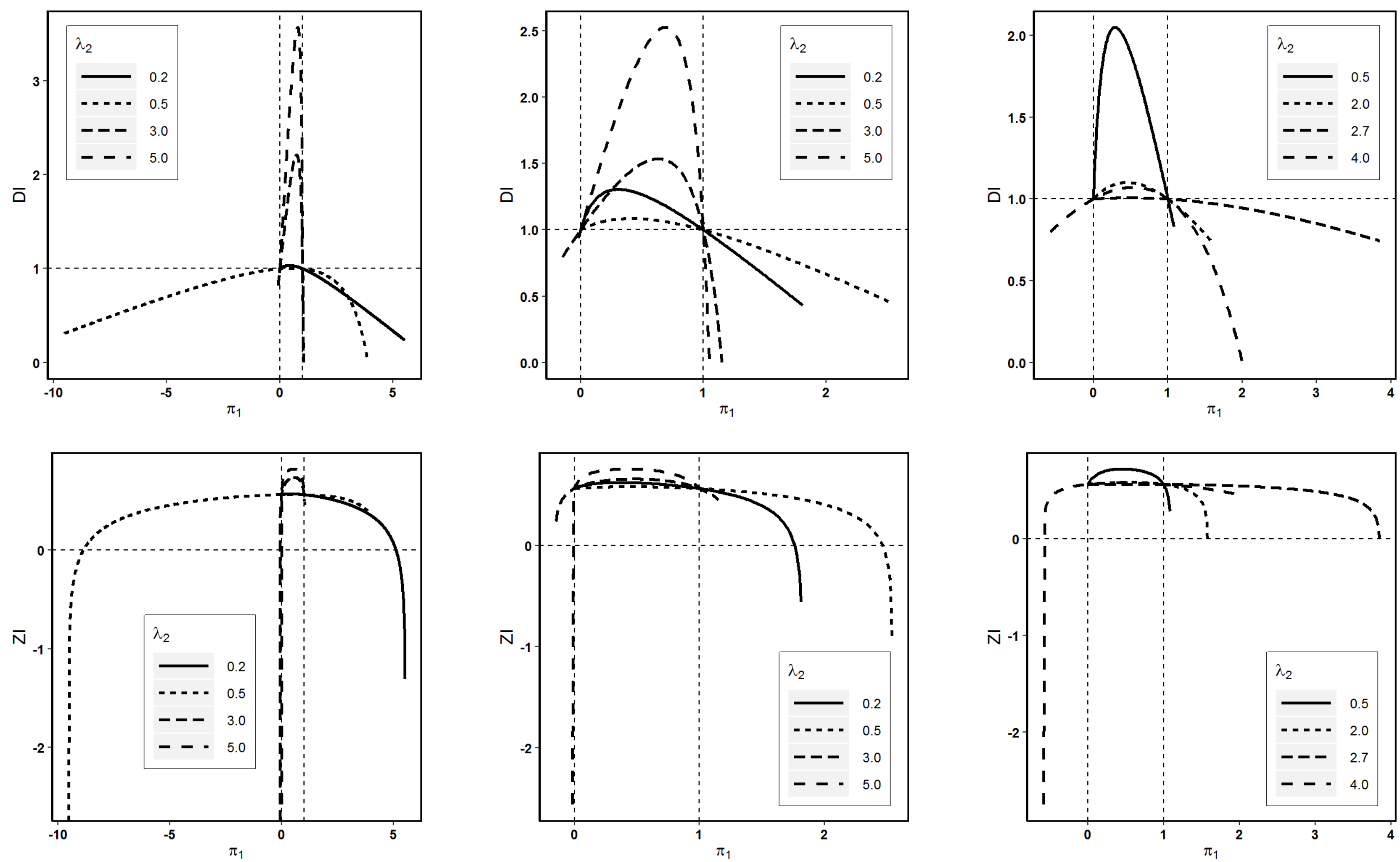
- Characteristics:

$$\text{DI} = \frac{\text{Var}(Y)}{\text{E}(Y)}, \qquad \text{ZI} = 1 + \frac{\log[p(Y = 0 \mid \lambda_1, \lambda_2)]}{\text{E}(Y)}.$$

- DI > 1: Overdispersion
- DI $= 1$: Equidispersion
- DI < 1: Underdispersion

- ZI > 0: Zero-inflation
- ZI $= 0$: No excess of zeros
- ZI < 0: Zero-deflation

(a) $\lambda_1 = 0.4$ (b) $\lambda_1 = 1$ (c) $\lambda_1 = 3$



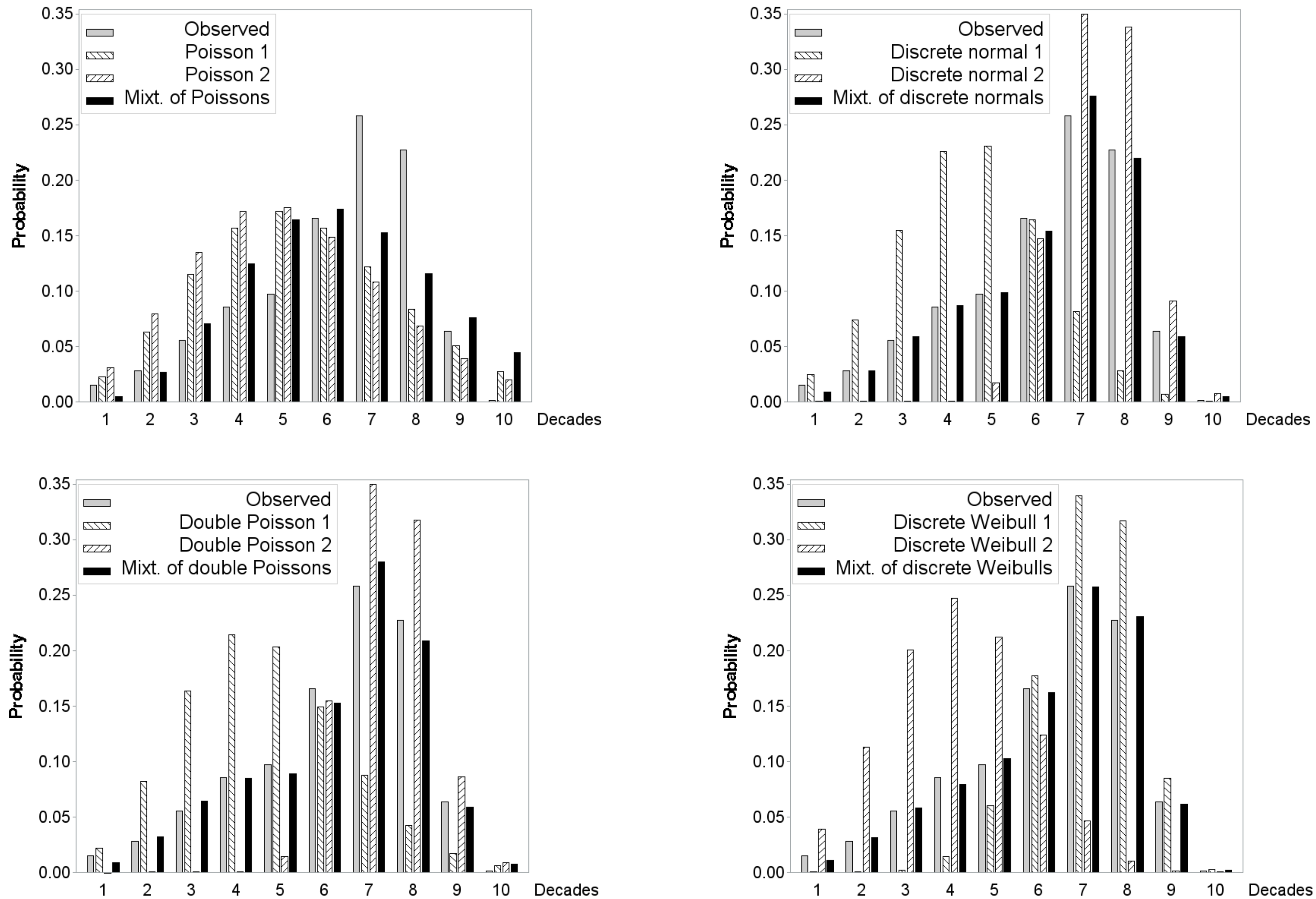
Part 3:

Analyzing the Moerzeke data

3.1 Findings with the extended FMM approach

- Mixtures of 2 similar elementary components are considered

		Mixt. Poissons	Mixt. discrete normals	Mixture double-Poissons	Mixt. discrete-Weibulls
Effect	Par.	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)
Intensity 1	λ_1	5.4661 (0.1113)	4.5533 (0.2484)	4.6775 (0.1675)	0.9999 ($3.2E-8$)
Std. dev. 1	σ_1	$--$ ($--$)	1.6430 (0.1174)	$--$ ($--$)	$--$ ($--$)
Dispersion 1	ϕ_1	$--$ ($--$)	$--$ ($--$)	1.3853 (0.1064)	$--$ ($--$)
	ρ_1	$--$ ($--$)	$--$ ($--$)	$--$ ($--$)	8.3960 (0.6059)
Intensity 2	λ_2	5.0918 (0.1307)	7.3614 (0.0626)	7.3376 (0.0484)	0.9956 (0.0014)
Std. dev. 2	σ_2	$--$ ($--$)	0.8862 (0.0438)	$--$ ($--$)	$--$ ($--$)
Dispersion 2	ϕ_2	$--$ ($--$)	$--$ ($--$)	8.4797 (0.6981)	$--$ ($--$)
	ρ_2	$--$ ($--$)	$--$ ($--$)	$--$ ($--$)	3.3182 (0.2914)
Mixing prob.	π_1	3.1892 (1.2438)	0.3833 (0.0449)	0.3956 (0.0305)	0.7194 (0.0702)
-2 log-lik.		5814.1	5324.2	5358.9	5310.9
AIC		5820.1	5334.2	5368.9	5320.9
BIC		5835.7	5360.3	5395.0	5347.0


Observed
Poisson 1
Poisson 2
Mixt. of Poissons
Probability
Decades
Observed
Discrete normal 1
Discrete normal 2
Mixt. of discrete normals
Probability
Decades
Observed
Double Poisson 1
Double Poisson 2
Mixt. of double Poissons
Probability
Decades
Observed
Discrete Weibull 1
Discrete Weibull 2
Mixt. of discrete Weibulls
Probability
Decades


danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
bedankt
hvala
mauruuru
dziękuję
mochchakkeram
go raibh maith agat
sagolun
sukriya
kop khun krap
obrigado
grazie
arigatô
takk
dakujem
мерси
terima kasih
감사합니다
ευχαριστώ
merci