Alternative estimation framework: from single population to cause-of-death

Antoine Burg, Université Paris Dauphine (CEREMADE), SCOR

Joint work with Christophe Dutang, LJK, Grenoble INP - UGA.

SCOR Chair on mortality research, 2024/11/07

Plan

1. New framework for single population
2. How to derive multi-populations model
3. Multivariate extension for Causes-of-Death

Plan

1. New framework for single population
2. How to derive multi-populations model
3. Multivariate extension for Causes-of-Death

Generalized Age Period Cohort framework

General framework for single population following [AMV+18] (StMoMo R package).

- the *random component*: The number of deaths $D_{x,t}$ is assumed to follow a Poisson or a Binomial distribution:

$$D_{x,t} \sim \mathcal{P}(E^c_{x,t} m_{x,t}) \quad \text{or} \quad D_{x,t} \sim \mathcal{B}(E^0_{x,t} q_{x,t}),$$

- the *systematic component*: the effects of age x, calendar year t and year-of-birth (cohort) $c = t - x$ are captured through a *predictor* $\eta_{x,t}$ defined as:

$$\eta_{x,t} = \alpha_x + \sum_i \beta^i_x \kappa^i_t + \beta^0_x \gamma_{t-x}.$$

- The *link function* g associating the random and the systematic components:

$$g\left(\mathbb{E}\left(\frac{D_{x,t}}{E_{x,t}}\right)\right) = \eta_{x,t}.$$

- The *set of parameter constraints*: as most stochastic models are not identifiable, some constraints may be needed to ensure uniqueness of the parameters α_x, κ_t, γ_{t-x}.

Generalized Age Period Cohort framework - current limitations

Designed for a single-population

- Univariate Poisson or Binomial assumption;
- Multi-populations models are built based on a 2-steps procedure;
- Fail to handle multivariate distributions (e.g. for Causes-of-deaths modelling).

Relies on Maximum Likelihood Estimation (MLE)

- Involves Iterative Weighted Least Squares algorithm (no closed-form formulas);
- Can be computationally intensive.

Leveraging categorical variables

- **Assumptions:** age and time are two categorical variables with dummy structure
  age ranges in $[x_1; x_{d_1}]$ and is encoded by $\left(z_x^{(1)}\right)_{x\in[x_1;x_{d_1}]}$: $z_{x_j}^{(1)} = 1_{x=x_j}$;
  time ranges in $[t_1; t_{d_2}]$ and is encoded by $\left(z_x^{(2)}\right)_{t\in[t_1;t_{d_2}]}$: $z_{t_k}^{(2)} = 1_{t=t_k}$;
- **Systematic component** $\eta_{x,t}$ can be rewritten as

$$
\begin{aligned}
\eta_{x,t} = \theta_0 &+ \sum_{x'=x_1\dots x_{d_1}} z_{x'}^{(1)}\theta_{x'} + \sum_{t'=t_1\dots t_{d_2}} z_{t'}^{(2)}\theta_{t'} && \text{intercept and single effect} \\
&+ \sum_{x'=x_1\dots x_{d_1}} \sum_{t'=t_1\dots t_{d_2}} z_{x'}^{(1)} z_{t'}^{(2)} \theta_{x',t'} && \text{double effect.}
\end{aligned}
$$

$\theta = (\theta_0, \theta_{x_1}, \dots, \theta_{x_{d_1}}, \theta_{t_1}, \dots, \theta_{t_{d_2}}, \theta_{1,1}, \dots, \theta_{x_{d_1} t_{d_2}})$ parameters vector

Leveraging categorical variables

Theorem

Following results from [BDR20, BDR22], we derive the alternative estimators:

$$\tilde{\theta} = \left(Q^\intercal Q + R^\intercal R\right)^{-1} Q^\intercal \tilde{\eta},$$

- Q is the unique matrix such that $\eta = Q\theta$;
- R a contrast matrix that ensures identifiability, $R\theta = 0$;
- $\tilde{\eta}$ is an estimator of $g(\mathbb{E}[D_{x,t}])$.

Corollary

If $D_{x,t} \sim \mathcal{B}(E^0_{x,t}, q_{x,t})$ with a logit link function, we have the single categorical estimator:

$$\forall x_j, t_k, \quad \tilde{\eta}_{x_j,t_k} = \text{logit}\left(\overline{q^{(x_j,t_k)}}\right), \quad \overline{q^{(x_j,t_k)}} = \frac{1}{m_{j,k}} \sum_x \sum_t q_{x,t} z_{x_j} z_{t_k}, \quad m_{j,k} = \sum_x \sum_t z_{x_j} z_{t_k}.$$

Example with x= 50..51 and t = 2023..2024 ($d_1 = d_2 = 2$):

$$\eta = \begin{pmatrix} \eta_{50,2023} \\ \eta_{50,2024} \\ \eta_{51,2023} \\ \eta_{51,2024} \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix},$$

Identifying classical models - APC (M3) model

We directly identify the coefficients of the M3 model, as γ_{t-x} play the same role as double effect coefficients $\theta_{x,t}$:

$$\eta_{x,t} = \alpha_x + \kappa_t + \gamma_{t-x} \quad (M3).$$

Example with $d_1 = d_2 = 2$:

$$Q_{M3} = \begin{array}{c} x,t \\ x_1,t_1 \\ x_2,t_1 \\ x_1,t_2 \\ x_2,t_2 \end{array} \begin{array}{c} \begin{array}{ccccccc} \alpha_{x_1} & \alpha_{x_2} & \kappa_{t_1} & \kappa_{t_2} & \gamma_{c_{-1}} & \gamma_{c_0} & \gamma_{c_1} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$R_{M3} = \begin{array}{c} \textit{constraints} \\ \sum_t \kappa_t = 0 \\ \sum_c \gamma_c = 0 \\ \sum_c c\gamma_c = 0 \end{array} \begin{array}{c} \begin{array}{ccccccc} \alpha_{x_1} & \alpha_{x_2} & \kappa_{t_1} & \kappa_{t_2} & \gamma_{c_{-1}} & \gamma_{c_0} & \gamma_{c_1} \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & c_{-1} & c_0 & c_1 \end{pmatrix} \end{array}$$

Identifying classical actuarial mortality models

We identify as well the coefficients of the M5 model:

$$\eta_{x,t} = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) \quad (M5).$$

Example with $d_1 = d_2 = 2$:

$$Q_{M5} = \begin{array}{c} x,t \\ x_1,t_1 \\ x_2,t_1 \\ x_1,t_2 \\ x_2,t_2 \end{array} \begin{array}{c} \begin{array}{cccc} \kappa_{t_1}^{(1)} & \kappa_{t_2}^{(1)} & \kappa_{t_1}^{(2)} & \kappa_{t_2}^{(2)} \end{array} \\ \begin{pmatrix} 1 & 0 & x_1 - \bar{x} & 0 \\ 1 & 0 & x_2 - \bar{x} & 0 \\ 0 & 1 & 0 & x_1 - \bar{x} \\ 0 & 1 & 0 & x_2 - \bar{x} \end{pmatrix} \end{array}, \quad R_{M5} = 0.$$

And similarly for

$$\eta_{x,t} = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_{t-x}^{(3)} \qquad (M6)$$

$$\eta_{x,t} = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}\left[(x - \bar{x})^2 - \hat{\sigma}_x^2\right] + \gamma_{t-x}^{(3)} \qquad (M7).$$

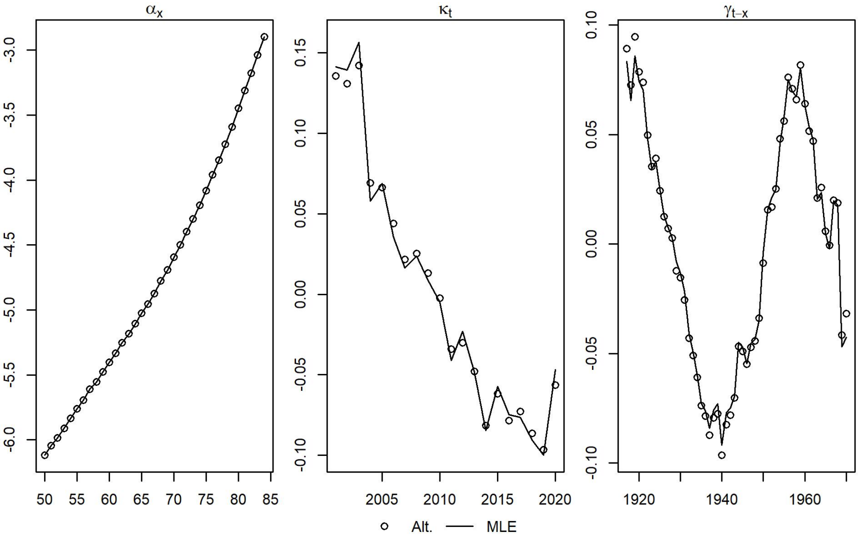
Numerical examples

- Country: France
- Age band: 50-84
- Period: 2001-2020
- Sex: Females
- Estimators: Maximum Likelihood Estimators (from StMoMo package) Vs Alternative Estimators $\tilde{\theta}$
- Source: Human Mortality Database (INSEE)

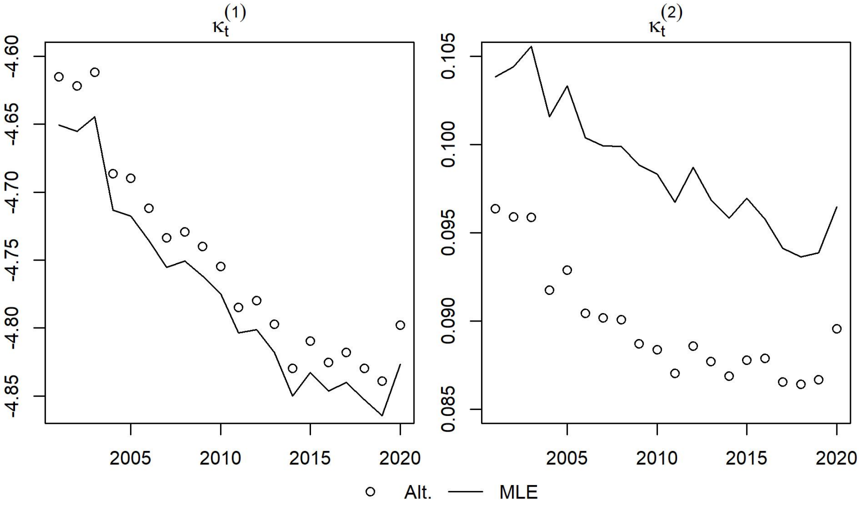
Numerical examples - Females, 50-84, 2001-2020

	Deviance	AIC	MSE	MAE	MAPE (%)
MLE.M3	8.41E+02	7.86E+03	6.85E-08	1.74E-04	1.75
Alt.M3	9.29E+02	7.94E+03	1.13E-07	1.99E-04	1.73
MLE.M5	5.06E+04	5.75E+04	6.20E-06	1.61E-03	1.42
Alt.M5	6.53E+04	7.22E+04	1.36E-05	1.93E-03	1.22
MLE.M6	1.84E+03	8.83E+03	1.17E-07	2.43E-04	2.70
Alt.M6	2.26E+03	9.24E+03	2.87E-07	3.11E-04	2.62
MLE.M7	7.37E+02	7.76E+03	6.30E-08	1.68E-04	1.63
Alt.M7	7.50E+02	7.77E+03	7.64E-08	1.76E-04	1.62

Numerical examples - M3 fit (Females, 50-84, 2001-2020)



Numerical examples - M5 fit (Females, 50-84, 2001-2020)



Plan

1. New framework for single population
2. How to derive multi-populations model
3. Multivariate extension for Causes-of-Death

Possibility to build multi-populations models

We try here a multi-population model following ideas from [LL05]. Two populations indexed with g_1 and g_2, with linear systematic component (here M3) and a common factor $K(t)$ such that

$$\begin{cases} \eta_{x,t,g_1} = \alpha_{x,g_1} + K(t) + \kappa_{t,g_1} + \gamma_{t-x,g_1} \\ \eta_{x,t,g_2} = \alpha_{x,g_2} + K(t) + \kappa_{t,g_2} + \gamma_{t-x,g_2} \end{cases}$$

Problem equivalent to

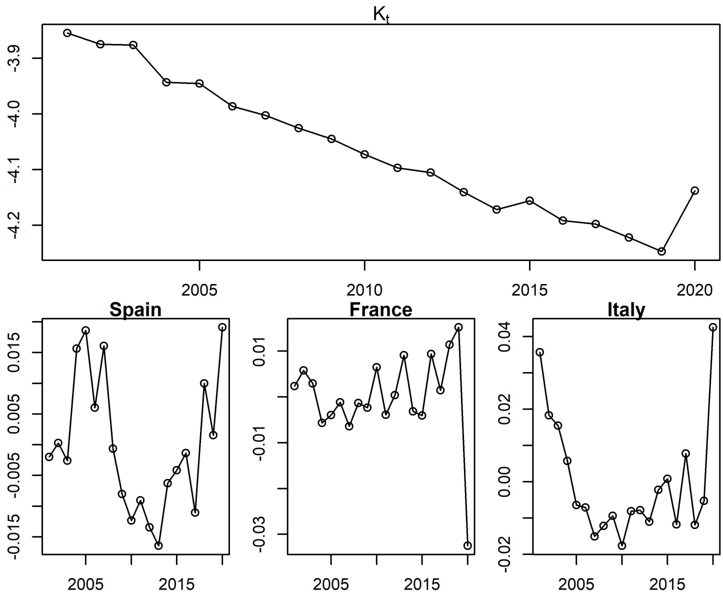
- Adding a third categorical variable for country;
- Adding on observation of a country g which is the sum of the sub-populations g_1 and g_2;
- Reference levels of $\alpha_{x,g}$ and $\gamma_{t-x,g}$ set to 0.

Possibility to build multi-populations models

Example with $d_1 = d_2 = 2$:

$$Q = \begin{array}{cc} & \begin{array}{cccc} \boldsymbol{\theta}_{g_1} & \boldsymbol{\theta}_{g_2} & K_{t_1} & K_{t_2} \end{array} \\ \begin{array}{c} g_1 \\ \\ \\ \\ g_2 \\ \\ \\ \\ g \\ \\ \\ \\ \end{array} & \begin{pmatrix} & 0 & 1 & 0 \\ & 0 & 1 & 0 \\ Q_{M3} & 0 & 0 & 1 \\ & 0 & 0 & 1 \\ 0 & & 1 & 0 \\ 0 & & 1 & 0 \\ 0 & Q_{M3} & 0 & 1 \\ 0 & & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}, \quad R = \begin{pmatrix} R_{M3} & 0 & 0 & 0 \\ 0 & R_{M3} & 0 & 0 \end{pmatrix}.$$

Possibility to build multi-populations models



Plan

1. New framework for single population

2. How to derive multi-populations model

3. Multivariate extension for Causes-of-Death

Multivariate GAPC framework

Generalization of the GAPC framework based on multinomial distribution:

- the *random component*: The random vector of death counts $\boldsymbol{D}_{x,t}$ is assumed to follow a multinomial distribution

$$\left(D_{x,t}^{(1)},\dots,D_{x,t}^{(d)}\right)\sim\mathcal{M}_d\left(E_{x,t}^{0},\boldsymbol{q}_{x,t}=\left(q_{x,t}^{(1)},\dots,q_{x,t}^{(d)}\right)\right).$$

- the *systematic component*: the effects of age x, calendar year t and year-of-birth (cohort) $c=t-x$ are captured through a *predictor* $\boldsymbol{\eta}_{x,t}$ defined as

$$\boldsymbol{\eta}_{x,t}=(\eta_{x,t}^{(1)},\dots,\eta_{x,t}^{(d)}),\quad \eta_{x,t}^{(j)}=\alpha_x^{(j)}+\sum_i\beta_x^{i,(j)}\kappa_t^{i,(j)}+\beta_x^{(0)}\gamma_{t-x}^{(j)}.$$

  The formulation of the predictor may differ for each cause j.

- The *link function* g associating the random and the systematic components is the canonical link

$$g\left(\mathbb{E}\left(\frac{\boldsymbol{D}_{x,t}}{E_{x,t}^{0}}\right)\right)=\boldsymbol{\eta}_{x,t},\quad\text{with}\quad g:p=(p_1,\dots,p_d)\rightarrow\left(\log\frac{p_1}{p_d},\dots,\log\frac{p_{d-1}}{p_d},0\right).$$

- The *set of parameter constraints*: as most stochastic models are not identifiable, some constraints may be needed to ensure uniqueness of the estimates.

Leveraging categorical variables for multinomial distribution

In multivariate context, we show that the alternative estimators still holds:

$$\tilde{\theta} = \left(Q^\intercal Q + R^\intercal R\right)^{-1} Q^\intercal \tilde{\eta}.$$

Theorem (from [BD24])

If $\left(D_{x,t}^{(1)}, \dots, D_{x,t}^{(d)}\right) \sim \mathcal{M}_d\left(E_{x,t}^0, \boldsymbol{q}_{x,t} = \left(q_{x,t}^{(1)}, \dots, q_{x,t}^{(d)}\right)\right)$ *with canonical link function, we have the single categorical estimator:*

$$\forall x_j, t_k, \forall l \in [1; d], \quad \tilde{\eta}_{x_j,t_k}^{(l)} = \log\left(\frac{\overline{q_l^{x_j,t_k}}}{1 - \sum_{l=1}^{d} \overline{q_l^{x_j,t_k}}}\right), \quad \overline{q_l^{x_j,t_k}} = \frac{1}{m_{j,k}} \sum_x \sum_t q_{x,t}^{(l)} z_{x_j} z_{t_k}.$$

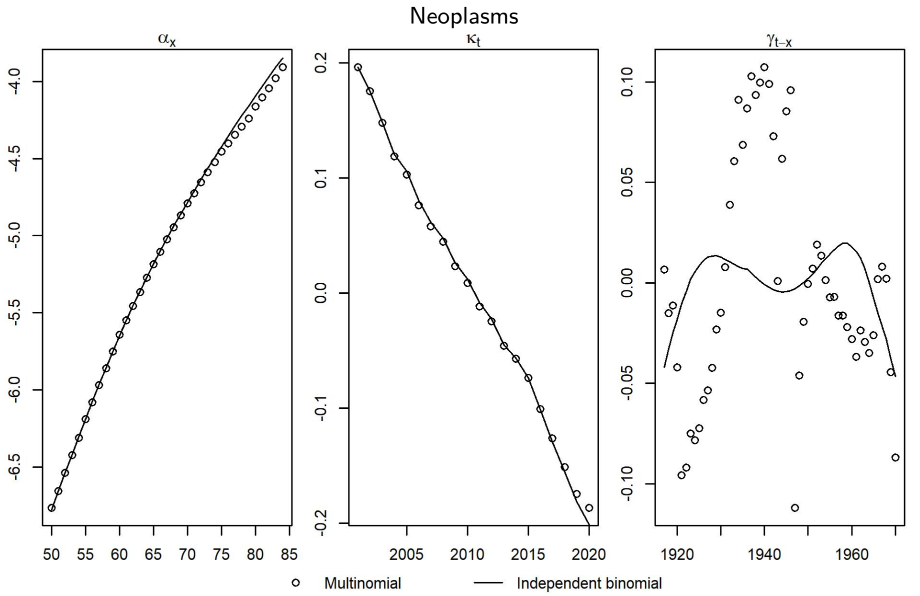
Numerical example

- Country: USA
- Age band: 50-84
- Period: 2001-2020
- Sex: Males
- Source: Human Mortality Database
- For each cause, M3 model with cause-specific parameters

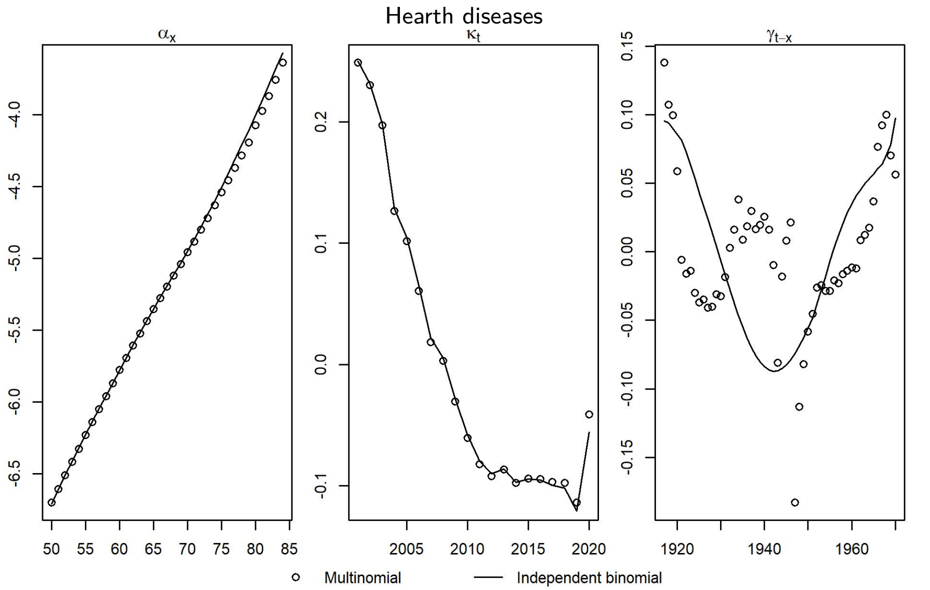
$$\eta_{x,t}^{(j)} = \alpha_x^{(j)} + \kappa_t^{(j)} + \gamma_{t-x}^{(j)}.$$

- Estimators: comparison between univariate independent single population VS multinomial assumption.

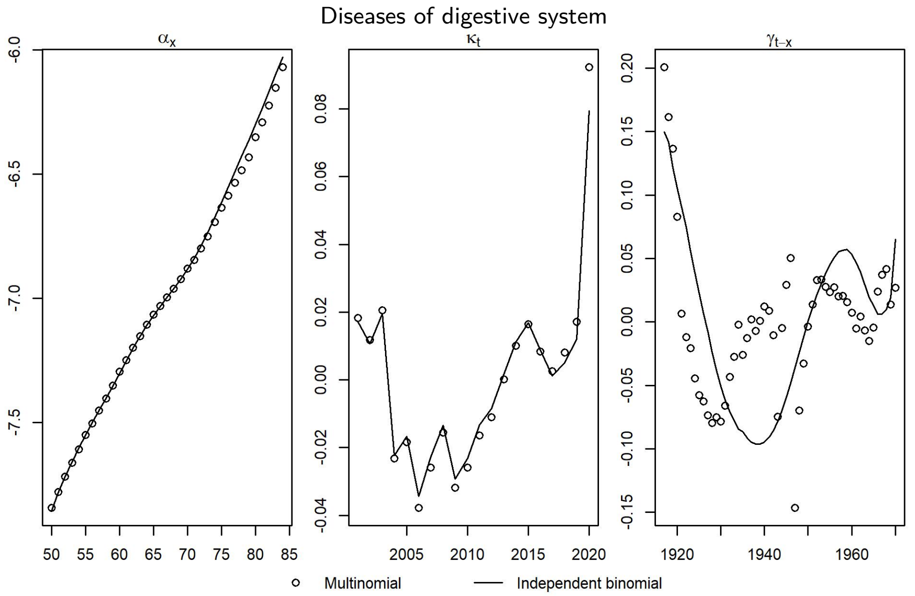
Numerical Example - US Males 50-84 2001-2020



Numerical Example - US Males 50-84 2001-2020



Numerical Example - US Males 50-84 2001-2020

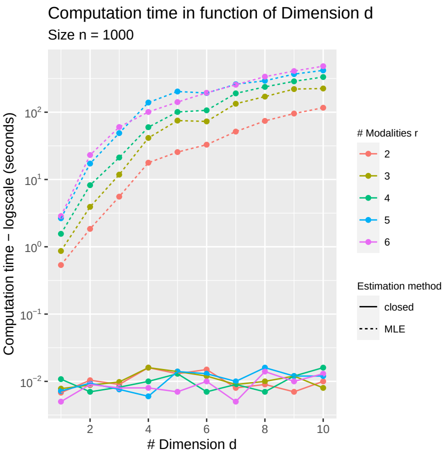


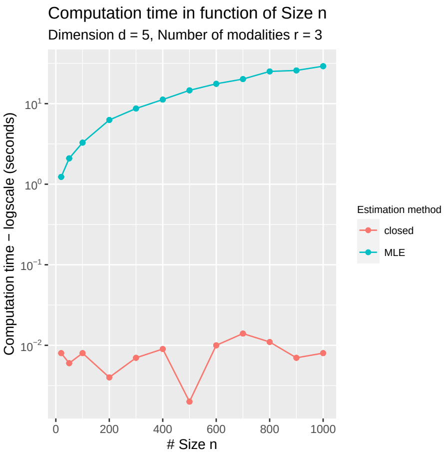
Numerical Example - US Males 50-84 2001-2020

Parameters	Deviance	AIC	MSE	MAE	MAPE (%)
Independent univariate	11883	107342	2.55E-07	1.58E-04	7,03%
Multivariate	5149	97172	8.88E-09	2.56E-05	2.14%

Life Expectancy Comparison (all-causes)

Computation time - closed form vs MLE for multinomial distribution





Key takeaways

Objectives achieved:

- One-step framework for multivariate models (e.g. causes-of-death);
- Unified and flexible framework for mortality modelling;
- Faster computation time, especially in high dimensions.

Limits:

- Only linear models - Lee-Carter model is out of scope due to the bilinear term $\beta_x \kappa_t$;
- Not MLE in general but not necessarily worse performance, depending on the metrics chosen.

Future work:

- Compare to other causes-of-deaths mortality models;
- Implement LeCam one-step procedure to improve performance of non-MLE estimators;
- Assess performance in forecasting;
- Parameters clustering with GLM-trees.

ICD10 classification

	Disease	ICD10
1	Certain infectious diseases	A00-B99
2	Neoplasms	C00-D48
3	Diseases of the blood and blood-forming organs	D50-D89
4	Endocrine, nutritional and metabolic diseases	E00-E88
5	Mental and behavioural disorders	F01-F99
6	Diseases of the nervous system and the sense organs	G00-G44, G47-H93
7	Heart diseases	I00-I51
8	Cerebrovascular diseases	G45, I60-I69
9	Other and unspecified disorders of the circulatory system	I70-I99, K64
10	Acute respiratory diseases	J00-J22, U04, U07
11	Other respiratory diseases	J30-J98
12	Diseases of the digestive system	K00-K63, K65-K92
13	Diseases of the skin	L00-M99
14	Diseases of the genitourinary system	N00-O99
15	Certains conditions originating in the perinatal period	
16	External causes	V01-Y89

Main references

Villegas Andres, Pietro Millossovich, Kaishev Vladimir, et al., *Stmomo: Stochastic mortality modeling in r*, Journal of Statistical Software **84** (2018), no. 3, 1–38.

Antoine Burg and Christophe Dutang, *Closed-form estimators for multivariate regressions models - a single categorical variable approach*, submitted (2024).

Alexandre Brouste, Christophe Dutang, and Tom Rohmer, *Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling*, Computational Statistics **35** (2020), 689–724.

______, *A closed-form alternative estimator for glm with categorical explanatory variables*, Communications in Statistics-Simulation and Computation (2022), 1–17.

Nan Li and Ronald Lee, *Coherent mortality forecasts for a group of populations: An extension of the lee-carter method*, Demography **42** (2005), 575–594.