Functional concurrent regression with compositional covariates and its application to the time-varying effect of causes of death on human longevity

Emanuele Giovanni De Paoli, Marco Stefanucci and Stefano Mazzuco
Department of Statistical Sciences,
University of Padova,
and
Department of Economics and Finance,
University of Rome Tor Vergata,

October 30, 2023

Abstract

Multivariate functional data that are cross-sectionally compositional data are attracting increasing interest in the statistical modeling literature, a major example being trajectories over time of compositions derived from cause-specific mortality rates. In this work, we develop a novel functional concurrent regression model in which independent variables are functional compositions. This allows us to investigate the relationship over time between life expectancy at birth and compositions derived from cause-specific mortality rates of four distinct age classes, namely 0–4, 5–39, 40–64 and 65+. A penalized approach is developed to estimate the regression coefficients and select the relevant variables. Then an efficient computational strategy based on an augmented Lagrangian algorithm is derived to solve the resulting optimization problem. The good performances of the model in predicting the response function and estimating the unknown functional coefficients are shown in a simulation study. The results on real data confirm the important role of neoplasms and cardiovascular diseases in determining life expectancy emerged in other studies and reveal several other contributions not yet observed.

keyword: Mortality by Cause, Life Expectancy, Functional Data Analysis, Compositional Data Analysis,Sparsity.

1 Introduction

There is still a considerable heterogeneity across countries (even if we focus on high-income countries only) in terms of longevity, and the variability of the time pattern with which the recent mortality levels have been reached, is even more heterogeneous. Several studies have investigated on these time patterns (see, for instance, Canudas-Romo, 2010), but recently some are trying to analyze the role of causes of death in determining them. For example, Bergeron-Boucher et al. (2020) try to determine which causes of death are associated with longevity extension. Woolf and Schoomaker (2019) attribute the recent stagnation of life expectancy in the USA to increasing midlife mortality caused by drug overdoses, alcohol abuses, suicides and some organ diseases. Mehta et al. (2020) contest these findings, arguing that cardiovascular diseases are mainly responsible for such stagnation. The idea of associating life expectancy (or other summary measures of mortality rates) with causes of death, is not new: many of these employ a decomposition method (see, for instance, Vaupel and Canudas-Romo (2003)). However much of these studies are limited to a single country (see Jasilionis et al. (2023); Mehta et al. (2020)) or to a single age group (see Remund et al. (2018)), others (see Canudas-Romo et al. (2020)) collapse time dimension into a single indicator, thus not considering the evolution of causes of death over the last decades. Recently Stefanucci and Mazzuco (2022) proposed a combination of Functional Data Analysis (FDA) and Compositional Data Analysis (CDA) to analyze the time pattern of causes of death, limiting to mortality at age 40–64. Although the study by Stefanucci and Mazzuco (2022) provides some useful insights on the evolution of cause-specific mortality, it remains of descriptive nature and limited to a specific age group, while it might be of interest to measure if and to what extent different compositions of causes of death are associated with the evolution of overall mortality in the latest years. Conducting such an analysis can prove highly beneficial in gaining valuable insights into the epidemiological experiences of different countries. Moreover, it allows for an indirect association with trends in risk factors, such as the prevalence of smoking.

We suggest that this can be performed by regressing the evolution of overall mortality (measured in terms of life expectancy at birth) with causes of death composition of mortality as defined by Stefanucci and Mazzuco (2022). Sun et al. (2020) have recently proposed a log-contrast regression model with functional compositional covariates but limiting to the case of a scalar response variable. Although of great interest, their model is not specifically tailored to our purposes, thus we extend the previous work to cope with the functional essence of our response variable (life expectancy over time). Such an extension consists of a concurrent specification of the function-on-function linear regression model, with appropriate constraints due to the compositional nature of the covariates. Four age groups of causes of death are considered i.e., 0–4, 5–39, 40–64, and 65+, thus giving rise to four different compositions, each with many components – not necessarily the same ones, as shown in Table 1. Since it is reasonable that only few of them are relevant to predict the outcome, the model

Table 1: Classifications of causes of death used and age groups for which they are considered.

Classifications of causes of death	Age classes			
Congenital anomalies (CONG)	0–4			
Infancy related causes, excluded congenital anomalies (INFA)	0–4			
Certain infectious and parasitic diseases (INFE)	0–4	5–39	40–64	65+
Neoplasms (NEOP)	0–4	5–39	40–64	65+
Respiratory diseases (RESP)	0–4	5–39	40–64	65+
External causes of death (EXT)	0–4	5–39	40–64	65+
Diseases of nervous system (NERV)	0–4	5–39	40–64	65+
Digestive system diseases (DIG)		5–39	40–64	65+
Mental disorders (MENT)		5–39	40–64	65+
Endocrine, nutritional and metabolic diseases (END)		5–39	40–64	65+
Circulatory system diseases (CIRC)		5–39	40–64	65+
Diseases of urogenital system (UROG)			40–64	65+
Lung cancer (LUNG)			40–64	65+
Diseases of skin, musculoskeletal system and connective tissue system (SKIN)			40–64	65+

Table 2: Considered countries.

Area	Country
North Eur.	Denmark, Finland, Norway, Sweden
West Eur.	Austria, Belgium, Switzerland, France, Ireland, Netherlands, UK
East Eur.	Hungary, Poland, Lithuania, Estonia, Latvia, Russia, Ukraine
South Eur.	Italy, Spain
Extra Eur.	USA, Japan, New Zealand, Canada, Australia

specification assumes sparsity of the regression coefficients. In this way, variable selection is performed and interpretable results are obtained. An efficient computational strategy based on an augmented Lagrangian algorithm is also described to estimate the proposed model, and the performances of the method are illustrated through a simulation study.

The article proceeds as follow: in Section 2 we describe the analyzed data and formalize all the relevant quantities, in Section 3 we introduce a novel concurrent functional regression model with compositional covariates and discuss its estimation. The results of a simulation study are presented in Section 4 and the results on real data are extensively commented in Section 5. Finally, Section 6 concludes the article.

2 Data and problem setup

For each cause i, age x and calendar year t, we consider cause-specific mortality rates that can be written as

$$ {}^{i}m_x^t = m_x^t \frac{{}^{i}D_x^t}{D_x^t}, $$

where ${}^{i}D_x^t$ is the number of deaths for cause i at age x and time t, D_x^t is the number of deaths for all causes at age x and time t, ${}^{i}m_x^t$ and m_x^t are the cor-

respondent rates. For a given age x, compositions of mortality rates can be regarded as compositions of ${}^{i}m_x^t$ using m_x^t as normalization constant. Otherwise, data with unit-sum constraints may be obtained from ${}^{i}D_x^t$, using $\sum_{x,i} {}^{i}D_x^t$ as normalization constant. The latter approach was adopted by Oeppen (2008) and Kjærgaard et al. (2019) to model and forecast age-at-death distributions. In this way, the parts of the composition are related to different ages and the results could be difficult to interpret. Although it is not a problem for forecasting purposes, it is a major drawback for our perspective. The exact opposite of the previous approach is to study ${}^{i}m_x^t$ directly, that is, different compositions for each age. This would result in many predictors, making estimation problematic, especially for limited sample sizes. Moreover, as before, interpreting the results could be challenging. For these reasons, we focus on four age classes: 0–4, 5–39, 40–64, 65+, giving rise to four different compositions. From a demographic point of view, they account for infant, premature, early-adult and senescent mortality causes of death patterns, respectively. The underlying idea is that not only the causes of death composition changes among age groups, but also its effect on life expectancy varies with age. Age stratification allows us to control for different age structures across countries and over time. Life expectancy is a summary indicator of overall mortality that is independent on age structure of population but the compositions of causes of death are potentially affected by age structure changes, since some causes might be negligible at very young ages and highly relevant for old ones (e.g. dementia) and others (e.g. congenital anomalies) may be the other way round. Therefore, by considering a distinct composition for each age group, we can take into account the changing significance of different causes of death according to age. Consequently, certain causes may become irrelevant for specific age classes. Data on causes of death come from the WHO mortality database (Organization, Organization) and from the Human Causes of Death database (HCD) (for Demographic Studies (France) and for Demographic Research (Germany), for Demographic Studies (France) and for Demographic Research (Germany)) which contain time series of age-specific and cause-specific deaths for several countries. A primary issue is that the International Classification of Diseases (ICD) has changed significantly over the years, determining potentially biased results. Following Canudas-Romo et al. (2020) and Stefanucci and Mazzuco (2022), we use broad categories of causes, which are minimally affected by the classification revisions. The categories considered are shown in Table 1: the number of causes is higher with respect to Stefanucci and Mazzuco (2022), who limit their analysis to age group 40–64. Here, we also consider causes that are specific to infant ages (e. g., congenital anomalies) and senescent ones (e. g., mental disorders, including dementia and Alzheimer's disease). As can be seen in Table 1, only some of the 14 causes are included in the composition of a specific age group. For example, age 0–4 includes only 7 causes; the others are ignored as their role for that age group is negligible. On average, our classification accounts for 88% of the total number of deaths for the age class 0–4, 92% for the age class 5–39, and 98% for the age classes 40–64 and 65+. Regarding the countries used in this work, after some preliminary analyses, we decided to limit the study to the $n = 25$ nations

reported in Table 2, with a population size exceeding one million and good data quality. In order to consider the same time window for each nation, we restrict the study to the years 1965–2012. Some years are still missing for a few countries, that is, 2005 for Australia, 1996—1997 for Poland, and 2000 for the UK. This is not an issue, since our methodology also works for a non-equispaced time grid. Furthermore, a small number of zero counts is present in the age class 0–4 for external causes, neoplasms, infectious, respiratory and nervous diseases, as well as for mental and digestive diseases in the age class 5–39 and mental diseases in the other two age groups. Since the data have to be log-transformed, we replace them by the maximum rounding error of 0.5, which is a common practice in CDA (Aitchison, 2003). Concerning life expectancy at birth, we use data from life tables from the Human Mortality Database (HMD) University of California and for Demographic Research (Germany) (University of California and for Demographic Research (Germany)), which contains detailed, consistent, and high-quality data on human overall mortality, with no distinction among causes (Barbieri et al., 2015).

3 Methods

The main objective is to analyze the time-varying effect of causes of death on human longevity, studying whether variations in the causes of death composition can be predictive of life expectancy at birth. Since life expectancies in a given year are calculated based on age-specific mortality rates for the same year, we assume a concurrent relationship between the response variable and the covariates. We formulate the statistical problem in a very general way, considering an arbitrary number q of age classes and the possible inclusion of time-varying control variables. Let $\boldsymbol{y}(t) = [y_1(t), \ldots, y_n(t)]^\top \in \mathbb{R}^n$ be the response vector whose i-th component is the life expectancy at birth at time $t \in \mathcal{T}$ for the i-th country, with $i = 1, \ldots, n$. Let $\boldsymbol{x}_{ij}(t) = \left[x_{ij1}(t), \ldots, x_{ijp_j}(t)\right]^\top \in \mathbb{S}^{p_j - 1}$ be the composition of p_j cause-specific mortality rates for the i-th nation and j-th age class at time t, with $j = 1, \ldots, q$, and $\mathbb{S}^{p-1} = \left\{[x_1, \ldots, x_p]^\top \in \mathbb{R}^p, x_j > 0, \sum_{j=1}^p x_j = 1\right\}$ denoting the positive simplex lying in $\mathbb{R}^p$. Also, let $\boldsymbol{x}_i(t) = \left[\boldsymbol{x}_{i1}(t)^\top, \ldots, \boldsymbol{x}_{iq}(t)^\top\right]^\top \in \mathbb{R}^q$ be the vector containing all the q compositions, with $p = \sum_{j=1}^q p_j$, and let $\boldsymbol{X}(t) = [\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_n(t)]^\top \in \mathbb{R}^{n \times p}$ be the matrix of functional predictors at time t. Finally, $\boldsymbol{Z}_c(t) \in \mathbb{R}^{n \times (p_c + 1)}$ is the matrix of control variables at time t, where the first column is a vector of ones $\mathbf{1}_n$, to estimate the functional intercept. The observed life expectancies and compositions of mortality rates at each calendar year can be considered as discrete observations from $\boldsymbol{y}(t)$ and $\boldsymbol{X}(t)$, respectively.

3.1 Linear log-contrast model

Since the pioneering work of Aitchison and Bacon-Shone (1984), log-contrast models have been very popular for regression problems with compositional co-

variates. Suppose that we observe a response vector $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ and a design matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times p}$ with $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^\top \in \mathbb{S}^{p-1}$, for $i = 1 \ldots, n$. Because of the unit-sum constraint, each row of the matrix $\boldsymbol{X}$ cannot vary freely and the classical regression model is subject to identification problems. A naive solution is to omit one of the parts of the composition, but the method is not invariant to the choice of the removed component and the resulting coefficients are difficult to interpret. The idea of Aitchison and Bacon-Shone (1984) is to perform a log-ratio transformation of the compositional data so that the transformed data admit the familiar Euclidean geometry in $\mathbb{R}^{p-1}$. For a given reference component $r \in \{1, \ldots, p\}$, let $\boldsymbol{Z}_r = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^\top \in \mathbb{R}^{n \times (p-1)}$ be the associated design matrix, where the j-the element of $\boldsymbol{z}_i$ is given by $z_{ij} = \log(x_{ij}/x_{ir})$, for $j = 1, \ldots, r-1, r+1, \ldots, p$. The resulting linear log-contrast model is

$$\boldsymbol{y} = \mathbf{1}_n \beta_0 + \boldsymbol{Z}_r \boldsymbol{\beta}_r + \boldsymbol{e}, \tag{1}$$

where β_0 is the intercept, $\boldsymbol{\beta}_r \in \mathbb{R}^{p-1}$ is the regression coefficient, and $\boldsymbol{e} \in \mathbb{R}^n$ is the error vector, independent from $\boldsymbol{Z}_r$ and distributed as $\mathcal{N}(0, \sigma^2)$. The log-contrast model can be written in the symmetric form

$$\boldsymbol{y} = \mathbf{1}_n \beta_0 + \boldsymbol{Z} \boldsymbol{\beta} + \boldsymbol{e}, \quad \text{s.t. } \mathbf{1}_p^\top \boldsymbol{\beta} = 0, \tag{2}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ is the matrix resulting from log-transforming each element of the matrix $\boldsymbol{X}$, β_0 and $\boldsymbol{e}$ are the same as in model (1), and the regression coefficient $\boldsymbol{\beta}_r$ is the subvector obtained from $\boldsymbol{\beta}$ by removing the r-th component. The log-contrast model obeys a landmark concept in CDA, called subcompositional coherence (Aitchison, 2003): if the j-th coefficient of $\boldsymbol{\beta}$ is zero, then the results are unchanged if the model is applied to the subcomposition without the j-th component.

In the classical regression framework, the least squares estimation can be performed indifferently assuming the model (1) or the constrained form in (2). However, in a high-dimensional setup where variable selection is required, the use of a Lasso penalization method (Tibshirani, 1996) determines the loss of equivalence between the symmetric and non-symmetric form. For example, consider the inclusion of a L_1 penalty term for model (1), determining the optimization problem

$$\underset{\boldsymbol{\beta}_r, \beta_0}{\arg\min} \frac{1}{2} ||\boldsymbol{y} - \mathbf{1}_n \beta_0 + \boldsymbol{Z}_r \boldsymbol{\beta}_r||_2^2 + \lambda ||\boldsymbol{\beta}_r||_1. \tag{3}$$

The solution of problem (3) is not invariant to the choice of the reference category r and, in general, is different from that of the Lasso criteria related to the symmetric model (2), which determines the optimization problem

$$\underset{\boldsymbol{\beta}, \beta_0}{\arg\min} \frac{1}{2} ||\boldsymbol{y} - \mathbf{1}_n \beta_0 + \boldsymbol{Z} \boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1, \quad \text{s.t. } \mathbf{1}_p^\top \boldsymbol{\beta} = 0. \tag{4}$$

The latter is proposed and studied in the context of gut microbiome and metagenomic data by Lin et al. (2014), who also provide theoretical guarantees for the resulting estimator. Moreover, the zero-sum constraint makes the model subcompositional coherent.

3.2 Sparse functional concurrent log-contrast regression

Although in practice the functional compositional predictors and the response variable are observed at each calendar year, here we assume that $\boldsymbol{X}(t)$ and $\boldsymbol{y}(t)$ are observed for each $t \in \mathcal{T}$. Following the notation of Section 3.1 and Section 2, let $\boldsymbol{Z}(t) \in \mathbb{R}^{n \times p}$ be the matrix resulting from log-transforming each element of the matrix $\boldsymbol{X}(t)$ at time t, and recall that $\boldsymbol{y}(t) \in \mathbb{R}^n$ is the functional response and $\boldsymbol{Z}_c(t) \in \mathbb{R}^{n \times (p_c+1)}$ is the functional matrix of control variables, including a vector of ones $\mathbf{1}_n$ as the first column. The matrix $\boldsymbol{Z}(t)$ contains q compositions and thus we need to impose q zero-sum constraints to achieve subcompositional coherence. Following Lin et al. (2014) and Sun et al. (2020), we propose the functional concurrent log-contrast regression model

$$\boldsymbol{y}(t) = \boldsymbol{Z}_c(t)\boldsymbol{\beta}_c(t) + \boldsymbol{Z}(t)\boldsymbol{\beta}(t) + \boldsymbol{e}(t), \quad \text{s.t. } \boldsymbol{L}\boldsymbol{\beta}(t) = \mathbf{0}_q \quad \forall t \in \mathcal{T}, \tag{5}$$

where $\boldsymbol{\beta}(t) = \left[\boldsymbol{\beta}_1(t)^\top, \ldots, \boldsymbol{\beta}_q(t)^\top\right]^\top \in \mathbb{R}^p$ is the functional regression coefficient, with $\boldsymbol{\beta}_j(t) = [\beta_{j1}(t), \ldots, \beta_{jp_j}(t)]^\top \in \mathbb{R}^{p_j}$ for $j = 1, \ldots, q$, $\boldsymbol{\beta}_c(t) \in \mathbb{R}^{p_c+1}$ is the functional regression coefficient related to the control variables, and $\boldsymbol{e}(t) \in \mathbb{R}^n$ is the vector of functional errors distributed as $\mathcal{N}(0, \sigma^2)$. The set of linear constraints is represented by the matrix

$$\boldsymbol{L} = \begin{bmatrix} \mathbf{1}_{p_1} & \mathbf{0}_{p_1} & \cdots & \mathbf{0}_{p_1} \\ \mathbf{0}_{p_2} & \mathbf{1}_{p_2} & \cdots & \mathbf{0}_{p_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{p_q} & \mathbf{0}_{p_q} & \cdots & \mathbf{1}_{p_q} \end{bmatrix}^\top \in \mathbb{R}^{q \times p}.$$

For our study, it is reasonable to assume that the effects of causes of death on life expectancy are smooth over years. To achieve smoothness, each coefficient curve is represented by a linear expansion of k known basis functions, such that

$$\boldsymbol{\beta}(t) = \boldsymbol{B}\boldsymbol{\Phi}(t), \quad \boldsymbol{\beta}_c(t) = \boldsymbol{B}_c\boldsymbol{\Phi}(t),$$

where $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p]^\top \in \mathbb{R}^{p \times k}$ and $\boldsymbol{B}_c = \left[\boldsymbol{b}_0, \boldsymbol{b}_{c_1}, \ldots, \boldsymbol{b}_{c_p}\right]^\top \in \mathbb{R}^{(p_c+1) \times k}$ are the coefficient matrices, and $\boldsymbol{\Phi}(t) = [\phi_1(t), \ldots, \phi_k(t)]^\top \in \mathbb{R}^k$ is the vector of basis functions. For simplicity and since it is usually sufficient in practice, here we assume the same number k of basis functions for each predictor and control variable, obtained considering an equispaced grid of knots. Moreover, we assume that the elements of $\boldsymbol{\Phi}(t)$ are B-splines of order d (De Boor, 1978). A B-spline of order d is a piecewise polynomial function of degree $d-1$ and is defined by a set of knots, which are the points where the functions meet. The choice is not restrictive, and other basis functions can be adopted: see Ramsay and Silverman (2005) for a detailed discussion. The same consideration applies to the number of basis k, which can be assumed to be different for each coefficient curve.

Another reasonable assumption is that some compositional components have no effect on life expectancy. To enable variable selection, we induce sparsity by

using a L_1 penalization method. For model (5), the functional sparsity of the coefficient curves in $\boldsymbol{\beta}(t)$ translates into the row sparsity of the coefficient matrix $\boldsymbol{B}$. Many penalization methods have been proposed in Statistics and Machine Learning literature to induce sparsity, among which the Lasso (Tibshirani, 1996) is probably the most famous. The Group Lasso (Yuan and Lin, 2006) is an extension which considers the concept of groups of coefficients and fits for the purpose, since it allows the whole coefficient vectors $\boldsymbol{b}_j$, for $j = 1, \ldots, p$, to be selected rather than their individual components.

To formulate the optimization problem, the zero-sum constraints and the coefficient curves have to be expressed in terms of the elements of the matrices $\boldsymbol{B}$ and $\boldsymbol{B}_c$. For this purpose, it is convenient to express the problem in terms of $\boldsymbol{b} = \text{vec}(\boldsymbol{B}^\top) \in \mathbb{R}^{pk}$ and $\boldsymbol{b}_c = \text{vec}(\boldsymbol{B}_c^\top) \in \mathbb{R}^{(p_c+1)k}$. It can be easily seen that imposing $\mathbf{1}_{p_j}^\top \boldsymbol{b}_j(t) = 0$, for $j = 1, \ldots q$ and $\forall t \in \mathcal{T}$, is equivalent to imposing zero-sum constraints on the columns of the matrix $\boldsymbol{B}$, that is, $(\boldsymbol{L} \otimes \mathbf{I}_k)\boldsymbol{b} = \tilde{\boldsymbol{L}}\boldsymbol{b} = \mathbf{0}_{qk}$ with $\tilde{\boldsymbol{L}} \in \mathbb{R}^{qk \times pk}$. Moreover, we have that

$$\boldsymbol{\beta}(t) = \left(\mathbf{I}_p \otimes \boldsymbol{\Phi}(t)^\top\right) \boldsymbol{b} = \tilde{\boldsymbol{\Phi}}(t)\boldsymbol{b},$$

with $\tilde{\boldsymbol{\Phi}}(t) \in \mathbb{R}^{p \times pk}$ and, similarly, $\boldsymbol{\beta}_c(t) = \tilde{\boldsymbol{\Phi}}_c(t)\boldsymbol{b}_c$, with $\tilde{\boldsymbol{\Phi}}_c(t) \in \mathbb{R}^{(p_c+1)\times(p_c+1)k}$. In accordance with the above considerations, we propose to estimate the parameters to solve the optimization problem

$$\frac{1}{2} \underset{\boldsymbol{b}, \boldsymbol{b}_c}{\arg\min} \int \boldsymbol{r}(t)^\top \boldsymbol{r}(t) dt + \lambda \sum_{j=1}^{p} ||\boldsymbol{b}_j||_2, \quad \text{s.t. } \tilde{\boldsymbol{L}}\boldsymbol{b} = \mathbf{0}_{qk}, \tag{6}$$

where $\boldsymbol{r}(t) = \boldsymbol{y}(t) - \boldsymbol{Z}_c(t)\tilde{\boldsymbol{\Phi}}(t)\boldsymbol{b}_c - \boldsymbol{Z}(t)\tilde{\boldsymbol{\Phi}}(t)\boldsymbol{b} \in \mathbb{R}^n$ and λ is a tuning parameter that controls the strength of the group-Lasso penalization. The proposed estimator has the same desirable properties as its counterparts in the classical regression framework (Lin et al., 2014) and in the functional case with scalar response (Sun et al., 2020). The zero-sum constraints for each composition guarantee that the estimator would remain unchanged under the transformation $\boldsymbol{X}(t) \longmapsto \boldsymbol{S}\boldsymbol{X}(t)$, where $\boldsymbol{S} = \text{diag}(s_1, \ldots, s_n)$, with $s_i > 0$ for $i = 1, \ldots, n$. Furthermore, the constraints ensure that the proposed methodology is subcompositional coherent: if we knew that some coefficient curves of $\boldsymbol{\beta}(t)$ are zero and estimated the model using the compositions formed by excluding the parts associated with those curves, then the resulting estimator would be unchanged. Finally, a direct consequence of the symmetric formulation of the problem (6) is that the solution is invariant under any permutation of the components of each composition.

3.3 Computation

We propose to solve the convex optimization problem (6) using an augmented Lagrangian algorithm (Bertsekas, 1982). For a detailed review of the method and its extensions with applications in Statistics and Machine Learning, see

Boyd et al. (2011). The problem (6) can be rewritten as

$$\begin{aligned}\arg\min_{\boldsymbol{b},\boldsymbol{b}_c} \frac{1}{2}\boldsymbol{b}^\top \boldsymbol{K}\boldsymbol{b} - \boldsymbol{b}^\top \boldsymbol{J} + \frac{1}{2}\boldsymbol{b}_c^\top \boldsymbol{M}\boldsymbol{b}_c - \boldsymbol{b}_c^\top \boldsymbol{P} + \boldsymbol{b}_c^\top \boldsymbol{Q}\boldsymbol{b} \\ + \lambda \sum_{j=1}^{p} ||\boldsymbol{b}_j||_2, \quad \text{s.t. } \tilde{\boldsymbol{L}}\boldsymbol{b} = \boldsymbol{0}_{qk},\end{aligned} \tag{7}$$

where the matrices containing functional inner products with weighting functions are denoted by $\boldsymbol{K} = \int \tilde{\boldsymbol{\Phi}}(t)^\top \boldsymbol{Z}(t)^\top \boldsymbol{Z}(t)\tilde{\boldsymbol{\Phi}}(t)dt \in \mathbb{R}^{pk\times pk}$, $\boldsymbol{J} = \int \tilde{\boldsymbol{\Phi}}(t)^\top \boldsymbol{Z}(t)^\top \boldsymbol{y}(t)dt \in \mathbb{R}^{pk}$, $\boldsymbol{M} = \int \tilde{\boldsymbol{\Phi}}_c(t)^\top \boldsymbol{Z}_c(t)^\top \boldsymbol{Z}_c(t)\tilde{\boldsymbol{\Phi}}_c(t)dt \in \mathbb{R}^{(p_c+1)k\times(p_c+1)k}$, $\boldsymbol{P} = \int \tilde{\boldsymbol{\Phi}}_c(t)^\top \boldsymbol{Z}_c(t)^\top \boldsymbol{y}(t)dt \in \mathbb{R}^{(p_c+1)k}$ and $\boldsymbol{Q} = \int \tilde{\boldsymbol{\Phi}}_c(t)^\top \boldsymbol{Z}_c(t)^\top \boldsymbol{Z}(t)\tilde{\boldsymbol{\Phi}}(t)dt \in \mathbb{R}^{(p_c+1)k\times p_c k}$.

Since $\boldsymbol{b}_c$ is involved neither in the penalty term nor in the constraint, the optimization problem can be restated as

$$\arg\min_{\boldsymbol{b}} \frac{1}{2}\boldsymbol{b}^\top \tilde{\boldsymbol{K}}\boldsymbol{b} - \boldsymbol{b}^\top \tilde{\boldsymbol{J}} + \lambda \sum_{j=1}^{p} ||\boldsymbol{b}_j||_2, \quad \text{s.t. } \tilde{\boldsymbol{L}}\boldsymbol{b} = \boldsymbol{0}_{qk}, \tag{8}$$

where $\tilde{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{Q}^\top \boldsymbol{M}^{-1}\boldsymbol{Q} \in \mathbb{R}^{pk\times pk}$ and $\tilde{\boldsymbol{J}} = \boldsymbol{J} - \boldsymbol{Q}^\top \boldsymbol{M}^{-1}\boldsymbol{P} \in \mathbb{R}^{pk}$. Then, once the solution $\widehat{\boldsymbol{b}}$ is obtained, the estimate of the coefficient associated with the control variables is $\widehat{\boldsymbol{b}}_c = \boldsymbol{M}^{-1}(\boldsymbol{P} - \boldsymbol{Q}\widehat{\boldsymbol{b}})$.

The augmented Lagrangian associated with problem (8) is

$$L_\rho(\boldsymbol{b}, \boldsymbol{u}) = -\boldsymbol{b}^\top \tilde{\boldsymbol{J}} + \frac{1}{2}\boldsymbol{b}^\top \tilde{\boldsymbol{K}}\boldsymbol{b} + \lambda \sum_{j=1}^{p} ||\boldsymbol{b}_j||_2 + \frac{\rho}{2}||\tilde{\boldsymbol{L}}\boldsymbol{b}||_2^2 + \boldsymbol{u}^\top \tilde{\boldsymbol{L}}\boldsymbol{b},$$

where $\boldsymbol{u} \in \mathbb{R}^{qk}$ is the Lagrange multiplier and ρ is the penalty parameter. The augmented Lagrangian method finds the solution of the original problem iterating between a minimization step and a dual ascent step. The procedure for a fixed λ is summarized in Algorithm 1. We allow the penalty parameter ρ to increase in each iteration if the error does not decrease sufficiently over the previous iteration. The adjustment scheme follows the guidelines in Bertsekas (1982, p. 123). The first step of the algorithm updates

$$\boldsymbol{b}^k \leftarrow \arg\min_{\boldsymbol{b}} L_{\rho^{k-1}}(\boldsymbol{b}, \boldsymbol{u}^{k-1}),$$

and it is equivalent to solving a standard group-Lasso problem. In our implementation, we employ the Alternating Direction Method of Multipliers (Boyd et al., 2011), but other routines can be used to solve the problem. When the model is fitted for a path of λ, the solutions $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{b}}$ associated with the previous penalty term are used as a warm start for the subsequent iteration.

As noted before, the functional compositional predictors and the response variable are observed at each calendar year and not continuously $\forall t \in \mathcal{T}$. Therefore, all the integrals involved in the optimization problem have to be computed from discrete-time observations. In our study, we employ the trapezoidal rule, which is equivalent to approximate the discrete-time data to continuous-time curves by means of linear interpolation.

Algorithm 1 Augmented Lagrangian method to solve problem (8)

Require: $\boldsymbol{b}^0, \rho^0, \boldsymbol{u}^0, \epsilon, k_{\max}$
  $k \leftarrow 1$
  $\text{err}^0 \leftarrow \max \tilde{\boldsymbol{L}}\boldsymbol{b}^0$
  while $\text{err}^{k-1} > \epsilon$ & $k \leq k_{\max}$ **do**
    $\boldsymbol{b}^k \leftarrow \arg\min_{\boldsymbol{b}} L_{\rho^{k-1}}(\boldsymbol{b}, \boldsymbol{u}^{k-1})$
    $\text{err}^k \leftarrow \max \tilde{\boldsymbol{L}}\boldsymbol{b}^k$
    if $\text{err}^k > 0.25\text{err}^{k-1}$ **then**
      $\rho^k \leftarrow 10\rho^{k-1}$
    else
      $\rho^k \leftarrow \rho^{k-1}$
      $\boldsymbol{u}^k \leftarrow \boldsymbol{u}^{k-1} + \rho^k \tilde{L}\boldsymbol{b}^k$
    $k \leftarrow k + 1$

4 Simulations

We performed a simulation study in order to compare the performance of our proposal based on a constrained group Lasso (CGL) with two possible competitors. The first candidate is a baseline method, that is, a standard group Lasso in which the reference level r is chosen randomly (BGL). The second competitor is based on a naive approach, which consists of estimating the log-contrast regression model with the Lasso penalty of Lin et al. (2014) at each time $t \in \mathcal{T}$ and smoothing the resulting estimates.

We generate the compositional data similarly to the previous works of Lin et al. (2014), Shi et al. (2016), Sun et al. (2020). The discrete-time grid is equispaced within the interval $\mathcal{T} = [0, 1]$ and consists of 20 time points $t_1, \ldots, t_{20}$. We consider scenarios with $q = 4$ compositions, each with p_j components, $j = 1, \ldots, q$. To introduce dependence between the covariates, we use a compound symmetry correlation matrix $\boldsymbol{\Sigma}_X \in \mathbb{R}^{p_j \times p_j}$ with unit variances and correlations ρ_X. To account for time dependence, we consider a matrix $\boldsymbol{\Sigma}_T \in \mathbb{R}^{20\times 20}$ with first-order autoregressive structure, unit variance and autoregressive parameter ρ_T. For each observation $i = 1, \ldots, n$, the j-th composition over time is obtained by simulating $\boldsymbol{w}_{ij} = [\boldsymbol{w}_{ij}(t_1)^\top, \ldots, \boldsymbol{w}_{ij}(t_{20})^\top]^\top \sim \mathcal{N}(\boldsymbol{0}_{20p_j}, \sigma_X^2(\boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_X))$ and then normalizing the counts as

$$w_{ijl}(t_v) = \frac{\exp\{w_{ijl}(t_v)\}}{\sum_{l=1}^{p_j} \exp\{w_{ijl}(t_v)\}},$$

for $i = 1, \ldots, n$, $l = 1, \ldots, p_j$ and $v = 1, \ldots, 20$. The number of bases for cubic splines is set to $k = 5$ and the number of components p_j is the same across compositions and equal to p/q. Only 3 coefficients are non-null for each composition. The coefficient vectors are $\boldsymbol{b}_1 = [1, -1, 0, 0, 0]^\top$, $\boldsymbol{b}_2 = [0, 0, -0.5, 1, 0]^\top$, $\boldsymbol{b}_3 = [-1, 1, 0.5, -1, 0]^\top$, $\boldsymbol{b}_{p_1+1} = [0.5, 0, 0, -0.5, 1]^\top$, $\boldsymbol{b}_{p_1+2} = [0, 1, -1, 0, -1]^\top$, $\boldsymbol{b}_{p_1+3} = [-0.5, -1, 1, 0.5, 0]^\top$, $\boldsymbol{b}_{p_2+1} = [0.5, -1, -1, 1, 0]^\top$, $\boldsymbol{b}_{p_2+2} = [0, 1, 1, 0, 0]^\top$, $\boldsymbol{b}_{p_2+3} = [-0.5, 0, 0, -1, 0]^\top$,

Table 3: Means and standard errors (in parentheses) of false positive and false negative rates for the three methods with SNR = 2, based on 100 simulations.

Configuration					FPR(%)			FNR(%)		
ρ_X	ρ_T	n	p	q	CGL	BGL	Naive	CGL	BGL	Naive
0.2	0.2	50	40	1	0.04 (0.04)	0.39 (0.11)	1.54 (0.20)	3.58 (0.41)	3.67 (0.42)	10.42 (0.52)
		50	40	4	0.00 (0.00)	0.36 (0.11)	1.32 (0.23)	2.67 (0.39)	3.42 (0.41)	10.67 (0.52)
		50	100	4	4.06 (0.22)	7.18 (0.22)	8.84 (0.30)	0.25 (0.14)	0.42 (0.18)	8.83 (0.54)
0.2	0.6	50	40	1	0.14 (0.07)	0.64 (0.16)	1.68 (0.25)	3.50 (0.41)	3.67 (0.42)	10.67 (0.54)
		50	40	4	0.00 (0.00)	0.79 (0.17)	1.39 (0.23)	3.42 (0.43)	3.67 (0.43)	11.08 (0.46)
		50	100	4	4.08 (0.20)	7.28 (0.23)	9.12 (0.33)	0.75 (0.24)	1.50 (0.32)	8.58 (0.52)
0.6	0.2	50	40	1	0.00 (0.00)	0.32 (0.10)	1.50 (0.20)	3.92 (0.43)	4.00 (0.45)	10.00 (0.52)
		50	40	4	0.04 (0.04)	0.64 (0.15)	1.29 (0.22)	2.83 (0.40)	3.92 (0.45)	10.58 (0.44)
		50	100	4	3.91 (0.19)	6.98 (0.19)	9.41 (0.34)	0.92 (0.26)	1.17 (0.29)	9.00 (0.50)
0.6	0.6	50	40	1	0.00 (0.00)	0.71 (0.14)	1.43 (0.23)	4.33 (0.43)	4.75 (0.43)	10.67 (0.53)
		50	40	4	0.07 (0.05)	0.93 (0.19)	1.21 (0.20)	3.08 (0.40)	3.08 (0.40)	10.33 (0.53)
		50	100	4	4.39 (0.20)	7.57 (0.22)	8.99 (0.32)	0.92 (0.26)	1.17 (0.29)	9.92 (0.53)

$\boldsymbol{b}_{p_3+1} = [1, 0, 0.5, 0, -1]^\top$, $\boldsymbol{b}_{p_3+2} = [0, 0, -0.5, 0, 0]^\top$, $\boldsymbol{b}_{p_3+3} = [-1, 0, 0, 0, 1]^\top$. We also consider scenarios with $p = 40$ and $q = 1$, with the same coefficients and the same degree of sparsity as for $p = 40$ and $q = 4$. For simplicity, we do not include either the intercept or other control variables. The response variables are generated from the model (5), with error terms distributed as $\mathcal{N}(0, \sigma^2)$, where σ^2 set to achieve specific signal-to-noise ratios (SNR). We simulated different settings $(n, p, q) = (50, 40, 1), (50, 40, 4), (50, 100, 4)$ and several combinations of parameters $\sigma_X^2 = 9$, $\rho_T = (0.2, 0.6)$, $\rho_X = (0.2, 0.6)$, SNR $= (2, 4)$. The tuning parameters λ and k are selected by ten fold cross-validation and one-standard error rule (Hastie et al., 2009, p. 244)

We use four different measures to compare our proposal with competitors. The prediction error is calculated using the average prediction mean square error $\sum_{v=1}^{20} ||\boldsymbol{y}(t_v) - \mathbf{1}_n^\top \widehat{\beta}_0(t_v) - \boldsymbol{Z}(t_v)\widehat{\boldsymbol{\beta}}(t_v)||_2^2/(20n)$ computed from an independent test sample of size 1000. The estimation error is measured by $\sum_{j=1}^{p} \left(\int_{\mathcal{T}} |\widehat{\beta}_j(t) - \beta_j(t)|^2 dt \right)^{\frac{1}{2}} /p$. As variable selection measures, we use the false positive rate (FPR) and false negative rate (FNR), where positives and negatives refer to non-null and null coefficients, respectively. The naive method does not include a procedure for the selection of coefficient curves, but only a variable selection procedure at each time t, therefore, we select active predictors based on empirical evidence. Consequently, to have a fair comparison, we use the same criteria for all three methods. As in Sun et al. (2020), the estimated index set $\widehat{\mathcal{S}}$ of non-null coefficients is defined as

$$\widehat{\mathcal{S}} = \left\{ j : \frac{\left(\int_{\mathcal{T}} \widehat{\beta}_j^2(t) dt \right)^{\frac{1}{2}}}{\sum_{j=1}^{p} \left(\int_{\mathcal{T}} \widehat{\beta}_j^2(t) dt \right)^{\frac{1}{2}}} \geq \frac{1}{p}, j = 1, \ldots, p \right\}.$$

The means and standard errors of the performance measures for the scenario with SNR = 2 are reported in Table 3 and 4. From Table 3, we can see that the proposed CGL has a similar variable selection performance compared to BGL when $n > p$, although the latter has the tendency to have higher false

Table 4: Means and standard errors (in parentheses) of prediction and estimation errors for the three methods with SNR = 2, based on 100 simulations. Estimation errors are multiplied by 100.

Configuration					Prediction error			Estimation error		
ρ_X	ρ_T	n	p	q	CGL	BGL	Naive	CGL	BGL	Naive
0.2	0.2	50	40	1	8.45 (0.02)	8.49 (0.02)	13.92 (0.11)	3.90 (0.04)	3.97 (0.04)	7.77 (0.04)
		50	40	4	8.22 (0.02)	8.34 (0.03)	12.55 (0.08)	3.80 (0.04)	4.15 (0.06)	7.63 (0.05)
		50	100	4	8.35 (0.03)	8.59 (0.04)	14.64 (0.10)	2.01 (0.02)	2.25 (0.03)	3.97 (0.02)
0.2	0.6	50	40	1	8.46 (0.03)	8.48 (0.03)	14.04 (0.13)	4.06 (0.05)	4.13 (0.06)	7.90 (0.06)
		50	40	4	8.35 (0.03)	8.51 (0.03)	12.63 (0.09)	3.88 (0.04)	4.25 (0.06)	7.64 (0.06)
		50	100	4	8.35 (0.03)	8.68 (0.04)	14.64 (0.09)	2.03 (0.02)	2.31 (0.03)	3.98 (0.03)
0.6	0.2	50	40	1	4.18 (0.01)	4.21 (0.01)	7.33 (0.07)	3.91 (0.04)	4.00 (0.04)	7.87 (0.05)
		50	40	4	4.08 (0.01)	4.14 (0.01)	6.24 (0.04)	3.83 (0.04)	4.12 (0.05)	7.56 (0.04)
		50	100	4	4.30 (0.02)	4.43 (0.02)	7.62 (0.06)	1.96 (0.02)	2.21 (0.02)	3.99 (0.02)
0.6	0.6	50	40	1	4.22 (0.01)	4.24 (0.01)	7.23 (0.08)	3.93 (0.04)	4.06 (0.05)	7.86 (0.06)
		50	40	4	4.04 (0.01)	4.11 (0.01)	6.17 (0.03)	3.86 (0.04)	4.21 (0.06)	7.61 (0.04)
		50	100	4	4.34 (0.01)	4.48 (0.02)	7.59 (0.06)	2.01 (0.02)	2.29 (0.02)	3.97 (0.02)

Table 5: Means and standard errors (in parentheses) of false positive and false negative rates for the three methods with SNR = 4, based on 100 simulations.

Configuration					FPR(%)			FNR(%)		
ρ_X	ρ_T	n	p	q	CGL	BGL	Naive	CGL	BGL	Naive
0.2	0.2	50	40	1	0.00 (0.00)	0.00 (0.00)	0.21 (0.09)	1.75 (0.34)	1.50 (0.32)	7.17 (0.43)
		50	40	4	0.00 (0.00)	0.00 (0.00)	0.18 (0.08)	1.33 (0.31)	1.83 (0.35)	7.83 (0.33)
		50	100	4	1.14 (0.11)	3.48 (0.17)	4.57 (0.23)	0.00 (0.00)	0.00 (0.00)	5.08 (0.51)
0.2	0.6	50	40	1	0.00 (0.00)	0.00 (0.00)	0.14 (0.07)	2.50 (0.38)	2.33 (0.38)	6.67 (0.44)
		50	40	4	0.00 (0.00)	0.04 (0.04)	0.11 (0.06)	1.33 (0.31)	1.75 (0.34)	7.25 (0.37)
		50	100	4	1.37 (0.13)	3.93 (0.18)	5.12 (0.22)	0.17 (0.12)	0.08 (0.08)	6.25 (0.49)
0.6	0.2	50	40	1	0.00 (0.00)	0.07 (0.05)	0.18 (0.08)	2.58 (0.39)	2.67 (0.39)	7.50 (0.37)
		50	40	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	1.58 (0.33)	1.58 (0.33)	8.33 (0.37)
		50	100	4	1.30 (0.13)	3.53 (0.17)	4.23 (0.19)	0.00 (0.00)	0.17 (0.12)	5.17 (0.46)
0.6	0.6	50	40	1	0.00 (0.00)	0.00 (0.00)	0.14 (0.07)	2.67 (0.39)	2.67 (0.39)	7.08 (0.46)
		50	40	4	0.00 (0.00)	0.04 (0.04)	0.11 (0.06)	1.75 (0.34)	1.83 (0.35)	7.67 (0.35)
		50	100	4	1.24 (0.13)	3.98 (0.18)	5.28 (0.23)	0.17 (0.12)	0.33 (0.16)	5.67 (0.49)

positive rates. This behavior is due to the automatic inclusion of the randomly chosen baseline for BGL and is, in fact, even more pronounced for $q = 4$. The advantages of the proposed CGL can be appreciated for the scenarios with $p > n$, where it clearly outperforms competitors. As seen in Table 4, the proposed CGL performs slightly better in terms of prediction and estimation error and, as before, the difference with the competitors is emphasized for $p > n$. Furthermore, increasing the correlation between the components leads to lower prediction errors, regardless of the method. This is because a small correlation determines few dominating components in each composition. As expected, the naive method has inferior performance in terms of all the measures in all the settings, since it is an unsophisticated approximation of functional nature of the data. Another expected behavior can be seen from Tables 5 and 6, which show that increasing the SNR leads to improved performance.

Table 6: Means and standard errors (in parentheses) of prediction and estimation errors for the three methods with SNR = 4, based on 100 simulations. Estimation errors are multiplied by 100.

Configuration					Prediction error			Estimation error		
ρ_X	ρ_T	n	p	q	CGL	BGL	Naive	CGL	BGL	Naive
0.2	0.2	50	40	1	4.29 (0.01)	4.31 (0.01)	7.71 (0.08)	2.98 (0.03)	3.05 (0.03)	5.97 (0.04)
		50	40	4	4.30 (0.01)	4.36 (0.02)	6.78 (0.05)	2.85 (0.03)	3.08 (0.04)	5.71 (0.05)
		50	100	4	4.31 (0.01)	4.46 (0.02)	8.75 (0.09)	1.53 (0.02)	1.71 (0.02)	3.28 (0.02)
0.2	0.6	50	40	1	4.30 (0.02)	4.33 (0.02)	7.90 (0.11)	3.00 (0.03)	3.09 (0.03)	6.02 (0.05)
		50	40	4	4.28 (0.01)	4.34 (0.02)	6.71 (0.06)	2.86 (0.03)	3.09 (0.04)	5.69 (0.05)
		50	100	4	4.46 (0.02)	4.61 (0.02)	8.90 (0.08)	1.52 (0.02)	1.72 (0.02)	3.29 (0.02)
0.6	0.2	50	40	1	2.21 (0.01)	2.22 (0.01)	4.25 (0.05)	2.93 (0.03)	3.02 (0.03)	6.00 (0.04)
		50	40	4	2.17 (0.01)	2.19 (0.01)	3.49 (0.03)	2.85 (0.03)	3.04 (0.04)	5.72 (0.04)
		50	100	4	2.22 (0.01)	2.28 (0.01)	4.57 (0.04)	1.51 (0.02)	1.68 (0.02)	3.28 (0.02)
0.6	0.6	50	40	1	2.18 (0.01)	2.18 (0.01)	4.15 (0.05)	3.02 (0.04)	3.07 (0.04)	6.03 (0.04)
		50	40	4	2.11 (0.01)	2.14 (0.01)	3.36 (0.03)	2.92 (0.03)	3.17 (0.04)	5.81 (0.04)
		50	100	4	2.24 (0.01)	2.32 (0.01)	4.61 (0.05)	1.53 (0.02)	1.72 (0.02)	3.30 (0.02)

5 Results

The proposed functional concurrent regression model with compositional covariates is fitted separately for males and females, since the trajectories of the causes of death have profoundly different characteristics. We use cubic spline bases and the penalty parameter λ as well as the number of basis functions k are selected through leave-one-out cross-validation, due to the limited sample size, and one-standard error rule.

As a by-product of the regression model results, we can measure the relative importance of causes in the j-th age class by considering the relative squared L_2 norm of the group-specific coefficients between years t and $t+1$

$$\sum_{l=1}^{p_j} \int_t^{t+1} |\beta_{jl}(t)|^2 dt \Bigg/ \sum_{j=1}^{4} \sum_{l=1}^{p_j} \int_t^{t+1} |\beta_{jl}(t)|^2 dt \ .$$

The results are reported in Figure 1 and show that, for both men and women, the most important age class is 40–64. The reason can be attributed to the inclusion of countries from Eastern Europe, for which the compositional trajectories in the age group 40–64 are very different from the other high-longevity nations. The result is consistent with the demographic literature, in which traditional life expectancy decomposition methods are applied longitudinally for single countries. For example, Meslé (2004) shows that in many former Soviet countries, decreases in life expectancy in the period 1965–2000 for males can be attributed to the rise in mortality at working ages. This is also in line with the substantial sex difference in the contribution of the age group 5–39. Another expected finding is the decline in importance for the age group 0–4, regardless of sex, which is associated with a progressive reduction in infant mortality. We also notice an increasing importance of age class 65+ for men. This can be explained by the faster progress of men in reducing heart disease-related mortality in recent decades, a pattern observed by Feraldi and Zarrulli (2022).

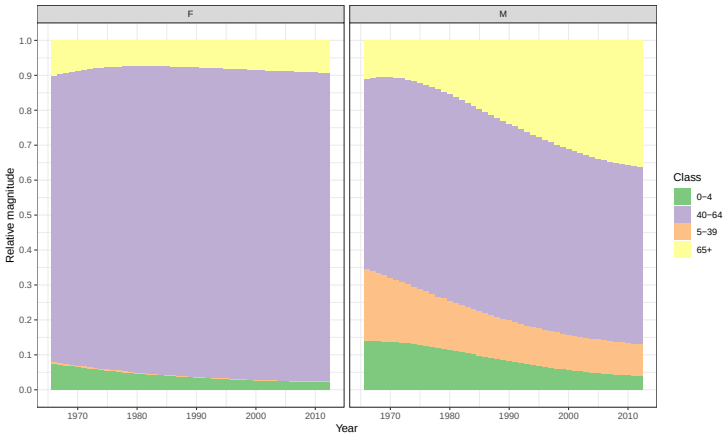
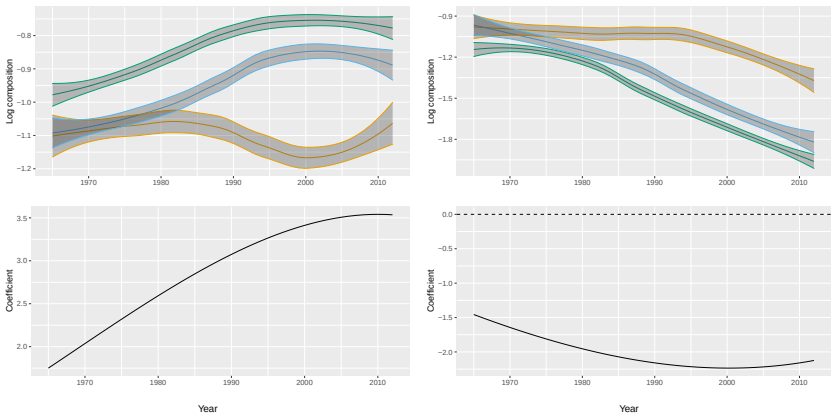


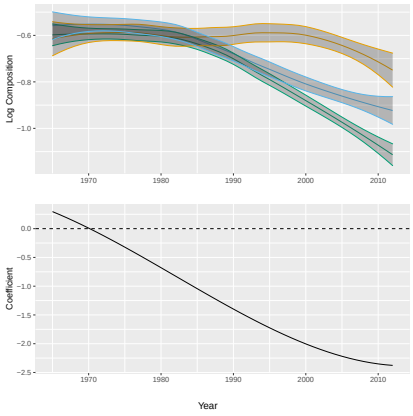
Figure 1: *Relative magnitude of the age group–specific coefficients for females and males.*

Regarding the results relative to specific causes, it is worth recalling that the interpretation of coefficients for the log-contrast model is different from the standard linear regression model. The main reason lies in the zero-sum constraint, which reflects the fact that one component increases its relative importance only if one or more of the others decreases (Coenders and Pawlowsky-Glahn, 2020). For the model (5), it can be shown that the following interpretation holds at each time t. Multiplying by a factor c the ratio of one component $\beta_{jl}(t)$ of the j-th composition over each of the other parts $\beta_{jm}(t), m = 1, \ldots, j-1, j+1, \ldots, p_j$ leads to a change of $\log(c)\beta_{jl}(t)$ in the expected value of the response variable. Equivalently, we can also interpret the coefficients jointly as follows. The expected value of the response variable grows when increasing the relative importance of components with positive coefficient and reducing that of components with negative coefficient. However, interpretation over time is not straightforward and we made use of additional plots to elucidate it, following Sun et al. (2020). The idea is to compare the smoothed trajectories of log compositions for three clusters of countries with the estimated coefficient curves. For each predictor and each year, the nations are divided into three groups characterized by low, medium and high life expectancy, thus giving rise to time-varying partitions. For each group, the smoothed values together with their 95% confidence bands are calculated using local regression. In this way, we can also check whether our model describes relationships encountered in raw data. Figure 2 shows the resulting plots for four relevant causes. The graphs show that our model provides realistic results. We observe that increases (decreases) in the difference of the prevalence of a cause of death between high and low longevity countries are reflected in increasing (decreasing) coefficient curves. For example, considering the age class 40–64 for females, in the '60s, countries with higher prevalence of death by neoplasms and lower by circulatory diseases have higher

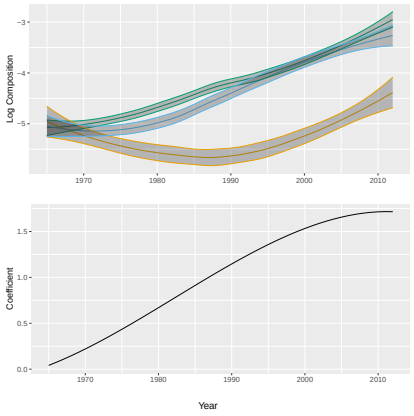


(a) NEOP, age class 40–64, females

(b) CIRC, age class 40–64, females



(c) CIRC, age class 65+, males



(d) NERV, age class 65+, males

Figure 2: *Smoothed curves of log composition of some causes of death for three clusters of countries, with the estimated coefficient curves below. For each predictor and year, the nations are divided into three groups characterized by low (in yellow), medium (in light blue) and high (in green) life expectancy.*

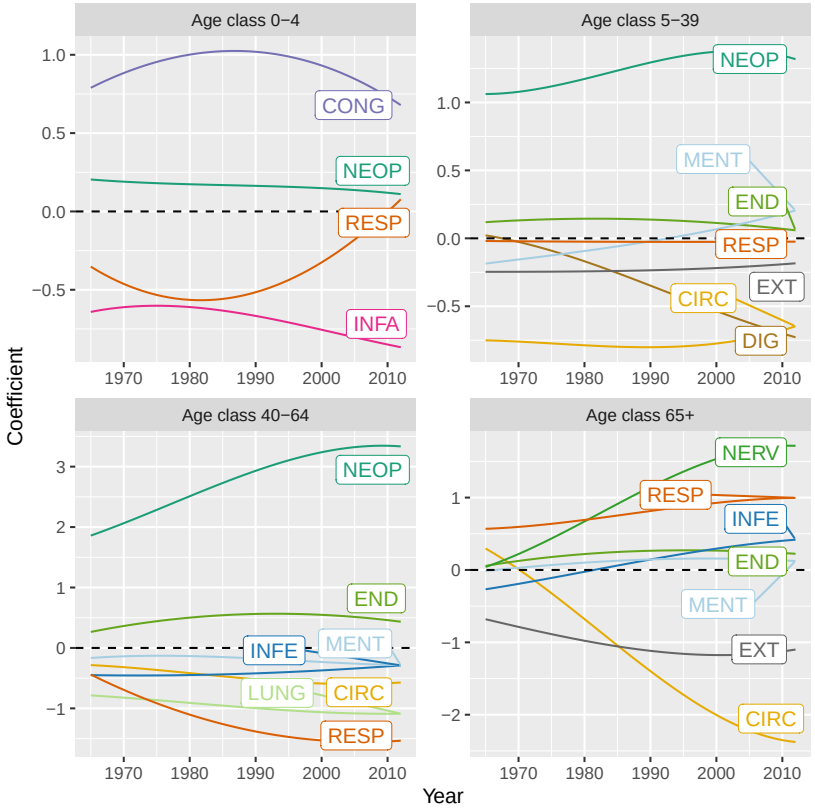


Figure 3: *Estimated coefficient curves for the four age classes, males.*

life expectancy. In subsequent years, the difference in terms of prevalence of neoplasm between high- and low-longevity countries increases and this is reflected in the increasing estimated curve, while the reverse holds for circulatory diseases.

The estimated coefficient curves for males are reported in Figure 3. The positive increasing trend of neoplasms in age classes 5–39 and 40–64 is a clear effect due to substitute mortality, which has been defined "that mortality which results from a decrease in another specific disease" (Van De Water, 1997). That is, in many countries with high longevity, cancer mortality has become the main cause of death due to the reduction of other conditions, such as those related to the circulatory system. In fact, circulatory diseases can be seen to have a negative effect for all age groups, excluded 0–4. Another cause with a negative decreasing effect in age class 5–39 is digestive diseases. It can be linked to the high incidence of this class of diseases, particularly liver cirrhosis, observed in early adulthood for Eastern European nations (Blachier et al., 2013) and other Commonwealth countries, such as the UK (Lewer et al., 2020). For the age

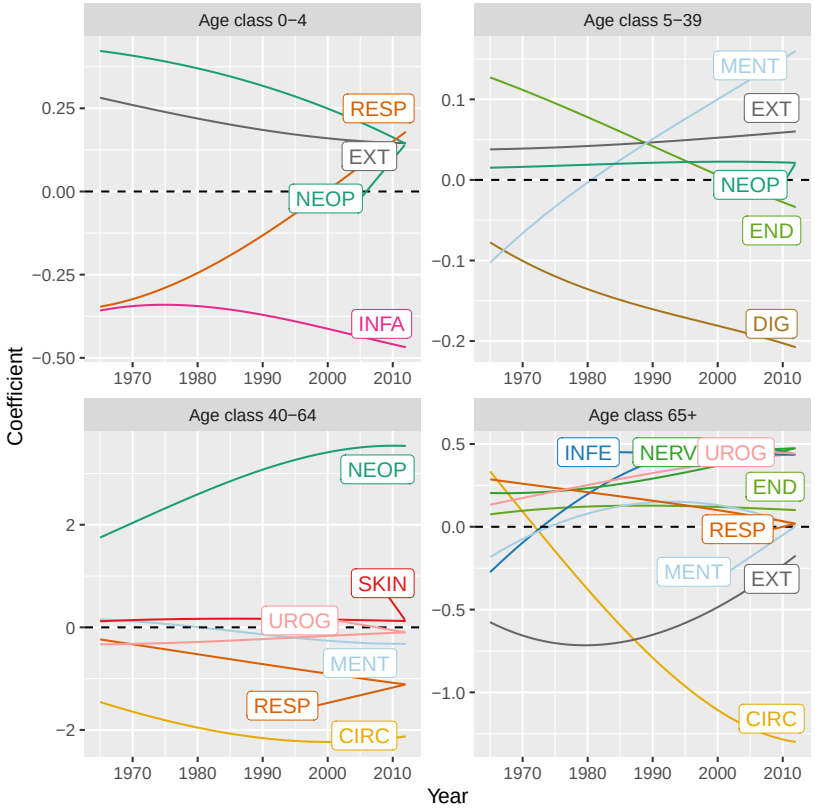


Figure 4: *Estimated coefficient curves for the four age classes, females.*

class accounting for senescent mortality, the effect of circulatory diseases is negative and strongly increasing, concurrently with the positive increasing effect of nervous, respiratory and infectious diseases. These are conditions whose susceptibility is higher in the elderly. The estimated positive increasing effect reflects the process of population aging, that is, the increase in proportion of population aged 65 and over, which is particularly vulnerable to the aforementioned diseases. It is interesting to highlight the sign change of infectious diseases, which means that in the first period this condition was associated with low-longevity countries.

The results for females are reported in Figure 4. Compared to males, the increasing positive effect of neoplasms and the increasing negative effect of circulatory diseases in age class 40–64 overshadow all others in terms of magnitude. In this age group, differently to males, skin and urogenital diseases are selected. On the contrary, endocrine and infectious diseases, as well as lung cancer, are not included. One possible explanation for the non-inclusion of lung cancer is its high mortality rate in both low- and high-life expectancy countries for women

(Jani et al., 2021). In the senescent age group, the effect of respiratory diseases is positive decreasing and, unlike males, there is an increase in the prevalence of urogenital diseases over time for high-longevity countries. This cause, which is also selected for age classes 40-64, appears to be a sex-specific cause.

To assess the stability of the selection procedure, we generated 500 bootstrap samples and used leave-one-out cross-validation to select the tuning parameters, as for the model estimated with the original data. The results reported in Figure 5 show that the variable selection is quite stable. In general, our proposal appears to be able to select the relevant predictors, at the cost of including some causes which may not have much effect on life expectancy. This is the case of external diseases in the age class 5–39 for both sexes, as well as neoplasms in the age class 5–39 for females and lung cancer and circulatory diseases in the age class 40–64 for males. On the other hand, infectious diseases for the age group 0–4 is selected in more than 70% of the bootstrap samples for both sexes, indicating that it may play an important role, although its coefficient is estimated zero.

6 Discussion

We introduced a functional regression model with compositional covariates in the spirit of the proposal by Sun et al. (2020), extending their work to the relevant framework of a functional response. The model allows us to explain the evolution of life expectancy at birth for several countries as a function of the compositions derived from cause-specific mortality rates of four distinct age groups. The method involves a B-spline expansion of the unknown functional coefficients coupled with a group-Lasso penalty, enabling variable selection at the function level and consequently high interpretability of the results. The methodology is implemented within the R package FCRC, available at `https://github.com/emanuelegdepaoli/fcrc`, where the code for reproducing the analysis, the simulation studies and all images of the paper is also included.

It is worth noting that causes of death cannot be regarded as causal drivers of overall mortality (life expectancy). The main reason is that the cause of death and mortality occur simultaneously, so one cannot be the cause of the other. It would be more sensible to include risk factors (e.g. life–styles, pollution, etc.) to assess a causal link with mortality. However, to our knowledge, there is no harmonized and sufficiently high quality cross–country data over time on risk factors to do that. Causes of death data are available, instead and while they cannot be considered really "drivers" of overall mortality we can see them as mediators between risk factors and life expectancy. Therefore such analysis can indirectly provide additional insights on the epidemiological trajectories of countries.

One major finding is that life expectancy is mainly driven by mortality at age 40–64 for women, while for men, also 65+ and 5–39 age groups are relevant. Not surprisingly, we found that circulatory diseases are increasingly relevant in determining the life expectancy of countries: the lower the relative

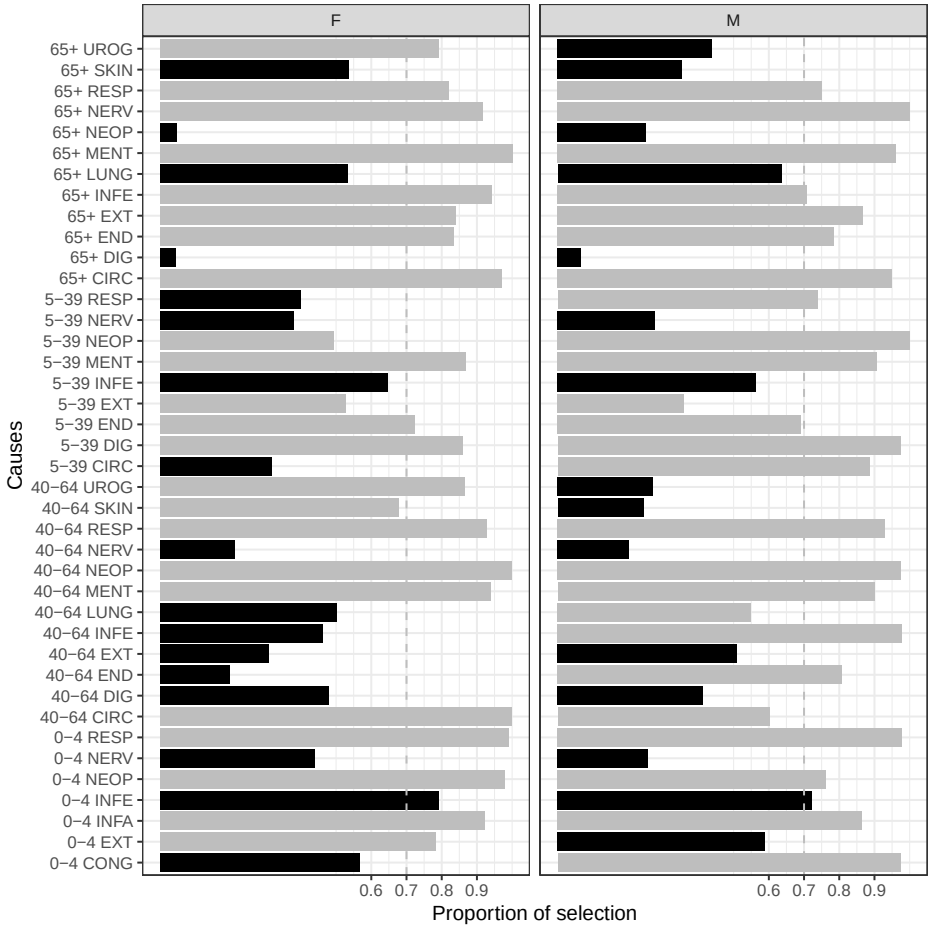


Figure 5: *Proportion of the causes of death selected in 500 bootstrap samples, for females and males. In gray, the bars of the selected predictors from fitting the model to the original data, in black the bars of the estimated null coefficients.*

importance of circulatory diseases, the higher the life expectancy. We also found an increasing relevance of digestive diseases for young men and women and of lung cancer for young men only. Other results, such as the increasingly positive effect of neoplasms at age 40–64 and of diseases of nervous system at age 65+ (that is, the higher the relative importance of these causes, the higher the life expectancy) can be explained in terms of "substitution effect", which means that the increasing relevance of these causes is an indirect effect of the reduction of other causes. We should keep in mind that the sample is made up of several countries with a different pattern of overall and cause-specific mortality. In particular, Eastern European countries that underwent a serious mortality crisis after the fall of the Soviet Union have a peculiar pattern that might have driven some of these results.

The proposed model allows us to simultaneously consider all causes of death and age groups in determining the evolution of overall mortality. This is increasingly important, since it has been observed that the composition of cause-specific mortality is getting increasingly diversified (Bergeron-Boucher et al., 2020), thus making analyses based on a single cause of death less reliable.

We consider the summary measure of life expectancy at birth, but other measures such as the modal age at death Canudas-Romo (2008), which is not affected by infant mortality, or lifespan disparity Vaupel and Canudas-Romo (2003), which is a measure of compression of age–specific mortality, can be used as a response variable, providing further insights on the evolution of mortality in high income countries.

Acknowledgments

The authors acknowledge financial support from the PRIN project "Unfolding the SEcrets of LongEvity: Current Trends and future prospects" (SELECT), project number 20177BRJXS. The authors also thank Emilio Zagheni, Ugofilippo Basellini and other scholars from the Max Planck Institute for Demographic Research for useful discussion during the presentation of earlier versions of this work.

References

Aitchison, J. (2003). *The statistical analysis of compositional data*. Caldwell, N.J. : Blackburn Press.

Aitchison, J. and J. Bacon-Shone (1984). Log contrast models for experiments with mixtures. *Biometrika 71*(2), 323–330.

Barbieri, M., J. R. Wilmoth, V. M. Shkolnikov, D. Glei, D. Jasilionis, D. Jdanov, C. Boe, T. Riffe, P. Grigoriev, and C. Winant (2015). Data resource profile: the human mortality database (hmd). *International Journal of Epidemiology 44*(5), 1549–1556.

Bergeron-Boucher, M.-P., J. M. Aburto, and A. van Raalte (2020). Diversification in causes of death in low-mortality countries: emerging patterns and implications. *BMJ Global Health 5*(7).

Bertsekas, D. P. (1982). *Constrained optimization and Lagrange multiplier methods*. Academic Press.

Blachier, M., H. Leleu, M. Peck-Radosavljevic, D.-C. Valla, and F. Roudot-Thoraval (2013). The burden of liver disease in europe: a review of available epidemiological data. *Journal of Hepatology 58*(3), 593–608.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Canudas-Romo, V. (2008). The modal age at death and the shifting mortality hypothesis. *Demographic Research 19*, 1179–1204.

Canudas-Romo, V. (2010). Three measures of longevity: time trends and record values. *Demography 47*(2), 299–312.

Canudas-Romo, V., T. Adair, and S. Mazzuco (2020). Reflection on modern methods: cause of death decomposition of cohort survival comparisons. *International Journal of Epidemiology 49*(5), 1712–1718.

Coenders, G. and V. Pawlowsky-Glahn (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, 201–220.

De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag New York.

Feraldi, A. and V. Zarrulli (2022). Patterns in age and cause of death contribution to the sex gap in life expectancy: a comparison among ten countries. *Genus 78*(23).

for Demographic Studies (France), F. I. and M. P. I. for Demographic Research (Germany). Human Causes of Death Database. Available at www.causeofdeath.org.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2nd Ed. Springer.

Jani, C., D. C. Marshall, H. Singh, R. Goodall, J. Shalhoub, O. Al Omari, J. D. Salciccioli, and C. C. Thomson (2021). Lung cancer mortality in europe and the usa between 2000 and 2017: an observational analysis. *ERJ Open Research 7*(4).

Jasilionis, D., A. A. van Raalte, S. Klüsener, and P. Grigoriev (2023). The underwhelming german life expectancy. *European Journal of Epidemiology*.

Kjærgaard, S., Y. E. Ergemen, M. Kallestrup-Lamb, J. Oeppen, and R. Lindahl-Jacobsen (2019). Forecasting causes of death by using compositional data analysis: the case of cancer deaths. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 68*(5), 1351–1370.

Lewer, D., W. Jayatunga, R. W. Aldridge, C. Edge, M. Marmot, A. Story, and A. Hayward (2020). Premature mortality attributable to socioeconomic inequality in england between 2003 and 2018: an observational study. *The Lancet Public Health 5*(1), e33–e41.

Lin, W., P. Shi, R. Feng, and H. Li (2014). Variable selection in regression with compositional covariates. *Biometrika 101*(4), 785–797.

Mehta, N. K., L. R. Abrams, and M. Myrskylä (2020). Us life expectancy stalls due to cardiovascular disease, not drug deaths. *Proceedings of the National Academy of Sciences 117*(13), 6998–7000.

Meslé, F. (2004). Mortality in central and eastern europe: long-term trends and recent upturns. *Demographic Research 2*, 45–70.

Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: a test using japanese cause of death data. Paper presented at the European Population Conference 2008, Barcelona, Spain.

Organization, W. H. Who mortality database. Available at www.who.int/data/data-collection-tools/who-mortality-database.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis*, Volume 2nd Ed. New York: Springer.

Remund, A., C. G. Camarda, and T. Riffe (2018). A cause-of-death decomposition of young adult excess mortality. *Demography 55*(3), 957–978.

Shi, P., A. Zhang, and H. Li (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics 10*(2), 1019–1040.

Stefanucci, M. and S. Mazzuco (2022). Analysing cause-specific mortality trends using compositional functional data analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 185*(1), 61–83.

Sun, Z., W. Xu, X. Cong, G. Li, and K. Chen (2020). Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *The Annals of Applied Statistics 14*(3), 1535–1556.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

University of California, B. U. and M. P. I. for Demographic Research (Germany). Human Mortality Database. Available at www.mortality.org or www.humanmortality.de.

Van De Water, H. P. (1997). Health expectancy and the problem of substitute morbidity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 352*(1363), 1819–1827.

Vaupel, J. W. and V. Canudas-Romo (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of nathan keyfitz's 90th birthday. *Demography 40*, 201–216.

Woolf, S. H. and H. Schoomaker (2019, 11). Life expectancy and mortality rates in the united states, 1959-2017. *Journal of the American Medical Association 322*(20), 1996–2016.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.