Phase-type regression

Prague 2021

Martin Bladt

Faculty of Business and Economics, University of Lausanne

November 19, 2021.

Basic concepts

- Let $(J_t)_{t\geq 0}$ denote a Markov jump process on a state space $\{1,\ldots,p,p+1\}$, where states $1,\ldots,p$ are transient and $p+1$ absorbing.
- Transition probabilities

$$p_{ij}(s,t) = \mathbb{P}(J_t = j | J_s = i), \quad i,j \in \{1,\ldots,p+1\},$$

- Intensity:

$$\boldsymbol{\Lambda}(t) = \begin{pmatrix} \boldsymbol{T}(t) & \boldsymbol{t}(t) \\ \mathbf{0} & 0 \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}, \quad t \geq 0,$$

implies that

$$\boldsymbol{P}(s,t) = \prod_s^t (\boldsymbol{I} + \boldsymbol{\Lambda}(u)du) := \boldsymbol{I} + \sum_{k=1}^{\infty} \int_s^t \int_s^{u_k} \cdots \int_s^{u_2} \boldsymbol{\Lambda}(u_1)\cdots\boldsymbol{\Lambda}(u_k)\, du_1 \cdots \mathrm{d}u_k,$$

for $s < t$, where $\boldsymbol{\Lambda}(t)$ is an intensity matrix.

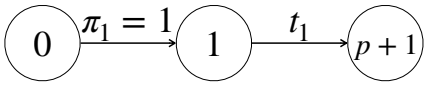
- If the matrices $\boldsymbol{T}(s)$ and $\boldsymbol{T}(t)$ commute for any $s < t$ we may write

$$\boldsymbol{P}(s,t) = \begin{pmatrix} \exp\left(\int_s^t \boldsymbol{T}(u)du\right) & \boldsymbol{e} - \exp\left(\int_s^t \boldsymbol{T}(u)du\right)\boldsymbol{e} \\ \mathbf{0} & 1 \end{pmatrix}, \quad s < t.$$

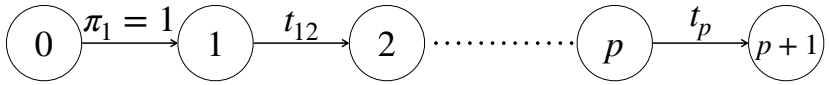
Basic Concepts II

- Initial distribution
$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p).$$
- **Inhomogeneous phase-type** distributed random variable:
$$Y = \inf\{t > 0 : J_t = p + 1\}.$$
- Assumption 1:
$$\boldsymbol{T}(t) = \lambda(t)\,\boldsymbol{T},$$
with $\lambda(t)$ some positive function.
- Assumption 2: the map
$$y \mapsto \int_0^y \lambda(s)ds \in (0, \infty), \quad \forall y > 0,$$
converges to infinity as $y \to \infty$.
- Assumption 3: The function λ is parametric.

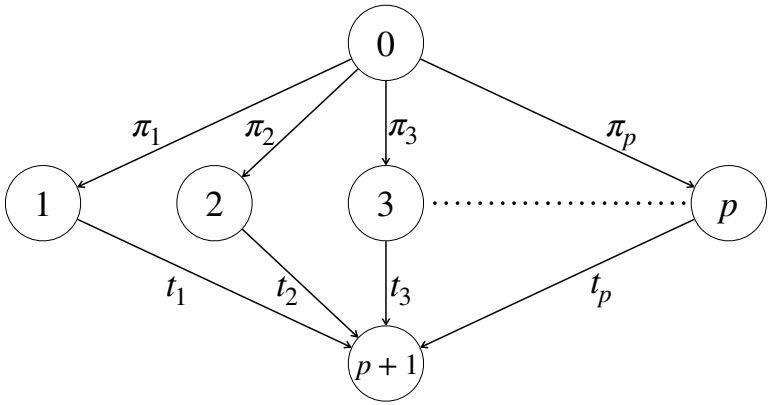
Sub-intensity structures



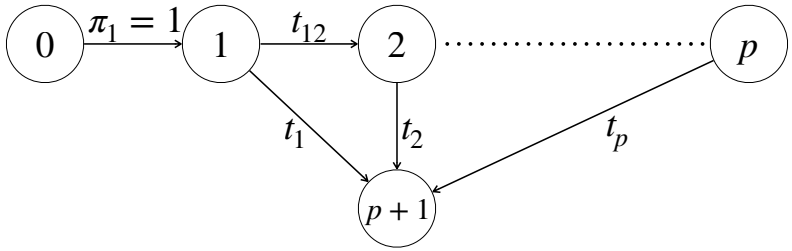
(a) Exponential



(b) Erlang



(c) Hyper-exponential



(d) Coxian

Figure: Underlying Markov structures. Names are borrowed from the corresponding PH representations, but apply to our inhomogeneous setup as well. The state 0 is added for schematic reasons, but is not part of the actual state-space of the chain.

Properties and tail behaviour

- Dense (weak convergence) on all positive distributions.

Proposition

Let $Y \sim IPH(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$. Then the survival function $S_Y = 1 - F_Y(y)$, density f_Y, hazard function h_Y and cumulative hazard function H_Y of Y satisfy, respectively, as $t \to \infty$,

$$S_Y(y) = \boldsymbol{\pi} \exp\left(\int_0^y \lambda(s)ds\ \boldsymbol{T}\right)\boldsymbol{e} \sim c_1[g^{-1}(y)]^{n-1}e^{-\chi[g^{-1}(y)]},$$

$$f_Y(y) = \lambda(y)\,\boldsymbol{\pi} \exp\left(\int_0^y \lambda(s)ds\ \boldsymbol{T}\right)\boldsymbol{t} \sim c_2[g^{-1}(y)]^{n-1}e^{-\chi[g^{-1}(y)]}\lambda(y),$$

$$h_Y(y) \sim c\lambda(y),$$

$$H_Y(y) \sim kg^{-1}(y),$$

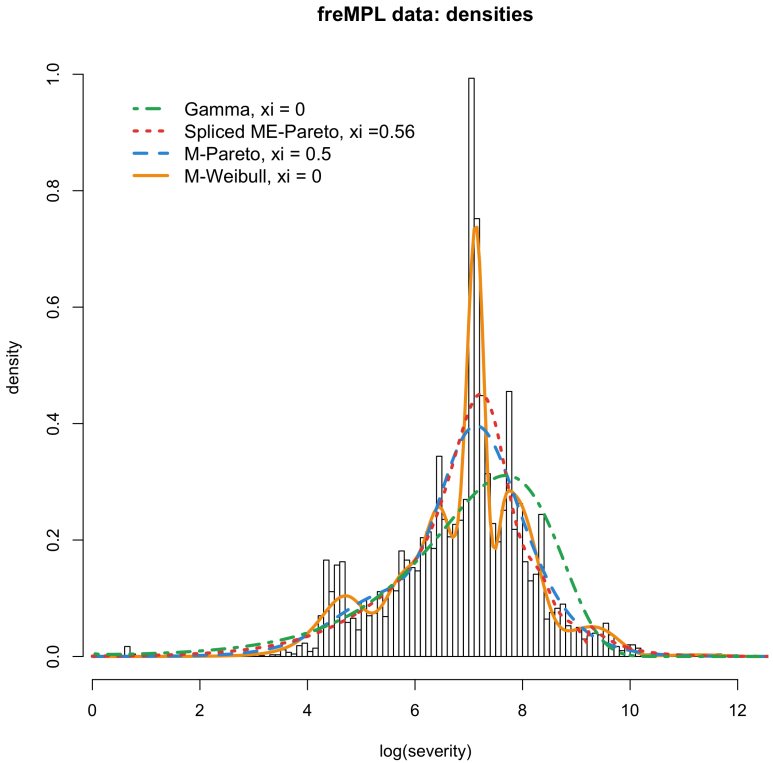
where c_1, c_2, c, k are positive constants, $-\chi$ is the largest real eigenvalue of $\boldsymbol{T}$ and n is the dimension of the Jordan block associated to χ. Here $g^{-1}(y) = \int_0^y \lambda(s)$.

Parametrizations

	$\lambda(t)$	$g(y)$	Parameters Domain
Matrix-Pareto	$(t+\beta)^{-1}$	$\beta\left(\exp(y)-1\right)$	$\beta > 0$
Matrix-Weibull	$\beta t^{\beta-1}$	$y^{1/\beta}$	$\beta > 0$
Matrix-Lognormal	$\frac{\gamma(\log(s+1))^{\gamma-1}}{s+1}$	$\exp(y^{1/\gamma})-1$	$\gamma > 1$
Matrix-Loglogistic	$\theta t^{\theta-1}/(t^{\theta}+\gamma^{\theta})$	$\gamma(\exp(y)-1)^{1/\theta}$	$\gamma, \theta > 0$
Matrix-Gompertz	$\exp(\beta t)$	$\log(\beta y+1)/\beta$	$\beta > 0$
Matrix-GEV	-	$\mu+\sigma(y^{-\xi}-1)/\xi$	$\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$

Example: FreMPL severity (Marginal)

Datasets `freMPL1`, `freMPL2`, `freMPL3`, `freMPL4`. Total 7008 claim severities for about 30,000 policies in 2004. 18 covariates (for later...)



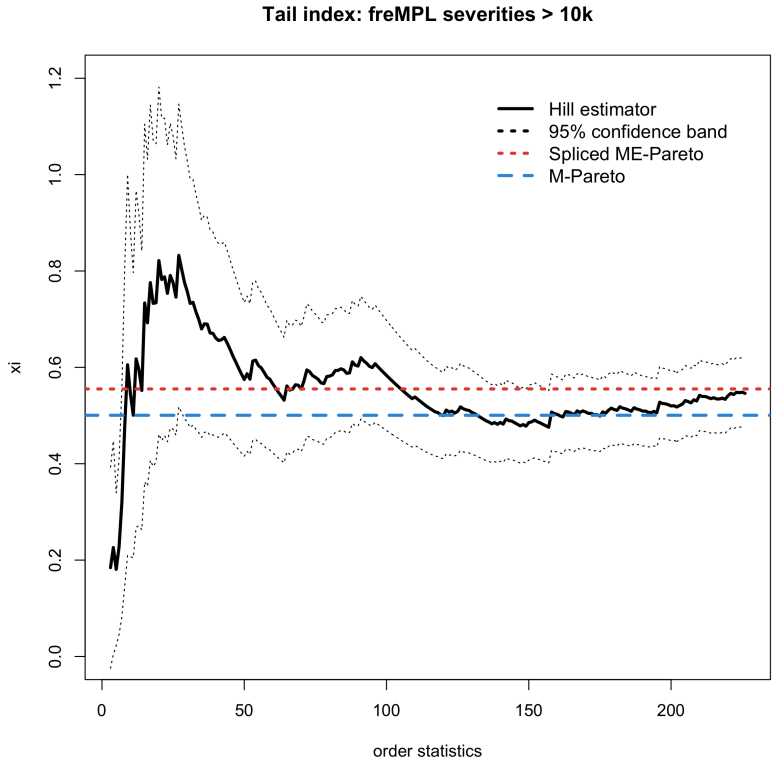


Figure: AIC: 121311; 119247; 119231; **119142**. $p = 3, 20$. Estimation using `matrixdist`, in CRAN.

From marginal to conditional specifications

$$S_Y(y) = \boldsymbol{\pi} \exp\left(\int_0^y \lambda(s)ds \; \boldsymbol{T}(\boldsymbol{X})\right) \boldsymbol{e}$$

Proportional Intensities (Non-life insurance)

From marginal to conditional specifications

$$S_Y(y) = \boldsymbol{\pi} \exp\left(\int_0^y \lambda(s, \boldsymbol{X}) \boldsymbol{T}(\boldsymbol{X}) ds \right) \boldsymbol{e}$$

Extended Proportional Intensities (Mortality modeling)

From marginal to conditional specifications

$$S_Y(y) = \boldsymbol{\pi}(\boldsymbol{X}) \exp\left(\int_0^y \lambda(s)ds\, \boldsymbol{T}\right)\boldsymbol{e}$$

Mixture-of-Experts (Non-life insurance)

Proportional intensities

- Since $\lambda(\,\cdot\,;\theta)$ is a parametric non–negative function depending on the parameter θ, we incorporate the predictor variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ by specifying

$$\lambda(t \mid \boldsymbol{X}, \boldsymbol{\beta}) = \lambda(t;\theta) m(\boldsymbol{X}\boldsymbol{\beta}), \quad t \geq 0, \tag{1}$$

  where $m\beta(\boldsymbol{X}\boldsymbol{\gamma})$ is a positive-valued function of the score $\boldsymbol{X}\boldsymbol{\beta}$.
- Example:

$$m(\boldsymbol{X}\boldsymbol{\beta}) = \exp(\boldsymbol{X}\boldsymbol{\beta})$$

  is a natural choice.
- The conditional mean can be written on the form

$$\mu(Y|\boldsymbol{X}) = \int_0^\infty \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}\boldsymbol{\beta}) \int_0^y \lambda(s;\theta) ds\, \boldsymbol{T} \right) \boldsymbol{e}\, dy. \tag{2}$$

Special cases

A simple special case is obtained by the following choices, giving a Gamma GLM with canonical link: take $\boldsymbol{T} = -1$, and $\lambda \equiv 1$ to receive

$$\mu(Y|\boldsymbol{X}) = \int_0^\infty \exp(-m(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})y)\, dy = \frac{1}{m(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})}.$$

Another slightly more complex special case is that of regression for Matrix-Weibull distributions, which contains the pure PH specification (when $\lambda \equiv 1$). In this setting it is not hard to see that

$$\mu(Y|\boldsymbol{X}) = \int_0^\infty \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})\boldsymbol{T}y^\theta\right)\boldsymbol{e}\, dy = \frac{\Gamma(1+\theta^{-1})\boldsymbol{\pi}\boldsymbol{T}^{-\theta^{-1}}\boldsymbol{e}}{m(\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta})^{\theta^{-1}}}. \tag{3}$$

Estimation

- Define g as

$$g^{-1}(y|\theta) = \int_0^y \lambda(s;\theta)ds$$

so that

$$Z = g^{-1}(Y/m(\boldsymbol{X}\boldsymbol{\beta}) \,|\, \theta) \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T}) \,. \tag{4}$$

- Define: B_k be the number of times that the underlying jump-process $\{J_t\}_{t\geq 0}$ initiates in state k.N_{ks} the total number of jumps from state k to s. N_k the number of times that we reach the absorbing state $p+1$ from state k. Z_k be the total time that the underlying Markov jump process spends in state k prior to absorption.
- Complete likelihood:

$$\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}; \boldsymbol{z}) = \prod_{k=1}^{p} {\pi_k}^{B_k} \prod_{k=1}^{p} \prod_{s\neq k} {t_{ks}}^{N_{ks}} e^{-t_{ks} Z_k} \prod_{k=1}^{p} {t_k}^{N_k} e^{-t_k Z_k} \,, \tag{5}$$

with explicit maximum likelihood estimators.

Estimation II

Step 1: (*E-step*) compute the statistics

$$\mathbb{E}(B_k \mid \boldsymbol{Z} = \boldsymbol{z}) = \sum_{i=1}^{N} \frac{\pi_k \boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}{\boldsymbol{\pi} \exp(\boldsymbol{T} x_i)\boldsymbol{t}}$$

$$\mathbb{E}(Z_k \mid \boldsymbol{Z} = \boldsymbol{z}) = \sum_{i=1}^{N} \frac{\int_0^{z_i} \boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}$$

$$\mathbb{E}(N_{ks} \mid \boldsymbol{Z} = \boldsymbol{z}) = \sum_{i=1}^{N} t_{ks} \frac{\int_0^{z_i} \boldsymbol{e}_s{}^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\,\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}$$

$$\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z}) = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e} y_k}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}.$$

Computational tricks are needed for M.E. and integrals of M.E.

Estimation III

Step 2: (*M-step*)

$$\hat{\pi}_k = \frac{\mathbb{E}(B_k \mid \boldsymbol{Z} = \boldsymbol{z})}{N}, \quad \hat{t}_{ks} = \frac{\mathbb{E}(N_{ks} \mid \boldsymbol{Z} = \boldsymbol{z})}{\mathbb{E}(Z_k \mid \boldsymbol{Z} = \boldsymbol{z})}$$

$$\hat{t}_k = \frac{\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z})}{\mathbb{E}(Z_k \mid \boldsymbol{Z} = \boldsymbol{z})}, \quad \hat{t}_{kk} = -\sum_{s \neq k} \hat{t}_{ks} - \hat{t}_k.$$

Step 3: (*Inhomogeneity optimization*)

$$(\hat{\theta}, \hat{\boldsymbol{\beta}}) = \arg\max_{(\theta, \boldsymbol{\beta})} \sum_{i=1}^{N} \log(f_Y(y_i; \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \theta, \boldsymbol{\beta}))$$

$$= \arg\max_{(\theta, \boldsymbol{\beta})} \sum_{i=1}^{N} \log\left(m(\boldsymbol{x}_i\boldsymbol{\beta})\lambda(y;\theta)\, \boldsymbol{\pi} \exp\left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \int_0^y \lambda(s;\theta) ds\, \boldsymbol{T} \right) \boldsymbol{t} \right)$$

and re-transform data.
Iterating the above three steps is an increasing EM algorithm.

Inference

Problem: $\boldsymbol{\pi}$ and $\boldsymbol{T}$ are non-identifiable. Proposed solution: use partial likelihood. Let $\ell_{\boldsymbol{y}}(\boldsymbol{\beta}, \theta)$ be the log-likelihood function of the observed severities $\boldsymbol{y} = (y_1, \ldots, y_N)$ with rating factors $\overline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. As $N \to \infty$,

$$\widehat{(\boldsymbol{\beta}, \theta)} \overset{d}{\approx} \mathcal{N}((\boldsymbol{\beta}, \theta), \mathcal{I}^{-1}),$$

where

$$[\mathcal{I}]_{jk} = \begin{cases} \sum_{i=1}^{N} G_1(i, j|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) G_1(i, k|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) & 1 \leq j, k \leq d, \\ \sum_{i=1}^{N} G_1(i, j|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) G_2(i|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) & 1 \leq j \leq d,\ k = d+1, \\ \sum_{i=1}^{N} G_2^2(i|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) & j = k = d+1. \end{cases} \tag{6}$$

and

$$G_1(i, j|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) = x_{ij} m'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \left(\frac{1}{m(\boldsymbol{x}_i^\top \boldsymbol{\beta})} + \frac{\boldsymbol{\pi} \exp\left(m(\boldsymbol{x}_i^\top \boldsymbol{\beta}) h(y; \theta) \boldsymbol{T}\right) h(y_i; \theta) \boldsymbol{T} \boldsymbol{t}}{\boldsymbol{\pi} \exp\left(m(\boldsymbol{x}_i^\top \boldsymbol{\beta}) h(y_i; \theta)\ \boldsymbol{T}\right) \boldsymbol{t}} \right),$$

$$G_2(i|\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \theta, \boldsymbol{y}, \overline{\boldsymbol{x}}) = \frac{\frac{d}{d\theta} \lambda(y_i; \theta)}{\lambda(y_i; \theta)} + \frac{\boldsymbol{\pi} \exp\left(m(\boldsymbol{x}_i^\top \boldsymbol{\beta}) h(y_i, \theta) \boldsymbol{T}\right) m(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \frac{d}{d\theta} h(y_i, \theta) \boldsymbol{T} \boldsymbol{t}}{\boldsymbol{\pi} \exp\left(m(\boldsymbol{x}_i^\top \boldsymbol{\beta}) h(y_i; \theta) \boldsymbol{T}\right) \boldsymbol{t}}$$

Example: FreMPL severity

Note: 18 covariates and all their levels were pre-processed using decision trees.

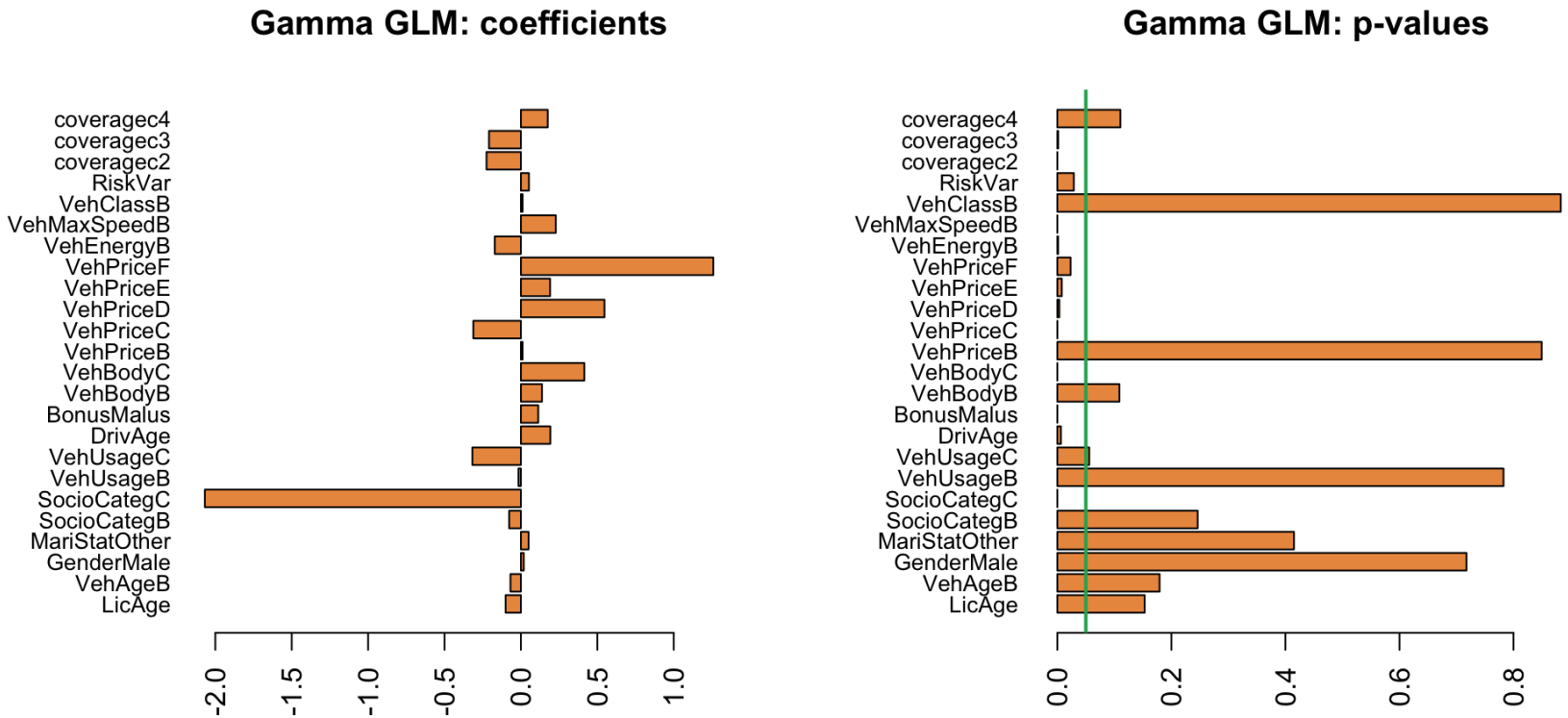


Figure: Coefficients and p-values of IPH and GLM regression. For display: IPH coefficients multiplied by -1 and intercept of GLM omitted.

Example: FreMPL severity

Note: 18 covariates and all their levels were pre-processed using decision trees.

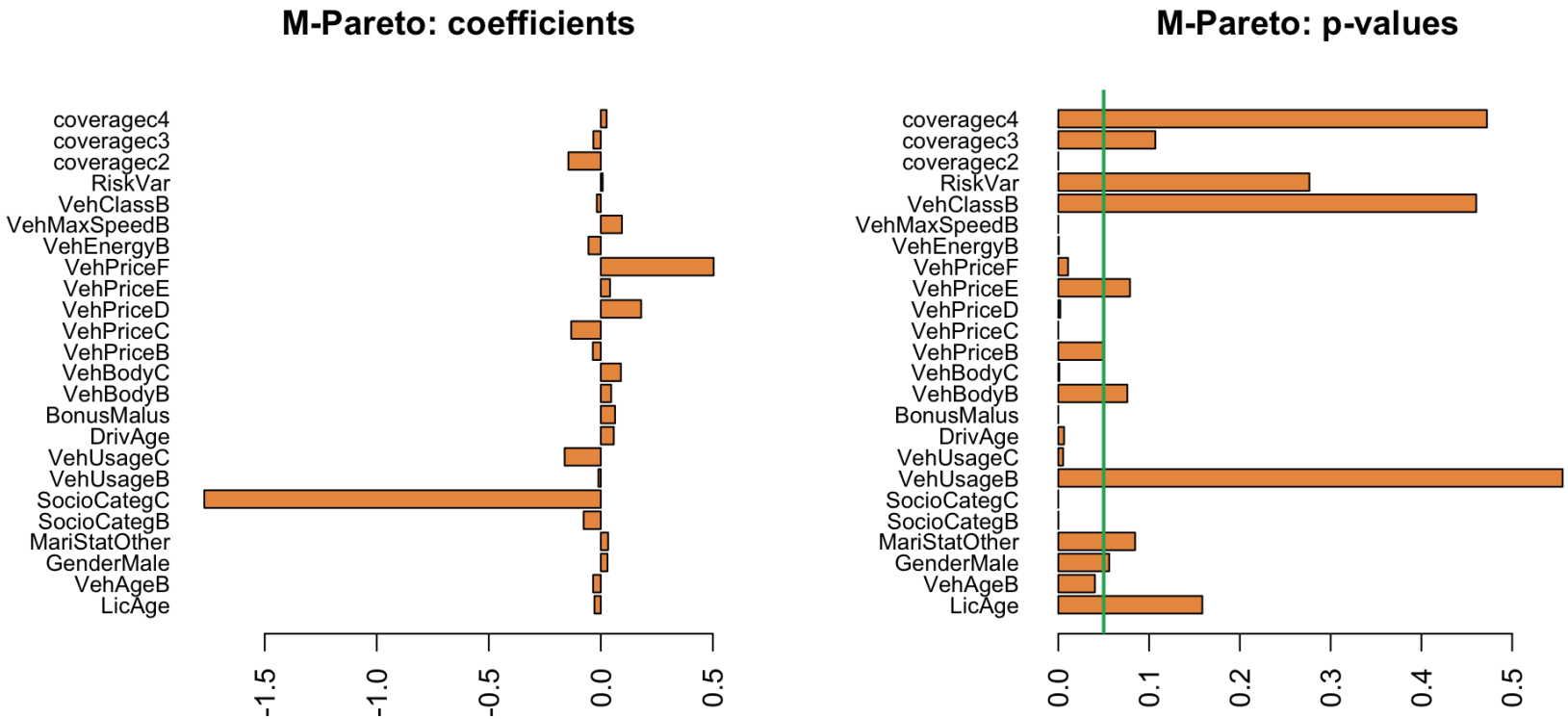


Figure: Coefficients and p-values of IPH and GLM regression. For display: IPH coefficients multiplied by -1 and intercept of GLM omitted.

Example: FreMPL severity

Note: 18 covariates and all their levels were pre-processed using decision trees.

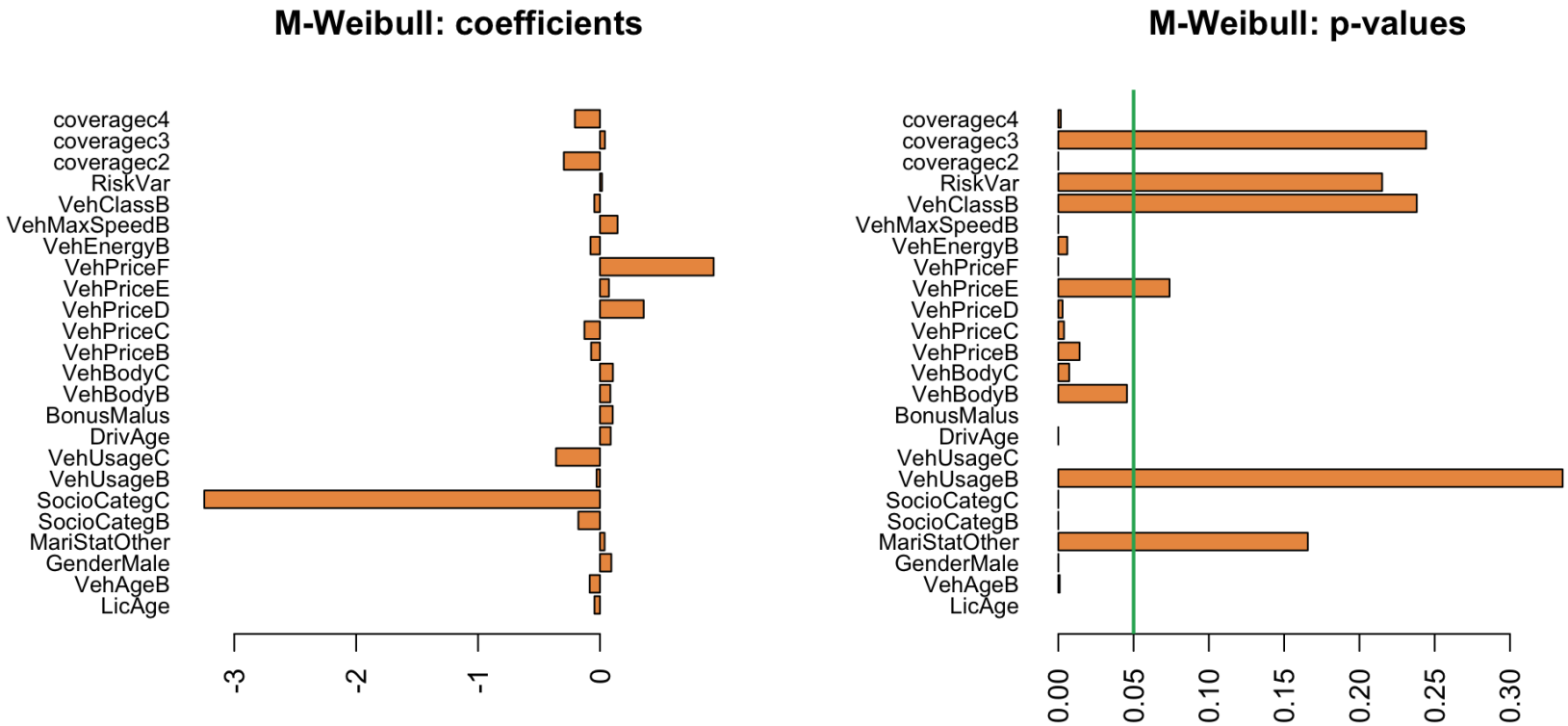


Figure: Coefficients and p-values of IPH and GLM regression. For display: IPH coefficients multiplied by -1 and intercept of GLM omitted.

Example: FreMPL severity

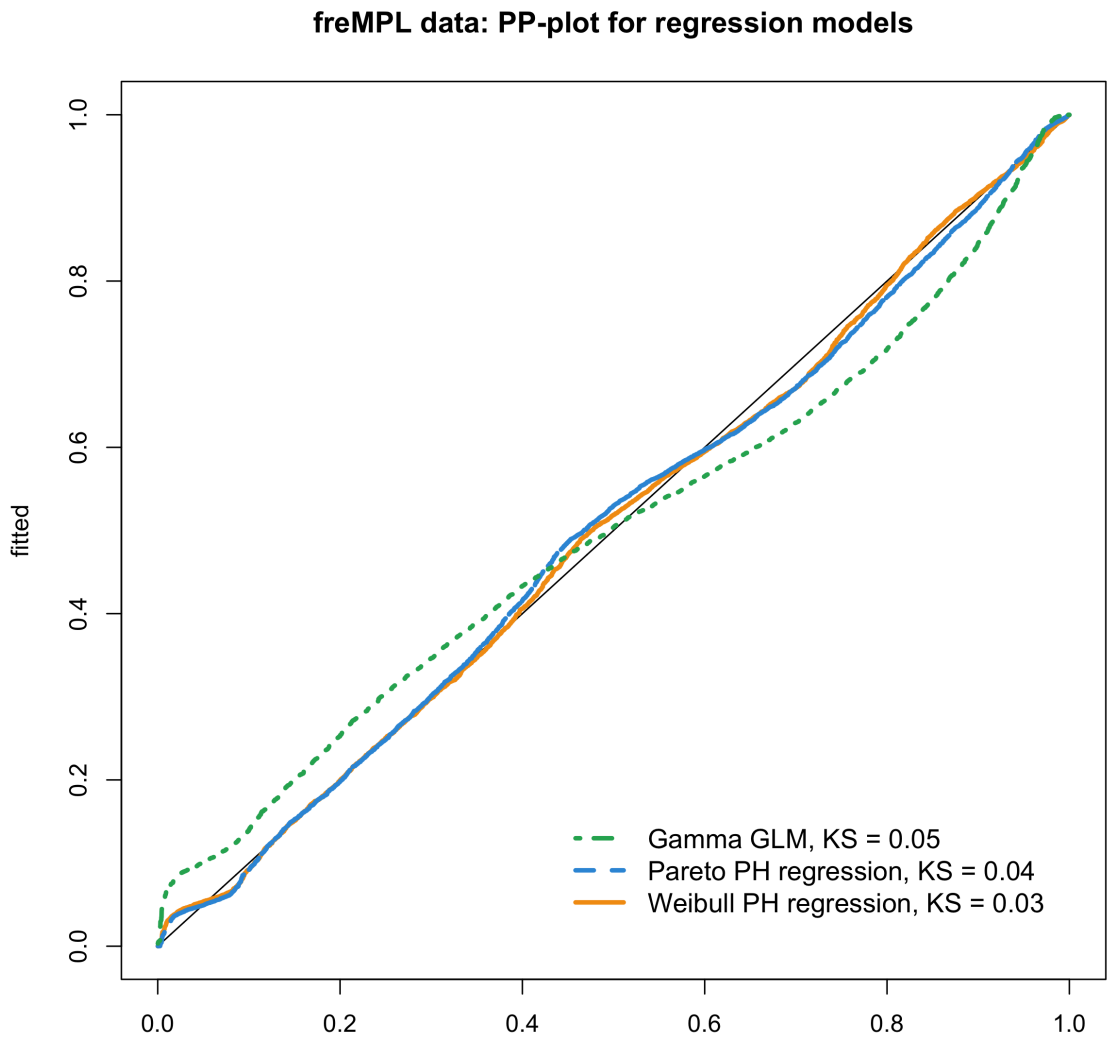
Even if Information Criteria are bad for PH:

Table: Summary for GLM and PH regression models for the freMPL dataset.

	Gamma GLM	Pareto PH reg.	Weibull PH reg.
Log Likelihood	$-60,368$	$-59,464$	$-59,446$
Degrees of freedom	26	34	34
AIC	$120,788$	$118,996$	**118,961**
BIC	$120,966$	$119,229$	**119,194**
Num. obs.	$7,008$	$7,008$	$7,008$
Loss-ratio (pure)	101.03%	105.18%	101.13%

Quantiles can be shown to provide much better performance.

Example: FreMPL severity (with rating factors)



Extended proportional intensities specification

- Now $\lambda(\,\cdot\,;\boldsymbol{\theta})$ is a parametric non–negative function depending on the vector $\boldsymbol{\theta}$, and we incorporate the predictor variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ by specifying

$$\lambda(t \mid \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda(t; \boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma}))m(\boldsymbol{X}\boldsymbol{\beta}), \quad t \geq 0, \tag{7}$$

  where $\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma})$ is a vector-valued function mapping the score $\boldsymbol{X}\boldsymbol{\gamma}$ to the parameter space of λ.

- Example:

$$\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma}) = \exp(\gamma_0 + \boldsymbol{X}\boldsymbol{\gamma})$$

  is a natural choice. Here, an intercept makes sense.

- λ according to a Gompertz tail will be used throughout (mortality!)

Estimation strategy

- The E and M steps remain the same, and the inhomogeneity optimization changes to

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\arg\max} \left\{ \sum_{i=1}^{N} \log \left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \lambda(y; \boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma})) \, \hat{\boldsymbol{\pi}} \exp \left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \int_0^y \lambda(s; \boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma})) ds \; \hat{\boldsymbol{T}} \right) \hat{\boldsymbol{t}} \right) \right.$$

- However, mortality modeling requires equal weighting for each age.
- Hence, we use the EM algorithm as a good initial guess, and we design another fitting procedure.

Estimation strategy II

- Define loss function:

$$\ell(\boldsymbol{\pi}, \boldsymbol{T}, f|\boldsymbol{\mu}) = \sum_{x=0}^{N} L\left(\mu_g(x), \mu_x\right),$$

where μ_x is the observed mortality at age x and μ_g is the hazard rate of the EPI model.

- We found that $L(\mu, \nu) = (\log(\mu) - \log(\nu))^2$ is a good choice. We get

$$\ell(\boldsymbol{\pi}, \boldsymbol{T}, f|\boldsymbol{\mu}) = \sum_{x=0}^{N} (\log(\mathcal{C}(x|\boldsymbol{\pi}, \boldsymbol{T}, f)) + \log(\mu_f(x)) - \log(\mu_x))^2,$$

with the correction factor

$$\mathcal{C}(x|\boldsymbol{\pi}, \boldsymbol{T}, f) = \frac{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s) ds\right) \boldsymbol{t}}{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s) ds\right) \boldsymbol{e}}.$$

- EPI model will have a lower likelihood than the one arising from the EM algorithm!

Example: Danish female mortality

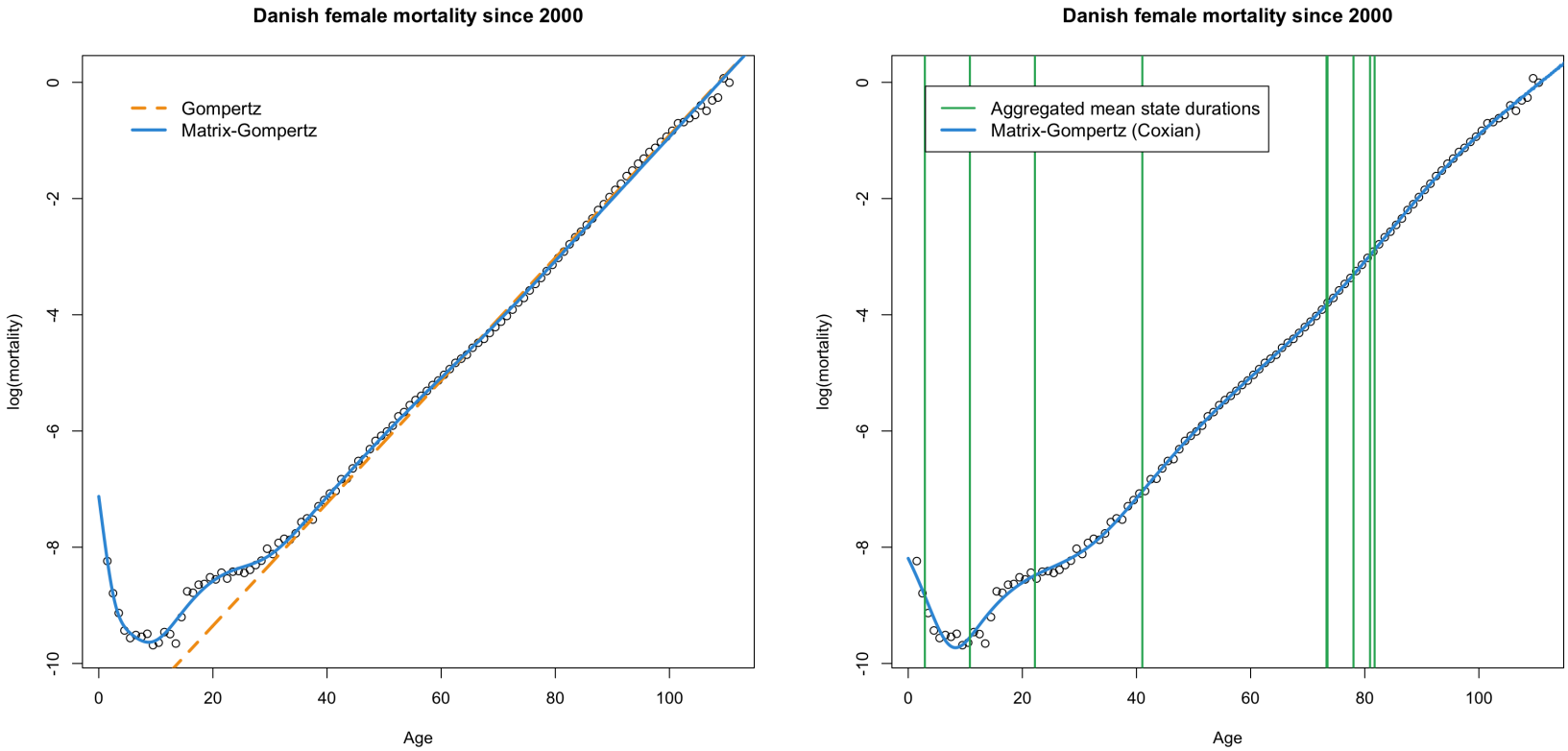


Figure: Fitted IPH distributions to Danish female mortality data.

The probability to reach last state $p = 10$ before death is 0.299.

Example: Japan vs USA female mortality

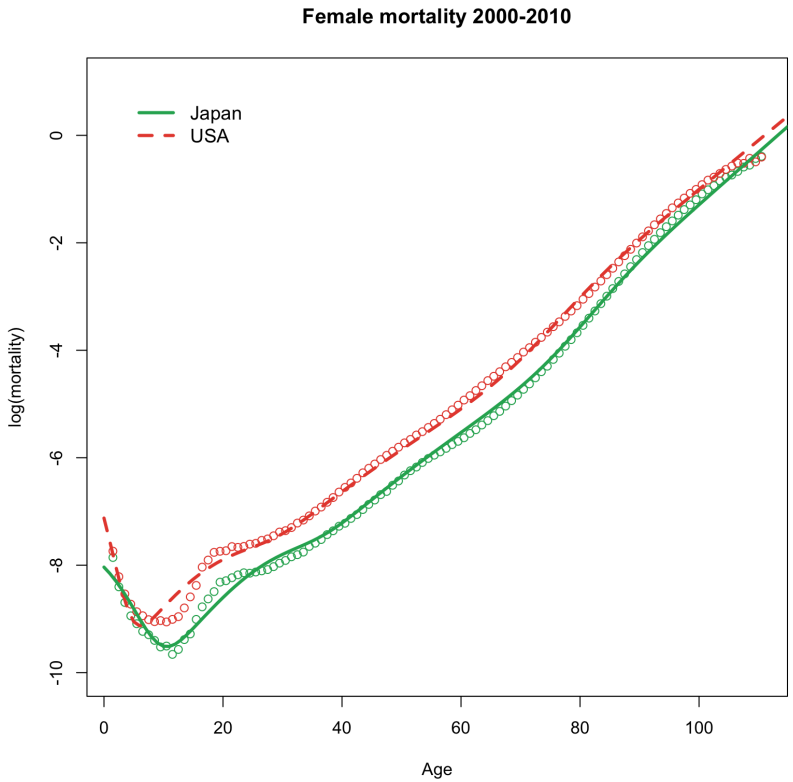


Figure: PI model applied to country as a covariate.

Regression parameters: $\boldsymbol{\beta} = (0.91)$, $\boldsymbol{\theta} = (2.28, -0.07)$. (USA=1,Japan=0)

Time as covariate

- Classical model Lee-Carter (**non-parametric**):

$$\log(\mu_{x,t}) = a_x + b_x k_t + \epsilon_{x,t},$$

where $\epsilon_{x,t}$ are Gaussian random variables. The a_x term is estimated as the average log-mortality over time at each age x, and then b_x and k_t are computed from a singular value decomposition of $\log(\mu_{x,t}) - a_x$.

- EPI model is **parametric**. As x grows,

$$\log(\mu_{x,t}) \approx a + b_{x,t} + k_t,$$

where $a = \log(c)$ is a constant which depends on the parameters $\boldsymbol{\pi}$ and $\boldsymbol{T}$, $b_{x,t} = \log(\lambda(x; \exp(\theta_0 + \theta_1 t^*))) = \exp(\theta_0 + \theta_1 t^*) \log(x)$, and $k_t = \beta_1 t^*$.

- For smaller x:

$$\log(\mu_{x,t}) = b_{x,t} + \log\left(\boldsymbol{\pi} \exp\left(\int_0^x \exp(b_{s,t} + k_t) ds\, \boldsymbol{T}\right) \boldsymbol{t}\right) - \log\left(\boldsymbol{\pi} \exp\left(\int_0^x \exp(b_{s,t} + k_t) ds\, \boldsymbol{T}\right) \boldsymbol{e}\right).$$

Danish females

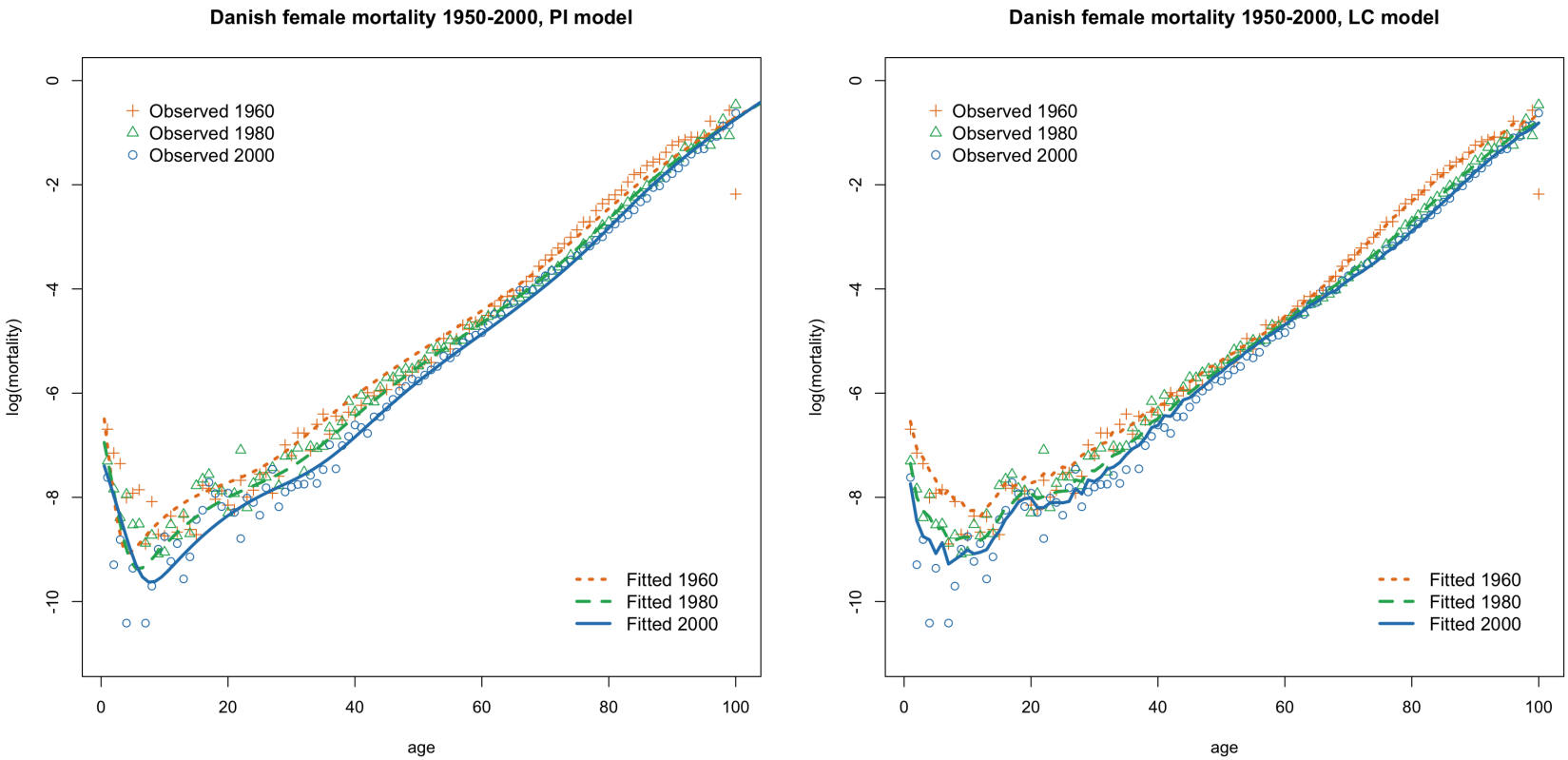


Figure: PI model using time as a covariate, plotted for 1960, 1980 and 2000.

Regression parameters: $\boldsymbol{\beta} = (-22.54) \quad \boldsymbol{\theta} = (1.92, 2.94)$. LC overfits!

PH Mixture-of-Experts

- Define the mapping
$$\boldsymbol{\pi} : D \subset \mathbb{R}^d \to \Delta^{p-1}\,,$$
where $\Delta^{p-1} = \{(\pi_1, \cdots, \pi_p) \in \mathbb{R}^p \mid \sum_k \pi_k = 1 \text{ and } \pi_k \geq 0 \text{ for all } k\}$ is the standard $(p-1)$-simplex.
- For any given $\boldsymbol{x} \in \mathbb{R}^d$, we endow the process with the initial probabilities
$$\mathbb{P}(J_0 = k) = \pi_k(\boldsymbol{x}) := (\boldsymbol{\pi}(\boldsymbol{x}))_k\,, \quad k = 1, \ldots, p\,,$$
and $\mathbb{P}(J_0 = p+1) = 0$.
- Note that
$$Y = \inf\{t > 0 : J_t = p+1\}\,,$$
satisfies that
$$Y \sim \text{IPH}(\boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{T}, \lambda) \quad \Leftrightarrow \quad J_0 \sim \boldsymbol{\pi}(\boldsymbol{x})\,.$$

Definition

Let $\boldsymbol{X}$ be a d-dimensional vector of covariates. Then we say that
$$Y|\,\boldsymbol{X} \sim \text{IPH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}, \lambda)$$
is a phase-type mixture-of-experts (PH-MoE) model.

Properties

- Mixture of IPH distributions:

$$\mathbb{P}(Y > y | \boldsymbol{X} = \boldsymbol{x}) = \sum_{k=1}^{p} \mathbb{P}(Y > y | J_0 = k) \pi_k(\boldsymbol{x}) \,.$$

- Marginals are always IPH:

Proposition

Let $\boldsymbol{X}$ be a random vector in a convex $D \subset \mathbb{R}^d$. Then the PH-MoE model has marginal distribution given by

$$IPH(\boldsymbol{\pi}(\boldsymbol{x}^*), \boldsymbol{T}, \lambda) \,,$$

for some $\boldsymbol{x}^* \in D$. In fact, $\boldsymbol{\pi}(\boldsymbol{x}^*) = \mathbb{E}(\boldsymbol{\pi}(\boldsymbol{X}))$.

- Hence, we use it in its conditional form.

Softmax parametrization

Definition

We say that the PH-MoE model with initial probabilities $\boldsymbol{\pi}(\boldsymbol{X};\boldsymbol{\alpha}) = (\pi_k(\boldsymbol{X};\boldsymbol{\alpha}))_{k=1,\ldots,p}$ given by

$$\pi_k(\boldsymbol{X};\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{X}^\top \boldsymbol{\alpha}_k)}{\sum_{j=1}^{p} \exp(\boldsymbol{X}^\top \boldsymbol{\alpha}_j)}\,, \quad k = 1,\ldots,p\,, \tag{8}$$

satisfies the softmax parametrization. Here, $\boldsymbol{\alpha}_k \in \overline{\mathbb{R}}^d$, $k = 1,\ldots,p$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top,\ldots,\boldsymbol{\alpha}_p^\top)^\top \in \overline{\mathbb{R}}^{(p\times d)}$.

For any $k, j \in \{1,\ldots,p\}$,

$$\log\left(\frac{\pi_k(\boldsymbol{X};\boldsymbol{\alpha})}{\pi_j(\boldsymbol{X};\boldsymbol{\alpha})}\right) = \boldsymbol{X}^\top(\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_j) = \sum_{i=1}^{d} X_i(\alpha_{ki} - \alpha_{ji})\,.$$

Denseness properties

Definition

Let $W_1, \ldots, W_n$ be positive and continuous random variables having otherwise arbitrary distributions, and let $\eta \in \{1, \ldots, n\}$ be a multinomial random variable, such that

$$W_i \perp\!\!\!\perp W_j \,, \ \forall i \neq j \,, \quad \text{and} \quad W_i \perp\!\!\!\perp_{\boldsymbol{X}} \eta \,, \ \forall i \,.$$

and such that $\boldsymbol{X}$ contains at least an intercept. Then we say that $W_\eta | \boldsymbol{X}$ follows a multinomial mixture distribution.

Proposition

Let $W | \boldsymbol{X}$ follow a multinomial mixture distribution. Then there exist PH-MoE models $(Y_m | \boldsymbol{X})_{m \geq 0}$ such that

$$Y_m | \boldsymbol{X} \xrightarrow{d} W | \boldsymbol{X} \,, \quad m \to \infty \,.$$

Moreover, the softmax parametrization may be chosen.

Denseness properties II

Definition

Let $\mathcal{A}$ be the set of possible values of the covariates $\boldsymbol{X}$. A severity regression model is the set of laws of

$$Y \,|\, \boldsymbol{X} = \boldsymbol{x}\,, \quad \boldsymbol{x} \in \mathcal{A}\,.$$

Condition

A regression model is said to satisfy the tightness and Lipschitz conditions on $\mathcal{A}$ if $\{\mathbb{P}(Y \in \cdot \,|\, \boldsymbol{X} = \boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{A}}$ is a tight family of distributions, and for each $y \geq 0$, the function $\boldsymbol{x} \mapsto \mathbb{P}(Y \leq y \,|\, \boldsymbol{X} = \boldsymbol{x})$ is Lipschitz continuous in $\mathcal{A}$.

Proposition

Let a regression model satisfy the tightness and Lipschitz conditions on $\mathcal{A} = \{1\} \times [a, b]^{d-1}$, $a, b \in \mathbb{R}$. Then there exists a sequence of PH-MoE regression models converging uniformly weakly to it.

Proof: Similar to Fung et al. (2019).

Estimation

- Completely observed likelihood:

$$\begin{aligned}
&\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}|\boldsymbol{z}, \overline{\boldsymbol{x}}) \\
&= \prod_{i=1}^{N} \mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}|z_i, \boldsymbol{X} = \boldsymbol{x}_i) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)} \prod_{k=1}^{p} \prod_{l \neq k} t_{kl}{}^{N_{kl}(\boldsymbol{x}_i)} \exp(-t_{kl} V_k(\boldsymbol{x}_i)) \prod_{k=1}^{p} t_k{}^{N_k(\boldsymbol{x}_i)} \exp(-t_k V_k(\boldsymbol{x}_i)) \\
&= \cdots \\
&= \left(\prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)} \right) \prod_{k=1}^{p} \prod_{l \neq k} t_{kl}{}^{N_{kl}} \exp(-t_{kl} V_k) \prod_{k=1}^{p} t_k{}^{N_k} \exp(-t_k V_k)\,,
\end{aligned}$$

with

$$N_{kl} := \sum_{i=1}^{N} N_{kl}(\boldsymbol{x}_i)\,, \quad V_k := \sum_{i=1}^{N} V_k(\boldsymbol{x}_i)\,, \quad N_k := \sum_{i=1}^{N} N_k(\boldsymbol{x}_i)\,.$$

Estimation II

- The E- and M-steps are very similar (although not identical) to the PI case.
- One distinction is the optimization of the $\boldsymbol{\pi}$ function:

$$\hat{\boldsymbol{\pi}}(\cdot) = \underset{\boldsymbol{\pi}(\cdot)\in\Delta^{p-1}}{\arg\max} \left(\prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{\mathbb{E}(B_k(\boldsymbol{x}_i)|Z=z_i,\boldsymbol{X}=\boldsymbol{x}_i)} \right),$$

  where as before Δ^{p-1} is the standard $(p-1)$-simplex.
- The latter step corresponds to a weighted multinomial regression.
- Then finally optimize the inhomogeneity function as before.

Asymptotic normality

Theorem

Let λ, $\boldsymbol{\eta} := (\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{T})$ *be such that the log-density*

$$y \mapsto \log \left[\boldsymbol{\pi}(\boldsymbol{\alpha}) \exp \left(\int_0^y \lambda(s; \boldsymbol{\theta}) ds \boldsymbol{T} \right) \boldsymbol{t}\, \lambda(x; \boldsymbol{\theta}) \right], \quad y > 0,$$

satisfies standard assumptions from Lehmann and Casella (2006). As $n \to \infty$,

1. *There exist consistent solutions* $\hat{\boldsymbol{\eta}}_n$ *to the likelihood equations.*
2. *The following convergence holds:*

$$\sqrt{n}\,(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}),$$

   where $\boldsymbol{\mathcal{I}}$ *is the information matrix.*

3. *The* j*-th parameter is asymptotically efficient:*

$$\sqrt{n}\,(\hat{\eta}_{jn} - \eta_j) \xrightarrow{d} \mathcal{N}(0, [\boldsymbol{\mathcal{I}}^{-1}]_{jj}).$$

In practice, we use partial likelihood.

Transforms

- Different tail behaviour can arise for subpopulations.
- We may use global models.
- We may use the following:

Definition

We say that a PH-MoE model is semi-composite if its intensity function is of the form

$$\lambda(t) = \begin{cases} \lambda_1(t)\,, & t \leq y_0\,, \\ \lambda_2(t)\,, & t > y_0\,, \end{cases}$$

for any two intensities λ_1, λ_2.

Proposition

For a semi-composite PH-MoE we have that

$$\overline{F}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \begin{cases} \boldsymbol{\pi}(\boldsymbol{x}) \exp(\boldsymbol{T} g_1^{-1}(y))\boldsymbol{e}\,, & y \leq y_0\,, \\ \boldsymbol{\pi}(\boldsymbol{x}) \exp((g_2^{-1}(y) + g_1^{-1}(y_0) - g_2^{-1}(y_0))\boldsymbol{T})\boldsymbol{e}\,, & y > y_0\,. \end{cases}$$

Simulation

We simulate 2000 observations divided into 4 groups of size 500, each having distributions as follows:

$$\text{Group A: } Y_i \sim \Gamma(\text{shape} = 1, \text{scale} = 3), \quad \text{Group B: } Y_i \sim \Gamma(\text{shape} = 3, \text{scale} = 9),$$
$$\text{Group C: } Y_i \sim \Gamma(\text{shape} = 1, \text{scale} = 9), \quad \text{Group D: } Y_i \sim \Gamma(\text{shape} = 3, \text{scale} = 3).$$

Estimated initial probabilities of a $p = 3$ homogeneous PH-MoE:

$$\boldsymbol{\pi}(\text{Group A}) = (0.976,\ 0.000,\ 0.013,\ 0.000,\ 0.011),$$
$$\boldsymbol{\pi}(\text{Group B}) = (0.000,\ 0.007,\ 0.000,\ 0.993,\ 0.000),$$
$$\boldsymbol{\pi}(\text{Group C}) = (0.328,\ 0.094,\ 0.324,\ 0.173,\ 0.080),$$
$$\boldsymbol{\pi}(\text{Group D}) = (0.000,\ 0.000,\ 0.976,\ 0.000,\ 0.024).$$

Simulation II

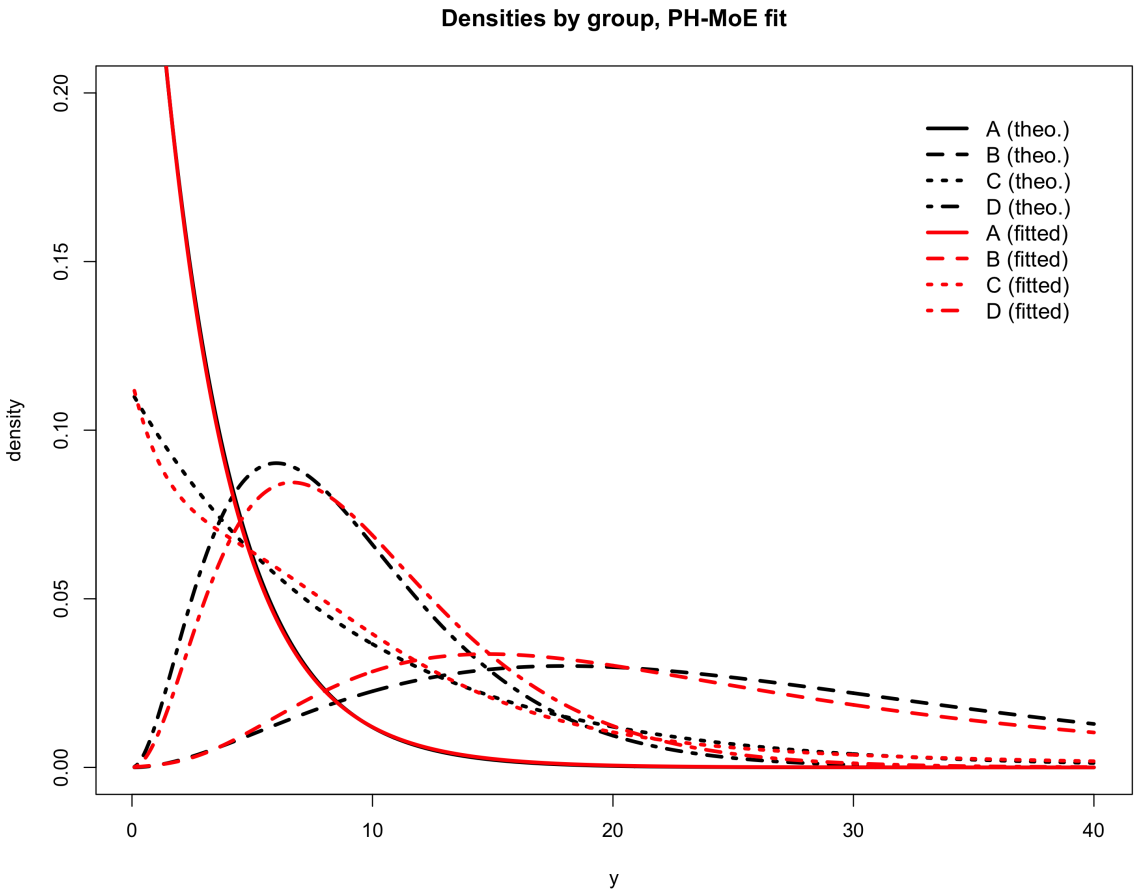


Figure: Fitted densities for each group of the simulated data, for the PH-MoE and GLM models.

Simulation II

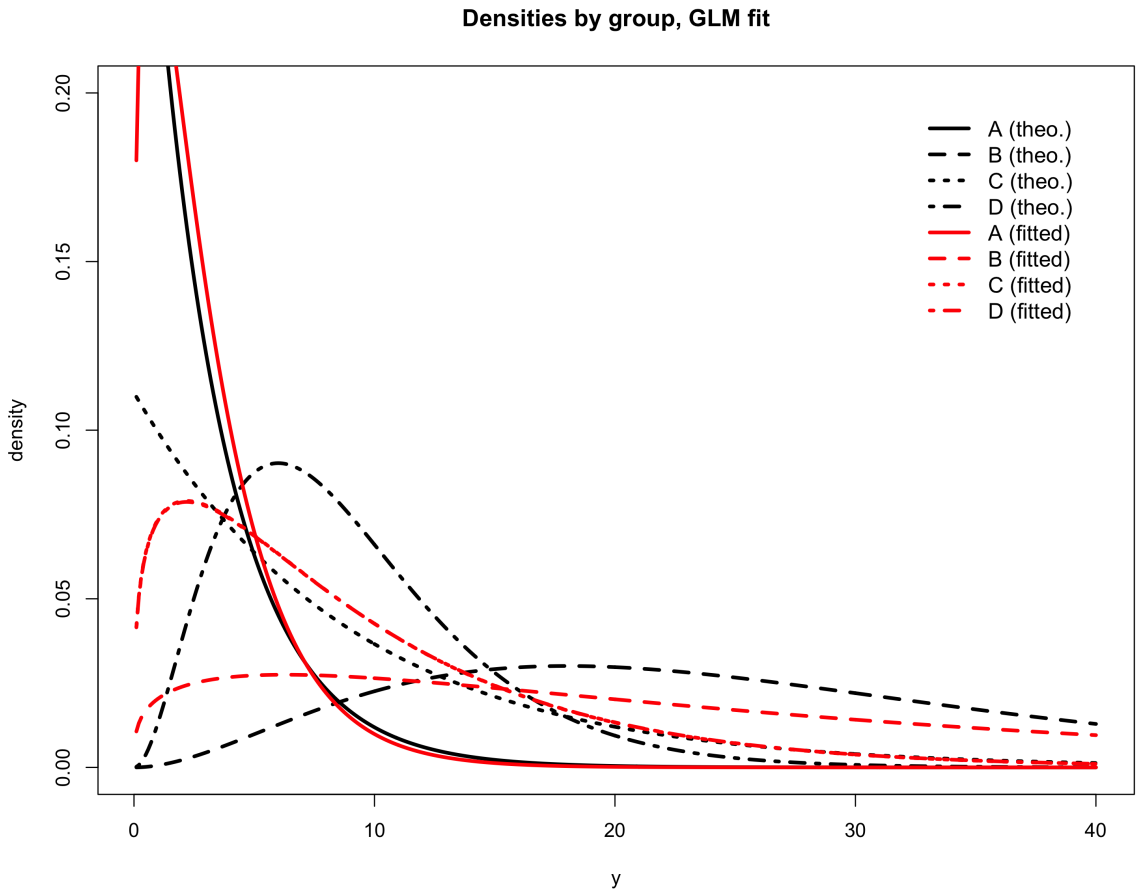


Figure: Fitted densities for each group of the simulated data, for the PH-MoE and GLM models.

FreMTPL severities

- French Motor Third Party Liability datasets `freMTPLfreq` and `freMTPLsev`. Total of 413, 169 motor insurance policies, with 15, 390 claims.
- For numerical reasons, we also divide the claim size by 10^4.
- $p = 5$ again. We select the following covariates:
  1. `Power`: The power of the car, an ordered categorical variable with values: d, e, f, g, h, i, j, k, l, m, n, o.
  2. `Region`: The policy region in France, based on the 1970-2015 classification. Possible values associated with the excesses are: Aquitaine, Basse-Normandie, Bretagne, Centre, Haute-Normandie, Ile-de-France, Limousin, Nord-Pas-de-Calais Pays-de-la-Loire, Poitou-Charentes.
- We consider excesses above a threshold: PH-MoE routine can be slow for large p (say, above 10), n (in the tens of thousands) and d (more than 20 covariates).

FreMTPL severities II

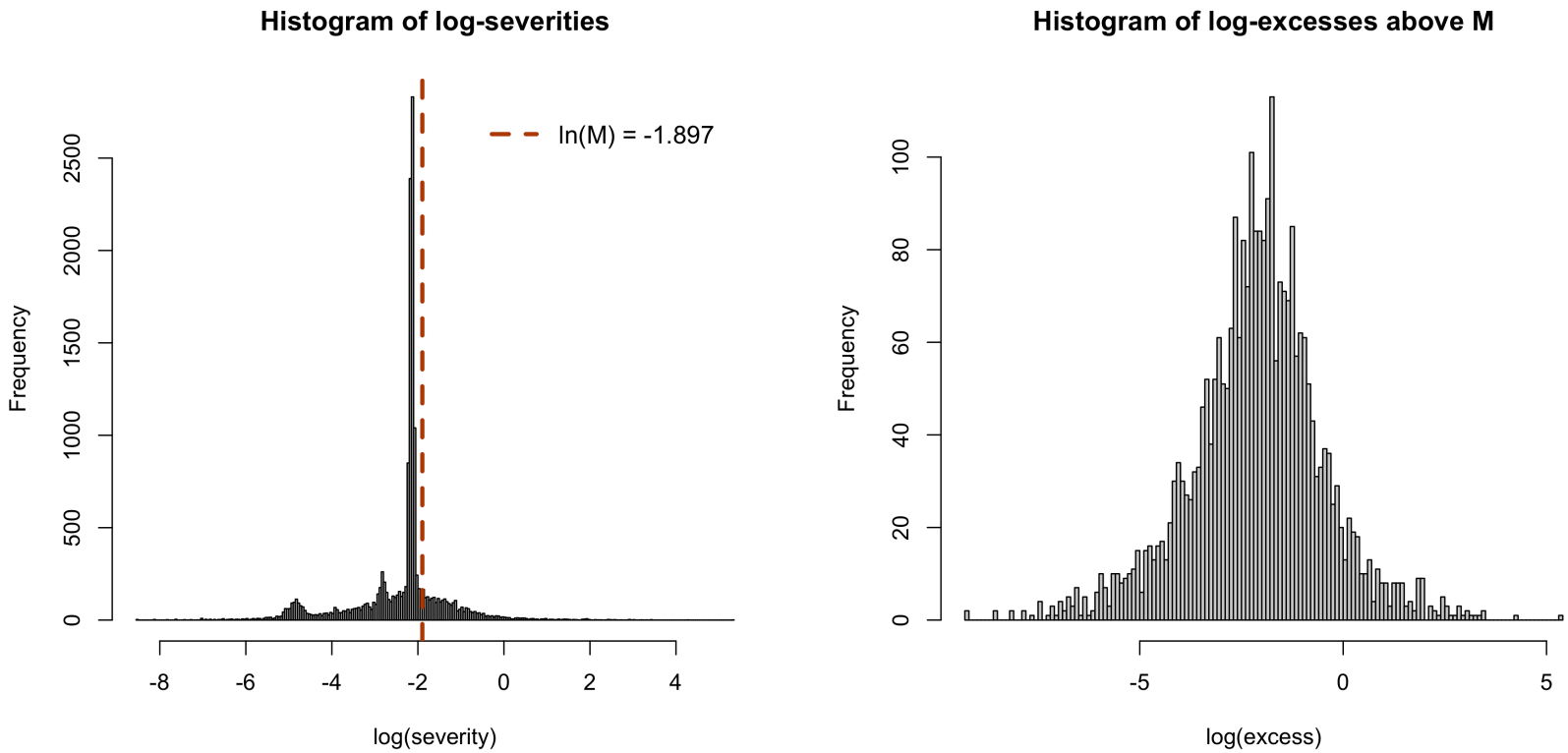


Figure: French MTPL full data.

FreMTPL severities II

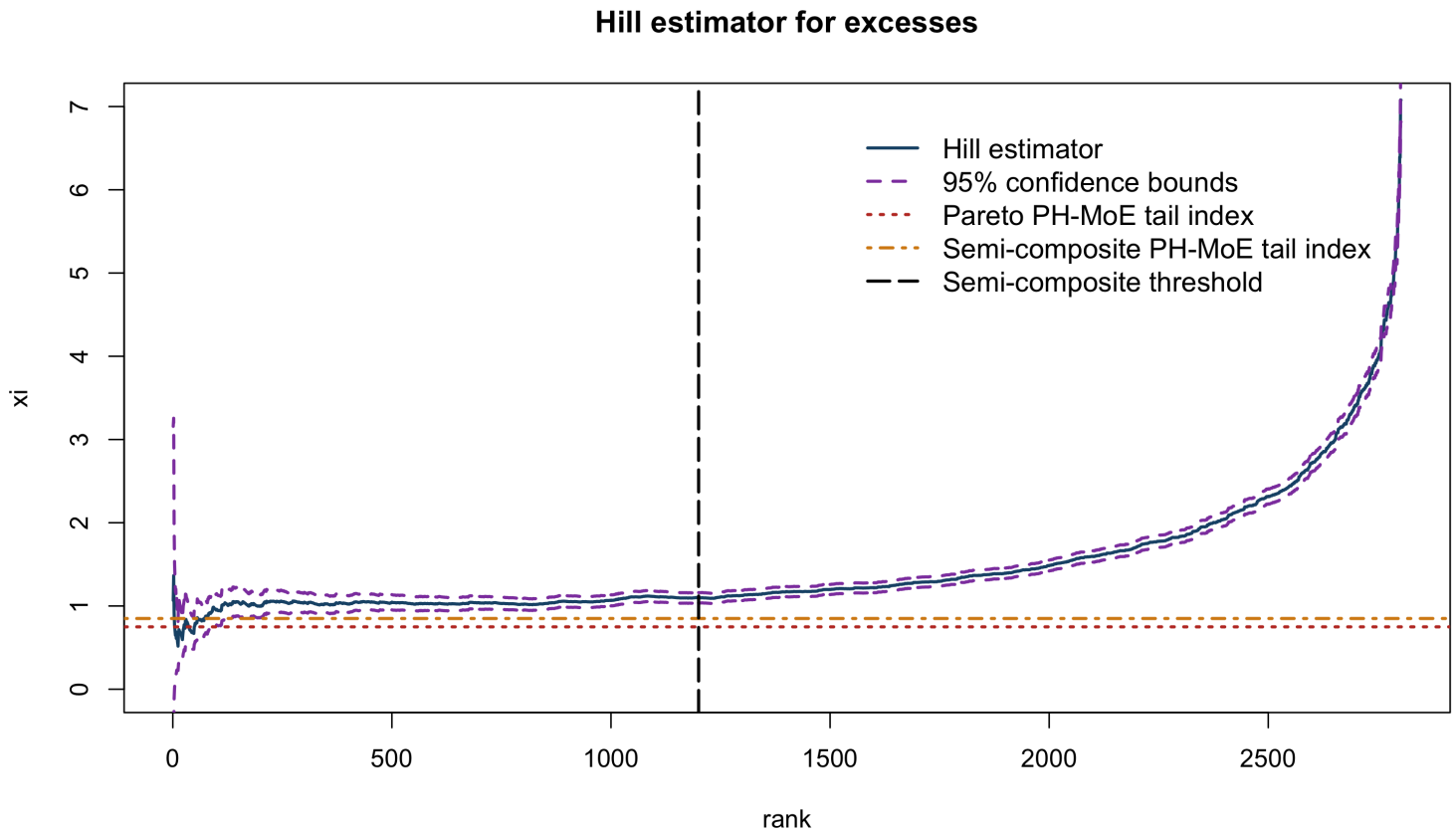


Figure: French MTPL full data.

FreMTPL severities III

	Pareto PH-MoE	Semi-composite PH-MoE	Log-normal
Log Likelihood	753.74	752.48	607.22
Degrees of freedom	110	110	22
AIC	**−1,287**	−1,284	−1,170
BIC	−634	−631	**−1,039**
Num. obs.	2,804	2,804	2,804

Table: Summary for PH-MoE and log-normal regression models for the freMTPL dataset.

FreMTPL severities IV

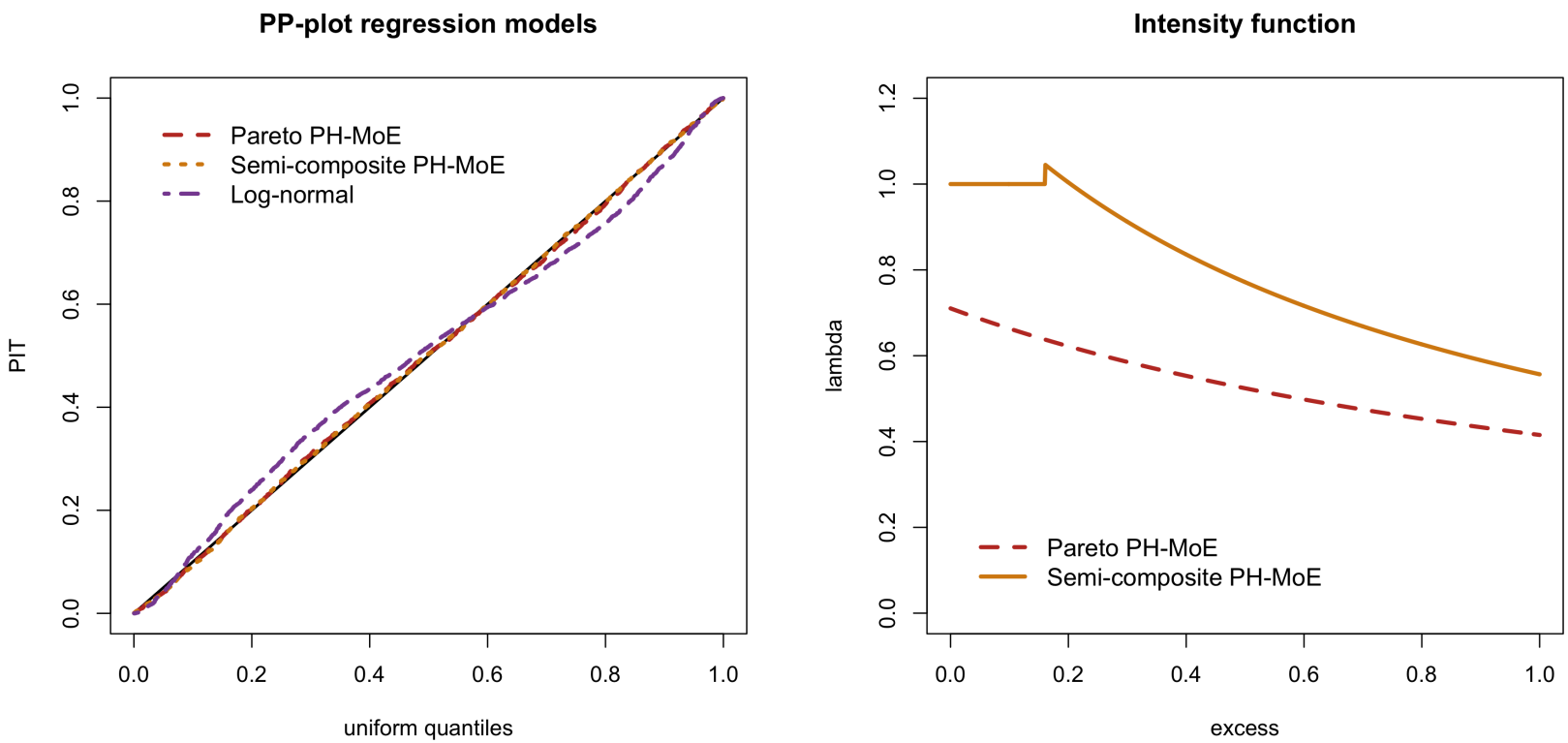


Figure: PP-plots for the fitted regression models (left panel), and fitted intensity functions for the PH-MoE models (right panel).

Conclusion and outlook

- If your data has an odd shape or tail, phase-type regression likely helps.
- Easy to interpret.
- Most software either already or soon-to-be in CRAN (package `matrixdist`).
- Speed gains required for regularization and automatic variable selection techniques.
- Multivariate case is currently being investigated.