Pitfalls and merits of cointegration-based mortality models

Jarner, Søren F.; Jallbjørn, Snorre

Published in:
Insurance: Mathematics and Economics

DOI:
10.1016/j.insmatheco.2019.10.005

Publication date:
2020

Document version
Publisher’s PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Pitfalls and merits of cointegration-based mortality models

Søren F. Jarner *, Snorre Jallbjørn

Danish Labour Market Supplementary Pension Fund (ATP), Kongens Vænge 8, 3400 Hillerød, Denmark
University of Copenhagen, Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

Abstract

In recent years, joint modelling of the mortality of related populations has received a surge of attention. Several of these models employ cointegration techniques to link underlying factors with the aim of producing coherent projections, i.e. projections with non-diverging mortality rates. Often, however, the factors being analysed are not fully identifiable and arbitrary identification constraints are (inadvertently) allowed to influence the analysis thereby compromising its validity. Taking the widely used Lee-Carter model as an example, we point out the limitations and pitfalls of cointegration analysis when applied to semi-identifiable factors. On the other hand, when properly applied cointegration theory offers a rigorous framework for identifying and testing long-run relations between populations. Although widely used as a model building block, cointegration as an inferential tool is often overlooked in mortality analysis. Our aim with this paper is to raise awareness of the inferential strength of cointegration and to identify the time series models and hypotheses most suitable for mortality analysis. The concluding application to UK mortality shows by example the insights that can be obtained from a full cointegration analysis.

1. Introduction

Mortality models have many applications in areas such as demography, epidemiology, economics and actuarial sciences. In some applications we are interested in a single life expectancy projection, e.g. a unisex projection for a given country, but in many cases we are more interested in simultaneous projections for groups of related (sub)populations. Examples of the latter include joint modelling of males and females in a population, coherent forecasts for countries in a given region, projecting the life expectancy of smokers and non-smokers, modelling of insured lives relative to a national population, and assessing the effectiveness of a mortality hedge with the presence of basis risk; see e.g. Chen and Millosovich (2018), Kleinhov (2015), Bergeron-Boucher et al. (2017), Janssen et al. (2013), Jarner and Kryger (2011), Cairns et al. (2011a), Dowd et al. (2011) and Cairns et al. (2011b).

Models applied independently to separate populations often lead to diverging forecasts. This is the case even for closely related populations. Tulipurkar et al. (2000) found that applying the model of Lee and Carter (1992) separately to the G7 countries over a 50-year forecast horizon resulted in a life expectancy gap between the countries as large as eight years; despite the countries sharing long-term trends in mortality and convergence of social and economic factors. The projected divergence is due to small differences in the timing and magnitude of historical improvements being magnified by the separate analyses.

Joint mortality models are based on an assumption of non-divergence, or coherence, of mortality rates of the group of populations under consideration. Coherence is typically achieved by imposing a specific structure of the joint time series model used for forecasting factors driving mortality improvements in each population. Formally, the factors are assumed to cointegrate, i.e. to exhibit stationary relations preventing them from diverging. Cointegration theory offers a rigorous statistical framework for identifying and testing such stationary relations. However, this framework is rarely exploited in full since the structure is often imposed rather than tested, see e.g. Li and Lee (2005), Li and Hardy (2011), Dowd et al. (2011) and Cairns et al. (2011a).

The purpose of the present paper is twofold. First, we wish to advocate that cointegration analysis has more to offer than assuring coherence. Indeed, we will demonstrate the insights and “surprising” models that can arise from a full analysis. Second, we wish to highlight some of the pitfalls and limitations of cointegration analysis when applied to factors that are not fully identifiable, e.g. the mortality index of the popular Lee-Carter model. The overall message is that cointegration-based mortality models have much to offer as an inferential tool, but also that extreme care must be exercised when dealing with semi-identifiable factors often encountered.

* Corresponding author.
E-mail address: sj@atp.dk (S.F. Jarner).

https://doi.org/10.1016/j.insmatheco.2019.10.005
© 2019 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1.1. Cointegration

Cointegration is rooted in econometrics. It was introduced by Engle and Granger (1987) as a methodology for testing for stationary relations between non-stationary time series. The basic idea is that if two variables share a (stochastic) trend, it might be possible to find a linear combination of the variables that cancels the trend resulting in a stationary process. The linear combination is referred to as a cointegrating relation. The Engle–Granger methodology is limited to only a single cointegrating relation, while the more general and comprehensive setup developed by Johansen (1995) allows for an arbitrary number of variables and cointegrating relations, at least in principle. In effect, each (linearly independent) cointegrating relation reduces the dimension of the “driving forces” of the system by one.

While the aim of an econometric analysis is to infer and interpret the (economic) system, the typical focus of a mortality analysis is to produce a plausible forecast with proper quantification of the uncertainty. With the aim of improving gender-specific mortality forecasts, Carter and Lee (1992) suggested cointegration as a possible tool. In recent years, there has been a proliferation of papers using cointegration techniques to obtain coherent forecasts, see e.g. Darkiewicz and Hoedemakers (2004), Lazar and Denuit (2009), Njenga and Sherris (2011), Gaille and Sherris (2011), Yang and Wang (2013), Hyndman et al. (2013), Zhou et al. (2014), Hunt and Blake (2015c), Salhi and Loisel (2017), and Li and Lu (2017). Many of the authors arrive at complex, high-dimensional models which are difficult to interpret, but potentially good at forecasting.

We find that, although cointegration can certainly be used as a tool to impose coherence, the real strength of cointegration lies in inference and interpretation. We believe that this aspect is largely absent in the actuarial literature and that important subject knowledge can be gained from a more statistical approach to mortality modelling, a point also made by Arnold and Sherris (2016). A number of the cited papers do in fact test for cointegration rank as part of their model selection, but formulating and testing hypotheses on parameters is not part of the analysis. The primary aim of this paper is to demonstrate the value of cointegration-based inference in a mortality context.

Cointegration theory is a technically sophisticated field and some preliminary work is needed to establish the type of cointegration models and hypotheses suitable for mortality modelling. Once established, we present a cointegration analysis of male and female UK mortality. We consider both a two-dimensional analysis based on the Lee–Carter model and a four-dimensional analysis based on the logistic two-factor model of Cairns et al. (2006). In principle, any number of factors can be analysed, but to aid interpretation and the formulation of hypotheses it is useful to consider only a moderate number of factors.

1.2. Identifiability

Many mortality models, including the Lee–Carter model and its many variants, are overparametrized and parameters are therefore only identifiable after adding one or more constraints. In the Lee–Carter model, for example, two constraints are needed to ensure identification of the time-varying index and the age-specific parameters. By definition, the fitted mortality rates are unaffected by the identification scheme, but the forecasted mortality rates are not necessarily unaffected. Forecasts are based on time series models for the time-varying parameters and these models might not be invariant to the identification scheme.

Several recent papers have addressed the issue of identifiability and forecasting in mortality models, see in particular Nielsen and Nielsen (2014), Kuang et al. (2008), Hunt and Blake (2015a,b, 2017), and Beutner et al. (2017). In summary, this body of work shows that for forecasts to be unaffected by the identification scheme the time series model should be flexible enough to “preserve” reparameterizations, i.e. forecasting and reparameterizing should be interchangeable operations. As Hunt and Blake (2017) point out, this seemingly innocent requirement can in fact be at odds with model structures imposed to achieve coherence. We will return to this point later in the paper.

Identifiability issues affect the interpretation of parameters and – unless properly addressed – might lead to conclusions resting entirely on arbitrary constraints. Non-trivial issues arise even in the standard setting of an age–period model of the Lee–Carter type, and the complexity of the issues increase rapidly with the number of time-varying indices, see Hunt and Blake (2015b). Further issues arise in joint models with cointegrating parameters, which is the focus of this paper. In this case, the semi-identifiability of the parameters severely limits the choice of meaningful time series models and hypotheses.

To guarantee identification invariant inference, Nielsen and Nielsen (2014) advocate an approach based on maximal invariants of reduced dimensionality in terms of which all estimation and forecasting must be formulated. Although theoretically elegant, researchers might be reluctant to adopt this idea, since the interpretation of the original parametrization is lost. Also, we fear that the approach adds to the impression held by many that identifiability concerns are an esoteric topic which overcomplicate simple problems. In contrast to Nielsen and Nielsen (2014), we do not develop any formal theory in this paper, but rather illustrate by examples some of the pitfalls and problems of semi-identifiability. Hopefully, our exposition will be both accessible and illuminating to a wide audience. We also prefer to retain the original parametrization of the models to make the examples as relevant and familiar as possible. In these respects, our work is similar in spirit to the analysis of the gravity model presented in Hunt and Blake (2017).

1.3. Outline

The rest of the paper is organized as follows. In Section 2 we introduce the mortality models we will use as examples throughout and we discuss identification issues in the familiar setup of the Lee–Carter model. Section 3 covers background information on cointegration theory, while Section 4 specializes the discussion of cointegration to mortality models and illustrates the problems with applying cointegration to semi-identifiable parameters. Section 5 contains a comprehensive analysis of UK mortality data applying cointegration techniques to both the semi-identifiable Lee–Carter model and the fully identified model of Cairns et al. (2006). Finally, Section 6 offers some concluding remarks.

2. Mortality modelling

The object of study for mortality models is the age-specific death rates (ASDR’s) in a given population. We assume that data consist of death counts, $D_{t,x}$, and corresponding exposures, $E_{t,x}$, over time–age cells of the form $[t, t+1) \times [x, x+1)$ for a range of calendar years $t$ and (integer) ages $x$. We also assume that the underlying force of mortality, $\mu_{t,x}$, is constant over each of these cells. It then follows that $E[D_{t,x}] = \mu_{t,x}E[E_{t,x}]$, where $E$ denotes the expectation operator. The observed death rate is defined as

More precisely, $E$ denotes the conditional expectation given $\mu_{t,x}$, since $\mu_{t,x}$ is itself a stochastic quantity. Thus, $E$ averages over the random times of death given the force of mortality. Note that the exposure is also a stochastic quantity since it depends on the life spans. For modelling purposes, we (implicitly) condition on the exposures by treating them as fixed.
the ratio $m_{t,x} = D_{t,x}/E_{t,x}$. This is commonly used as an (empirical) estimate of the underlying force of mortality. When considering more than one population we add an identifying superscript, e.g., $\mu_{t,x}^i$ denotes the force of mortality at time $t$ and age $x$ in population $i$.

Most mortality models capture the time evolution of mortality rates (the period effect) by one, or more, time-varying indices (factors), $k_t$. Below we introduce the one-factor model of Lee and Carter (1992) and the two-factor model of Cairns et al. (2006).

### 2.1. The Lee–Carter model

The single population model proposed by Lee and Carter (1992) is used in a great number of mortality studies due to its simplicity and ease of interpretation; the ASDR’s are modelled by a log-linear relation where an age-dependent $a_x$ describes the general shape of the force of mortality and a single time-varying index, $k_t$, describes the speed of mortality improvements, governed by an age response $b_x$. The model is given by

$$
\log m_{t,x} = \log \mu_{t,x} + \varepsilon_{t,x} = a_x + b_x k_t + \varepsilon_{t,x},
$$

(2.1)

where $\varepsilon_{t,x}$ are homoscedastic error terms with mean 0 and variance $\sigma^2$.

We note that the model is invariant under the parameter transformations

$$\{a_x, b_x, k_t\} \mapsto \{a_x - b_x c, b_x / d, d(k_t + c)\},
$$

(2.2)

where $c \in \mathbb{R}$ and $d \in \mathbb{R} \setminus \{0\}$, in the sense that these transformations all yield the same model for $\log m_{t,x}$. In other words, the parameters are not fully identifiable since $k_t$ is only determined up to a linear transformation, $b_x$ up to a multiplicative constant, and $a_x$ up to a shift proportional to $b_x$. To ensure identification, the parameters are typically subject to the constraints

$$
\sum_i k_i = 0, \quad \sum_x b_x = 1.
$$

(2.3)

In Lee and Carter (1992) the parameters are estimated by ordinary least squares (OLS), i.e., by minimization of the quantity $\sum_i \sum_x [(\log m_{t,x}) - a_x - b_x k_t]^2$. Under the constraints (2.3), $\hat{a}_x$ equals the time average of $\log m_{t,x}$, and $\hat{b}_x$ and $\hat{k}_t$ can be obtained from the first component of a singular value decomposition of the matrix $[\log m_{t,x} - \hat{a}_x 1_x]$. The parameter estimates thus obtained establish the maximum likelihood estimates under the additional assumption that the errors are normally distributed.\(^2\)

The assumption of homoscedastic errors is questionable as we would expect observed death rates to fluctuate more when death counts are low. In addition, the use of OLS has the practical problem of how to handle cells with zero death counts frequent in small data sets. In the application to UK data we use instead the Poisson variant of the Lee–Carter model proposed by Brouhns et al. (2002). The two Lee–Carter variants have the same parametric structure and the points made later regarding identification issues in relation to forecasting and cointegration analysis therefore apply to both of them.

#### 2.1.1. Forecasting

The time-varying index is typically modelled as a random walk with drift,

$$
k_t = k_{t-1} + \theta + \epsilon_t,
$$

(2.4)

where $\theta$ is the drift and the $\epsilon_t$’s are i.i.d. $\mathcal{N}(0, \sigma^2)$. The drift and variance are estimated by the sample mean and sample variance, respectively, of the differences $k_t - k_{t-1}$.

Let $T$ denote the last year of data. Forecasting is based on the conditional distribution of $k_{T+h}$ given $k_T = k_T$; we have for $h \geq 1$

$$
k_{T+h} = k_T + h \theta + h \epsilon_T
$$

(2.5)

from which forecasts and confidence intervals can be derived. In particular, a (median) forecast of future mortality rates is given by

$$
\hat{\mu}_{T+h,x} = \exp \left( \hat{a}_x + \hat{b}_x \left( k_T + h \theta \right) \right) = \hat{\mu}_{T,x} \exp \left( h \hat{\epsilon}_T \right).
$$

(2.6)

### 2.1.2. Coherence

The concept of coherent forecasts was introduced by Li and Lee (2005) and formalized by Hyndman et al. (2013). Mortality forecasts for two populations are said to be coherent if the relative mortality rates converge for each age $x$,

$$
\frac{\hat{\mu}^{i}_{T,x}}{\hat{\mu}^{j}_{T,x}} \to R_x, \quad t \to \infty,
$$

(2.7)

for positive, age-specific constants $R_x$. When producing forecasts for populations with historically similar mortality evolutions, the concept of coherence formalizes the intuitively desirable property that the forecasts should reflect these similarities.

On the other hand, coherence is a rather strict requirement which will generally not be satisfied by forecasts obtained by applying e.g. the Lee–Carter model to separate populations. Indeed, it follows from (2.6) that in a Lee–Carter setting a necessary and sufficient condition for coherence is that $d \mu^{i} = \theta d^{i}$ and $b^{i}_x = b^{j}_x$ for all $x$. In practice, of course, this will never happen. With the aim of obtaining coherent forecasts for a group of populations, Li and Lee (2005) proposed the augmented common factor model

$$
\log m_{t,x}^{i} = \log (\mu_{t,x}^{i} + \epsilon_{t,x}^{i} + a^{i}_x + b^{i}_x k^{i}_t + \epsilon_{t,x}^{i}),
$$

(2.8)

This model produces coherent forecasts when the population specific indices, $k^{i}_t$, are modelled as stationary processes, e.g., AR(1)-processes. The common factor, $K_t$, can be non-stationary, e.g. a random walk with drift as in the original Lee–Carter model.

The notion of coherence has undoubtedly been very influential in setting the standard for joint forecasts. Indeed, many joint mortality models have been devised with the specific aim of achieving coherence, as mentioned in the introduction. The model of Li and Lee (2005) can be seen as an early and very direct way to ensure coherence by equating the driving factors, while the more recent approaches typically combine a specific structure with cointegrating relations, e.g. Dowd et al. (2011). Generally, cointegrating relations do not guarantee coherence, although in the Lee–Carter setting the two concepts are closely linked. We will return to this point in Section 4.3.

It can be argued that coherence is too strict a requirement and that models enforcing coherence risk violating the historic pattern of covariation between populations, see Hunt and Blake (2017). Arguably, it is better to identify cointegrating relations which restrict the joint forecasts in plausible ways, than to insist on coherence. This point will be illustrated in the application section.

#### 2.2. The Cairns–Blake–Dowd model

Originally made to accommodate the British pension market, the model of Cairns et al. (2006) focuses primarily on the post-age
60 mortality curve and the pricing of immediate life annuities. The model fits the mortality curve by a logistic curve\(^3\)

\[
\logit(q_{t,x}) \triangleq \log \left( \frac{q_{t,x}}{1-q_{t,x}} \right) = k_{1,1} + k_{1,2}(x-\bar{x}), \tag{2.9}
\]

where \(\bar{x} = \frac{1}{N} \sum \chi_i\) is the arithmetic mean of the \(N\) ages considered and \(q_{t,x} = 1 - \exp(-\mu_{t,x})\) is the probability for an individual aged \(x\) at time \(t\) to die before \(t+1\). Note that the logit transform is well-defined since \(q\) lies between 0 and 1. We also note that \(q_{t,x} \approx \mu_{t,x}\) for small \(\mu_{t,x}\).

Under the CBD model the logit-transformed curve of death probabilities is linear in age with time-varying parameters. The first index is the level of the line, and a decreasing trend in \(k_{1,1}\) thus represents an overall improvement in mortality over time. The second index is the slope of the line, and an increasing trend in \(k_{1,2}\) thus implies a steepening of the mortality curve. The model is fully identified, since there are no invariant parameter transformations.

As is customary, we treat the model as a generalized linear model (GLM) within the binomial family with its canonical logit-link function. In principle, parameters can also be estimated by maximum likelihood assuming Poisson distributed death counts, see Currie (2016) for a comparison of the two approaches. The points made in this paper apply regardless of how parameters are estimated.

2.2.1. Forecasting

Forecasting is performed assuming a bivariate random walk with drift for the two time-varying indices

\[
\begin{pmatrix}
k_{t,1} \\
k_{t,2}
\end{pmatrix} \sim \begin{pmatrix}
k_{t-1,1} \\
k_{t-1,2}
\end{pmatrix} + \begin{pmatrix}
\theta_1 \\
\theta_2
\end{pmatrix} + \epsilon_t, \quad \epsilon_t \sim N_2(0, \Sigma). \tag{2.10}
\]

This projection method entails a dependency structure between the two time-varying indices by allowing for covariance, but omits the possibility of the indices being directly affected by the previous value of one another. As for the Lee–Carter model, the argument for the use of a random walk with drift is that it is often adequate to describe the data. Of course, if deemed necessary, more complicated ARIMA models can be used.

In the application section we shall model female and male mortality by two CBD models with cointegrating parameters. This will give rise to a sort of “logit coherence” rather than the usual (log) coherence.

2.3. Identification invariance

Identification issues arise in many fields of statistics, and mortality modelling is no exception. The problem arises from the fact that many (mortality) models are overparameterized such that there exist different sets of parameters yielding the same fit. From a statistical point of view all these sets are equally good, but to perform the analysis in practice the researcher has to choose one specific set. Therefore, constraints are imposed identifying one of the equivalent parameter sets over the others. The question now arises whether the (arbitrary) choice of constraints influence the forecast, and more generally the statistical inference.

To illustrate the problem, consider the familiar case of the Lee–Carter model as introduced above. Let \((\bar{a}_1, \bar{b}_1, \bar{k}_1)\) denote parameter estimates under the constraints (2.3), and consider the equivalent parameters \((\bar{a}_2, \bar{b}_2, \bar{k}_2)\) for given \(c \in \mathbb{R}\) and \(d \in \mathbb{R} \setminus \{0\}\). We forecast the time-varying index by the random walk with drift of (2.4). Using the mean difference as estimator for the drift we have \(\hat{\theta}_t = \hat{d}_t\), and thereby \(\theta_b = \hat{\theta}_b\) for all \(x\). Since also \(\hat{\mu}_{T,x} = \tilde{\mu}_{T,x}\), it follows from (2.6) that for \(h \geq 1\)

\[
\hat{\mu}_{T+h,x} = \tilde{\mu}_{T,x} \exp \left( \hat{h}\hat{\theta}_b \right) = \tilde{\mu}_{T,x} \exp \left( \hat{h}\hat{\theta}_b \right) = \tilde{\mu}_{T+h,x}. \tag{2.11}
\]

This shows that the forecast obtained by the standard Lee–Carter method is in fact invariant to the chosen identification scheme.

However, consider now the case where the time-varying index is modelled as the random walk with drift of (2.4), but with a fixed drift term \(\theta = \theta_0\). We then have

\[
\hat{\mu}_{T+h,x} = \tilde{\mu}_{T,x} \exp \left( \hat{h}\hat{\theta}_b \right) = \tilde{\mu}_{T,x} \exp \left( \hat{h}\hat{\theta}_b \right). \tag{2.12}
\]

which only equals \(\tilde{\mu}_{T+h,x}\) when \(\theta_b = 0\). Hence, the forecasts will only be the same if \(\theta_0 = 0\) (or \(\hat{\theta}_b = 0\) for all \(x\)). Mathematically, the problem is that forecasting and reparameterization are no longer interchangeable operations or, in other words, the restricted forecasting model has different meaning for different parameterizations. One might argue that this is a contrived example as one would never consider this model, but very similar problems arise in cointegrated models where the identification of meaningful hypotheses is much less obvious.

Despite the ease with which specific problems related to lack of identification can be identified, it is surprisingly hard to formulate and justify a general principle that models and inferential procedures must adhere to.\(^4\) Indeed, suggested principles often sound a bit vague: Hunt and Blake (2017) use the term “well-defined” for models giving the same projected mortality rates for any set of identifiability constraints, and Nielsen and Nielsen (2014) talk about “avoiding arbitrariness resulting from the identification process”. We propose to use the term “identification invariance” when reparameterization and inference (including forecasting) are interchangeable, cf. Fig. 1. We consider identification invariance a fundamental property of a sound statistical analysis.

Schematically, identification invariance is similar to the classical notion of “parametrization invariance”, see e.g. Lindsey (1996). However, where parametrization invariance requires inferential invariance to all one-to-one reparameterizations, identification invariance requires only invariance to parameter transformations induced by different identification constraints. In practice, this amounts to inferential invariance to a specific set of

\[^4\] Part of the problem seems to be that some researchers consider the constraints as an intrinsic part of the model, and not merely as (arbitrary) mathematical constraints needed for identification. From this perspective, different constraints imply different models and therefore “naturally” lead to different forecasts (even though the models are statistically identical in terms of describing the observed data).

---

\(^3\) The use of a logistic model as a suitable choice for examining age patterns of human adult mortality is well established, see e.g. Thatcher (1999).
linear parameter transformations. Note that, since time-varying parameters are typically modelled by linear time series models, the inference is generally not invariant to non-linear parameter transformations, i.e. the inference is generally not fully parametrization invariant.

3. Cointegration theory

In the following we give a brief introduction to cointegrated vector autoregressive (VAR) models, including interpretation and testing of hypotheses; unless explicitly stated otherwise, the exposition relies on Johansen (1995). In the subsequent sections this framework will be applied to mortality modelling.

A p-dimensional VAR(k)-model is a model of the form

\[ y_t = \Pi^1 y_{t-1} + \cdots + \Pi^k y_{t-k} + \Phi D_t + \varepsilon_t, \tag{3.1} \]

where \( \varepsilon_t \sim N_p(0, \Sigma) \) are i.i.d. errors and \( D_t \) contains all determinant terms such as constant, trend and dummy variables. The evolution of each variable in the VAR-model is based on its own lagged values as well as the lagged values of the other variables in the system. This formulation highlights the short-term dynamics, while possible long-run relations between the variables are hard to discern. In order to study long-run relations we introduce the notion of cointegration. First a few preliminary definitions.

A linear process is defined by \( \bar{y}_t = \sum_{i=0}^{\infty} C_i y_{t-i} \), where \( C_i \) are i.i.d. with mean zero and finite variance and \( \mathbb{C}(z) = \sum_{i=0}^{\infty} C_i z^i \) is convergent for \(|z| < \delta\) for some \( \delta > 1\). An \( l(0) \) process is a linear process with the additional requirement that \( \sum_{i=0}^{\infty} C_i \neq 0 \), or such a process with a deterministic trend added.\(^5\)

The difference operator, \( \Delta \), is defined by \( \Delta y_t = y_t - y_{t-1} \). A stochastic process \( \bar{y}_t \) is called integrated of order 1, or \( l(1) \), if \( \Delta(1) \bar{y}_t \) is an \( l(0) \) process. Loosely speaking, the stochastic component of an \( l(1) \) process behaves like a random walk.

**Definition 3.1.** Let \( \bar{y}_t \) be integrated of order 1. We say that \( \bar{y}_t \) is cointegrated with cointegrating vector \( \beta \neq 0 \) if \( \bar{y}_t - \mathbb{E}[\hat{\beta}' \hat{y}_t] \) admits a stationary distribution. The cointegrating rank is the number of linearly independent cointegrating vectors, and the cointegration space is the space spanned by the cointegrating vectors.

3.1. The vector error correction model

We are interested in conditions for the VAR-model to be integrated of order 1. For this purpose we subtract \( y_{t-1} \) from both sides of (3.1) and rearrange terms to obtain the equivalent vector error correction model (VECM), where the increment is expressed in terms of differences, lagged differences and the level of the process itself

\[ \Delta y_t = \Pi^1 y_{t-1} + \cdots + \Pi^k y_{t-k} + \Phi D_t + \varepsilon_t, \tag{3.2a} \]

\[ \Pi = (\Pi^1 + \cdots + \Pi^k) - I, \tag{3.2b} \]

\[ \Gamma_i = - (\Pi^{i+1} + \cdots + \Pi^k), \quad i = 1, \ldots, k-1. \tag{3.2c} \]

The behaviour of \( \bar{y}_t \) is mostly studied in terms of its characteristic polynomial given by \( \mathbb{A}(z) = (1 - z)I - \Pi^1 z - \cdots - \Pi^k z^k \) with determinant \( \mathbb{A}(z) \). If \( \mathbb{A} \) has a unit root then \( \Pi = \mathbb{A}(1) \) is singular and the process is non-stationary. In this case, \( r = \text{rank}(\Pi) < p \) and there exist two \( p \times r \) matrices \( \alpha \) and \( \beta \) such that

\[ \Pi = \alpha \beta'. \tag{3.3} \]

This is essentially the requirement for \( l(1) \). However, to avoid explosive and seasonal roots and to ensure invertibility we also need the following technical condition.

**Condition 3.2.** If \( |\mathbb{A}(z)| = 0 \), then either \( |z| > 1 \) or \( z = 1 \). Further, the matrix \( \alpha \perp \Gamma \beta' \perp \) has full rank, where \( \Gamma = I - \sum_{i=1}^{k-1} \Gamma_i \) and \( \alpha \perp \beta \) are \( p \times (p-r) \) matrices spanning the orthogonal complement of \( \text{span}(\alpha) \) and \( \text{span}(\beta) \), respectively.

None of the matrices \( \alpha \), \( \alpha \perp \beta \) or \( \beta \) are uniquely defined, but the conditions and conclusions do not depend on which versions we use. Let \( \Gamma'z = 1 - \sum_{i=1}^{k-1} \Gamma_i z^i \) whereby \( \Gamma'(1) = \Gamma \) and \( \Gamma'(L)y_0 = y_0 - \sum_{i=1}^{k-1} \Gamma_i y_{-i} \), where \( L \) is the lag operator. We can now formulate the celebrated Granger Representation Theorem in a version due to Hansen (2005).

**Theorem 3.3.** If \( |\mathbb{A}(1)| = 0 \) and Condition 3.2 is satisfied, then \( \bar{y}_t \) can be represented as the sum of a random walk and a stationary process

\[ \bar{y}_t = \mathbb{C} \sum_{i=1}^{T} (\varepsilon_t + \Phi D_t) + \sum_{i=0}^{\infty} \mathbb{C}_i (\varepsilon_{t-i} + \Phi D_{t-i}) + \mathbb{C} \Gamma'(L)y_0, \tag{3.4} \]

where \( \mathbb{C} = \beta \perp (\alpha \perp \Gamma \beta' \perp)^{-1} \alpha \perp \beta \), and \( \mathbb{C}_i \) is defined recursively by

\[ \mathbb{C}_i = (I + \Pi) \mathbb{C}_{i-1} + \sum_{j=1}^{k-1} \Gamma_j \Delta \mathbb{C}_{i-j}, \quad i = 1, 2, \ldots. \]

with \( \mathbb{C}_0 = I - \mathbb{C} \) and \( \mathbb{C}_{-1} = \cdots = \mathbb{C}_{-k+1} = -\mathbb{C} \). In particular, if \( r > 0 \) then \( \bar{y}_t \) is a cointegrated \( l(1) \) process with cointegrating vectors \( \beta \).

Intuitively, the process evolves as a random walk in \( \text{span}(\beta) \) while at the same time it tries to establish the equilibrium relation for \( \beta y_t \) with a force that depends on the adjustment coefficients \( \alpha \) and the equilibrium error \( \beta y_t - \mathbb{E}[\hat{\beta}' \hat{y}_t] \). The factorization (3.3) defines the cointegration space, but the individual cointegrating relations are not unique without further normalization. Johansen (1995) suggests letting the first part of \( \beta \) be a \( r \)-dimensional identity matrix making for a just-identified normalization and we adopt this approach throughout without further notification.

3.2. Deterministic terms and trends

The deterministic term is an important part of the specification of the model affecting both the trend of \( \bar{y}_t \) and the test statistics for cointegration rank. Under the conditions of Theorem 3.3, the process \( \bar{y}_t \) has, in general, a trend of the form \( \mathbb{C} \Phi D_0 + \sum_{i=0}^{\infty} \mathbb{C}_i \Phi D_{-i} \). Note that we refer to the deterministic part of \( \bar{y}_t \) as a trend, regardless of its order.

In general, the deterministic terms accumulate to a trend one order higher. More precisely, however, it is only the terms \( \mathbb{C} \Phi D_t \) that accumulate to a higher order. To illustrate the implication of this in more detail, we consider deterministic terms of the form

\[ \Phi D_t = \theta_0 + \theta_t t, \tag{3.5} \]

for \( p \)-dimensional vectors \( \theta_0 \) and \( \theta_t \). Following Johansen (1995), we decompose each \( \theta_t \) as \( \theta_t = \alpha_1 y_t + \alpha_2 r_p \). Since \( \alpha_1 \) is 0, it follows from the Granger representation (3.4) that only \( \alpha_1 y_t + \alpha_2 r_p \) enters into the \( t \)th term of the random walk component. In particular, there is a quadratic trend in the level of the process with coefficient \( \alpha \beta = 1 \).
Table 1  
Trending behaviour of the VECM for five nested models for the deterministic term of form (3.3). Starting with no restrictions, the models are defined by successively setting \( \gamma_1, \gamma_2, \beta_1, \) and \( \beta_2 \) to zero.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deterministic term</th>
<th>Trend in ( y_t )</th>
<th>( \mathbb{E} [\Delta y_t] )</th>
<th>( \mathbb{E} [g y_t] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0(r) )</td>
<td>( \theta_0 )</td>
<td>Quadratic</td>
<td>Linear</td>
<td>Linear</td>
</tr>
<tr>
<td>( H_1(r) )</td>
<td>( \theta_0 + \gamma_1 )</td>
<td>Linear</td>
<td>Constant</td>
<td>Linear</td>
</tr>
<tr>
<td>( H_2(r) )</td>
<td>( \theta_0 + \gamma_1 + \beta_1 t )</td>
<td>Linear</td>
<td>Constant</td>
<td>Constant</td>
</tr>
<tr>
<td>( H_3(r) )</td>
<td>( \alpha \beta_1 )</td>
<td>Constant</td>
<td>Zero</td>
<td>Constant</td>
</tr>
<tr>
<td>( H_4(r) )</td>
<td>0</td>
<td>Zero</td>
<td>Zero</td>
<td>Zero</td>
</tr>
</tbody>
</table>

The decomposition of \( \theta_1 \) gives rise to five nested models defined by restricting the number of non-zero terms. The models and the trending behaviour they entail are summarized in Table 1.

3.3. Cointegration rank and parameter estimation

In some situations, the cointegration rank can be justified on the basis of prior knowledge. Often, however, the cointegration rank needs to be inferred from the data. Let \( H(r) \) denote the hypothesis that \( \Pi = \alpha \beta \) for two \( p \times r \) matrices \( \alpha \) and \( \beta \). Without further restrictions, this is equivalent to the hypothesis that \( \text{rank}(\Pi) \leq r \). This creates a set of nested hypotheses

\[
H(0) \subset \cdots \subset H(r-1) \subset \cdots \subset H(p).
\]

The cointegration rank can be determined by testing these hypotheses sequentially, starting from \( H(0) \) and stopping when the first acceptance is encountered. The cointegration rank is \( r \), say, if \( H(0), \ldots, H(r-1) \) are rejected, while \( H(r) \) is accepted. Johansen (1995) derives the likelihood-ratio test, known as the trace test, for performing these tests. The distribution of the test statistic is non-standard and it depends on the specification of the deterministic term. Critical values are tabulated in Section 15.3 of Johansen (1995) for the five models considered in Section 3.2. The testing procedure is complicated by the fact that we might need to infer the model for the deterministic term and the cointegration rank simultaneously. We will return to this point in the application section.

Given the (maximal) cointegration rank and specification of the deterministic term, the maximum likelihood estimates of the parameters are obtained by reduced rank regression. For completeness, the estimates and the trace test statistic can be found in the Appendix.

4. Cointegration in mortality models

In this section we first discuss the trend models most relevant in a mortality context. Next, we show how the identification issues of the Lee–Carter model severely limit the set of testable (cointegration) hypotheses. Finally, we comment on alternative approaches to cointegration within the Lee–Carter framework.

4.1. Linear trend models

We restrict our attention to the case of analysing the period effect within a given parametric mortality model. Assume that \( k_t \), consisting of the combined time-varying indices of (separate) age–period models, can be shown to form an \( l(1) \) process. This is a reasonable assumption, since we expect the period effect to cause at least one of the time-varying indices (for each population) to accumulate annual improvements over time and hence to behave like a random walk with drift. In general, the drift term itself could be time-varying giving rise to trends of all shapes and orders. Indeed, Arnold and Sherris (2016) find quadratic trends when analysing cause-specific mortality. However, for the purpose of this exposition we focus on linear trends only, which are, arguably, also the only type of trends suitable for robust forecasting.

To characterize the relevant models in more detail, let us consider deterministic terms of the form \( \Phi D_t = \theta_0 + \theta_1 t \), where \( \theta_1 = \gamma_1 + \alpha \rho_1 \) for \( i = 0, 1 \), cf. Section 3.2. The absence of a quadratic trend implies that \( \gamma_1 = 0 \). Hence, in the current context the largest model of interest is \( H_0(r) \) of Table 1. Under this model, \( k_t \) has the representation

\[
k_t = \gamma_0 + \theta_1 t + \alpha \rho_1 t \sum_{i=1}^{\infty} c_i \varepsilon_{i-1} + \Gamma' \{ \varepsilon_L \} \theta_0;
\]

expressions for the intercept, \( \gamma_0 \), and slope, \( \theta_1 \), can be found in Hansen (2005). It can be shown that the cointegrating relations are trend stationary, i.e. they can be decomposed as a linear trend and a stationary process, \( \epsilon \), as \( \beta' k_t = \beta' \gamma_0 + \theta_1 t + \epsilon \). Thus, parameters drift further and further apart over time even though they engage in an equilibrium correcting relationship.

The second model of interest for mortality modelling is \( H_1(r) \) of Table 1. This model has \( \rho_1 = \gamma_1 = 0 \), i.e. the previous model with the further restriction that the cointegrating relations do not trend. The level of the process still possesses a linear trend. Demographically, the lack of a trend in the cointegrating relations is appealing, but it cannot be assumed in advance. In the application section we will use the statistical setup of Johansen (1995) to test for \( \rho_1 = 0 \).

Technically, it is also possible to test for further model restrictions, i.e. model \( H_2(r) \) for absence of a linear trend altogether and model \( H_3(r) \) for zero mean. However, neither of these latter models are relevant for modelling mortality data with period effects.

As described in Section 2, applications of the Lee–Carter model and the CBD-model often employ a simple random walk with drift to describe the time-varying indices. In the spirit of preserving as much of the marginal structure as possible, the natural candidate for joint modelling is therefore the VECM with zero lagged differences and no quadratic trend

\[
\Delta k_t = \alpha (\beta k_{t-1} + \rho_1 t) + \theta_0 + \varepsilon_t = \alpha (\beta k_{t-1} + \rho_1 t + \alpha \gamma_0 + \varepsilon_t).
\]

For this model, the linear trend of (4.1) is given by \( \tau_1 = \theta_1 C \theta_0 - \alpha (\beta' \alpha)^{-1} \rho_1 \). Further, if \( \rho_1 = 0 \) then \( \tau_1 = C \theta_0 \) and \( \mathbb{E} [\beta' k] = -\beta' \theta_0 \), and we have

\[
\Delta k_t = \alpha \beta' k_{t-1} + \theta_0 + \varepsilon_t = \alpha (\beta' k_{t-1} - \mathbb{E} [\beta' k]) + \tau_1 + \varepsilon_t.
\]

Models (4.2)–(4.3) will be our workhorse models in the application section.

4.2. Identification invariance of cointegrated Lee–Carter models

The Lee–Carter model is the predominant single population mortality model. At first thought, it therefore seems natural to use the Lee–Carter model as the underlying model for a joint analysis of related populations. However, it turns out that the identifiability issues of the Lee–Carter model severely limit its usefulness for this purpose.

The lack of identifiability of the Lee–Carter model and the consequences for interpretation and forecasting have also been studied by other authors, see in particular Nielsen and Nielsen (2014) and Hunt and Blake (2017). In contrast to these contributions, we here focus on how the identifiability issues restrict the set of testable hypotheses, i.e. the hypotheses that can form part of a statistical analysis. More precisely, we are interested in characterizing the identification invariant cointegration models for the time-varying mortality index of two Lee–Carter models, i.e. the
cointegation models for which forecasting and reparametrization of the underlying Lee–Carter models are interchangeable, cf. Section 2.3.

Let \( \mathbf{k}_t = (k^1_t, k^2_t) \) denote the vector of mortality indices of two Lee–Carter models with given identification schemes. Assume that we choose to model this as a VECM with a linear trend of class \( H^2_0(r) \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Drift</th>
<th>Invariant</th>
<th>Hypothesis</th>
<th>Testable</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0(r) )</td>
<td>( \theta_0 + \beta_1 t )</td>
<td>✓</td>
<td>( \alpha = \Lambda \psi )</td>
<td>( \psi )^{\top}</td>
</tr>
<tr>
<td>( H^2_0(r) )</td>
<td>( \theta_0 + \alpha^1_1 t )</td>
<td>✓</td>
<td>( \beta = \Lambda \psi )</td>
<td>( \psi )^{\top}</td>
</tr>
<tr>
<td>( H_1(r) )</td>
<td>( \theta_0 )</td>
<td>✓</td>
<td>( \psi = \theta = \theta )</td>
<td>( \psi )^{\top}</td>
</tr>
<tr>
<td>( H^2_1(r) )</td>
<td>( \alpha^1_0 )</td>
<td>✓</td>
<td>( \rho_0 = \rho )</td>
<td>( \psi )^{\top}</td>
</tr>
<tr>
<td>( H^2_2(r) )</td>
<td>0</td>
<td>✓</td>
<td>( \rho_1 = \rho )</td>
<td>( \psi )^{\top}</td>
</tr>
</tbody>
</table>

Table 2: Overview of models for deterministic terms (drift) and hypotheses for parameter constraints. A check mark indicates that the model/hypothesis is identification invariant/testable in the context of a VECM for the mortality indices of two Lee–Carter models, and a minus sign that it is not.

In summary, obeying identification invariance we can infer the cointegration rank and distinguish between the two types of linear trend models most relevant for mortality modelling (and also between other less relevant models). However, within a given model class we can in general not restrict parameters further without violating identification invariance, i.e. essentially all hypotheses of interest are non-testable. The situation is summarized in Table 2. For someone interested purely in joint forecasting of two populations this might not pose a problem. However, if the aim is to analyse the nature of the joint behaviour in more detail, cointegrated Lee–Carter models are of limited use.

4.3. Alternative approaches of the Lee–Carter type

Arguably, the simplest way to avoid the problems due to lack of identifiability of the Lee–Carter model is to use another model. Indeed, if we base our (joint) analysis on a fully identifiable mortality model, e.g. the CBD model, all (joint) hypotheses are well-defined and testable. However, due to the familiarity and widespread use of the Lee–Carter model some researchers might be reluctant to follow this route. For that reason, we consider below two alternative approaches to obtain identification invariant inference within the Lee–Carter framework.

The first approach is to impose further restrictions on the parameters of the underlying Lee–Carter models, thereby implicitly restricting the set of invariant transformations of the mortality indices. As an example of this approach, Zhou et al. (2014) assume equality of the age response parameters of two Lee–Carter models, i.e. \( b_y = b^*_y = b^*_x \) for all \( x \).

\[
\log \mu^1_{t,x} = \log \mu^2_{t,x} + \epsilon^1_{t,x} = d^1_i + b^*_x k^2_i + \epsilon^1_{t,x} \quad \text{for} \quad i = 1, 2.
\]  

This is similar in spirit to the augmented common factor model of Li and Lee (2005), cf. (2.8). Let \( \mathbf{k}_t = (k^1_t, k^2_t) \) denote the vector of mortality indices for given identification scheme. Due to the constraint on the \( b \)-parameters, different identification schemes lead to mortality indices of the form \( \mathbf{k}_t = d(\mathbf{k}_t + \mathbf{c}) \), with \( d \in \mathbb{R} \setminus \{0\} \), and \( \mathbf{c} = (c^1, c^2)^\top \) with \( c^1, c^2 \in \mathbb{R} \).

Applying the same transformation to (4.4) yields

\[
\Delta \mathbf{k}_t = \sum_{i=1}^{l-1} \Gamma^i \Delta k_{t-i-1} + \mathbf{D} \left( \beta^\top (D^{-1} \Delta \mathbf{k}_{t-1} - \mathbf{c}) \right) + \mathbf{D} \theta_0 + \mathbf{D} \mathbf{e}_t
\]

(4.5)

\[
= \sum_{i=1}^{l-1} \Gamma^i \Delta k_{t-i-1} + \alpha (\beta^\top k_{t-1} + \rho_1 t) + \theta_0 + \mathbf{e}_t,
\]

(4.6)

where \( \Gamma^i = D^i \Gamma D^{-i} \), \( \alpha = \mathbf{D} \alpha \), \( \beta = D^{-1} \beta \), \( \theta_0 = \mathbf{D} \theta_0 = \mathbf{D} \mathbf{e}_t \), and \( \mathbf{e}_t \sim \mathcal{N}_2(0, \Sigma \mathbf{D}) \). Since \( \mathbf{D} \) and \( \mathbf{D}^{-1} \) have full rank, rank \( (\alpha \beta^\top) = \text{rank}(\tilde{\alpha} \tilde{\beta}^\top) \), and it follows that (4.6) belongs to \( H^2_0(r) \) for the same value of \( r \) as in (4.4). Thus, in the context of cointegrated Lee–Carter models, \( H^2_0(r) \) is identification invariant, cf. Fig. 1.

Similar calculations show that the first four model classes of Table 1 are all identification invariant, while the fifth, \( H^2_2(r) \), is not (unless \( r = 0 \)). In other words, constant and linear trends are preserved by linear transformations of the two indices being modelled—as long as we allow all other parameters to vary freely.

While it is generally valid to impose different models for the deterministic terms, e.g. \( H^2_0(r) \) or \( H^2_1(r) \), for freely varying parameters, it is generally not valid to test or impose further parameter constraints. Indeed, since \( \mathbf{D} \) is a diagonal matrix with an arbitrary, non-zero diagonal and since \( \tilde{\alpha} = \mathbf{D} \alpha \) and \( \tilde{\beta} = D^{-1} \beta \), constraints on either \( \alpha \) or \( \beta \) will typically not be satisfied by the transformed model. In particular, the hypothesis of prime interest, \( \beta \propto (1, -1)^\top \), is not testable, i.e. the corresponding model is not identification invariant. Nor can we test hypotheses on the (relative) magnitude of the adjustment coefficients. In fact, apart from the cointegration rank, the only testable hypothesis of some demographic interest is \( \rho_1 = \rho \) for given \( \rho \). The \( \rho_1 \) parameter is identifiable and can be interpreted as a ‘divergence’ measure between (related) populations.

In summary, obeying identification invariance we can infer the cointegration rank and distinguish between the two types of linear trend models most relevant for mortality modelling (and also between other less relevant models). However, within a given model class we can in general not restrict parameters further without violating identification invariance, i.e. essentially all hypotheses of interest are non-testable. The situation is summarized in Table 2. For someone interested purely in joint forecasting of two populations this might not pose a problem. However, if the aim is to analyse the nature of the joint behaviour in more detail, cointegrated Lee–Carter models are of limited use.
An alternative approach is to consider the vector of log mortality rates as a multivariate time series (of high dimension) and model this series directly, i.e. to model directly the N-dimensional series \( y_t = (\log m_{x_{1,t}}, \ldots, \log m_{x_{N,t}}) \). By construction, there are no factors and hence no identification issues related to factor identification, but due to the high dimension the time series models are more complex and harder to interpret.

Lazar and Denuit (2009) use the cointegration methodology of Section 2 to model \( y_t \). In this framework, the Lee–Carter model is a special case with cointegration rank \( N - 1 \) corresponding to one common stochastic trend. Lazar and Denuit (2009) focus on single-population modelling, but the approach extends readily to multi-population modelling by stacking the \( y \)-vectors. The VAR/VECM approach to modelling \( y_t \) is also explored in the recent papers by Salhi and Loisel (2017) and Li and Lu (2017).

As mentioned in Section 2.1.2, coherence has received much attention as a desirable property of mortality forecasts. In the Lee–Carter setting of Section 4.2 coherence corresponds to stationarity of \( b_1 k^{\prime}_t - b_2 k^{\prime}_t \) for all \( x \). Note that, cointegrated mortality indices, i.e. stationarity of \( \beta' k_t \), for some \( \beta \), do not in itself guarantee coherence. Indeed, for (strict) coherence we must have \( b_x \propto \beta \) for all \( x \), where \( b_x = (b_1^{\prime}, -b_2)^{\prime} \).

In practice, this will never be (strictly) satisfied, unless enforced by design as in Zhou et al. (2014). In contrast, when modelling \( y_t \) directly non-diverging rates for different populations (coherence) or for different ages within the same population can more easily be obtained while preserving model flexibility, see e.g. the model of Li and Lu (2017).

The VAR/VECM approach to modelling \( y_t \) directly certainly has its merits as a flexible method for forecasting capable of capturing the dependency structure across ages. The approach also provides a useful framework for determining the number of common stochastic trends, i.e. the dimension of the driving dynamics. However, due to the high dimension the resulting models are often very complex and hard to interpret. For the purpose of gaining demographic insights by formulating and testing hypotheses we find that more parsimonious models with a limited set of interpretable factors are better suited.

5. Applications to UK mortality data

In this section we present two applications of cointegration-based mortality models. The applications focus on the inferential procedure, in particular hypothesis testing and interpretation of the models, rather than on the resulting forecasts. We use UK mortality data for males and females retrieved from the Human Mortality Database (2019).\(^7\)

The first application is based on the Lee–Carter model and illustrates the care with which results must be interpreted due to semi-identifiability. The second application is based on the fully identified CBD-model for which a more detailed analysis is possible. In both cases, the analysis starts with a visual inspection of the mortality indices being modelled and tests for non-stationarity.

5.1. Lee–Carter application

The period remaining life expectancy is a useful summary measure for the mortality conditions of a population at a given point in time. Fig. 2 shows the period remaining life expectancy at birth and at age 60 for UK males and females, calculated using the observed death rates \( m_{x,t} = d_{x,t}/e_{x,t} \). The most striking feature is the remarkable increase in life expectancy over the period; an increase also seen in most other developed countries. Another prominent feature is the degree of similarity between the life expectancy evolution of the two genders. This prompts the scientific question of whether the joint behaviour is the result of two random walk-type processes with similar drift and correlated innovations, or whether the two processes do in fact engage in a cointegrating relation? As a first attempt at answering this question, we consider a cointegration analysis of the mortality indices of two Lee–Carter models.

From Fig. 2 we can identify a number of different periods in the life expectancy evolution. After the steep and erratic initial part, the life expectancy evolution changes character during the early 1950s and improvements are hereafter smoother. Around 1970 improvements in male life expectancy pick up speed and the gender gap begins to narrow. Finally, in the last part of the series, around 2010, improvements slow down and life expectancy flattens for both sexes.

The choice of data period for the analysis is a compromise between including as much information as possible versus using only data adequately described by the models. Balancing these concerns we choose to use the period 1960–2016; a period so long that it enables us to capture potential equilibrium relations. As previously advertised we use the Poisson regression version of the Lee–Carter model with the identification constraints of (2.3). The Lee–Carter model is estimated separately for males and females over the period 1960–2016 and ages 0–100. Fig. 3 shows the estimated mortality indices (\( k \)-index) and the age response parameters (\( \beta \)-parameters). The mortality indices evolve quite similarly over time, while the age response parameters are rather different between ages 20–80.

\(^7\) Specifically, we use data for United Kingdom with HMD country code GBR_NP.
5.1.1. Cointegration rank and deterministic structure

From unit root tests we conclude that \( k^c_t \) and \( k^g_t \) are indeed (1)-processes (test results not shown). The next objective of the analysis is to determine whether or not the two indices engage in an equilibrium correcting relationship, i.e. do \( k^c_t \) and \( k^g_t \) cointegrate?

To answer this question we employ the VECM of (4.2) and test for cointegration rank and deterministic structure in this model. For given deterministic structure, the test for cointegration rank is based on Johansen’s trace statistic, cf. (A.15) of Appendix. The results are shown in Table 3. Recall that the null hypothesis of at most \( r \) cointegrated relations, \( \text{rank}(\Pi) \leq r \), is tested against an unrestricted \( \Pi \). Since the test for cointegration rank is very dependent on the assumed deterministic structure, the two must be determined jointly.

We first note that the two models \( H_t(0) \) and \( H_t^*(0) \) are identical, since \( \text{rank}(\Pi) = 0 \) implies that \( \alpha \) is zero. Table 3 presents two tests for this model against, respectively, \( H_t(2) \) and \( H_t^*(2) \). Both of these are rejected at the 5%-significance level. Consequently, we conclude that the system cointegrates and we move on to test the two cointegration models \( H_t(1) \) and \( H_t^*(1) \). Both of these models are accepted and we therefore take the simpler of these, \( H_t(1) \), as describing our data

\[
\Delta k_t = \alpha \beta^* k_{t-1} + \theta_0 + \varepsilon_t = \alpha (\beta^* k_{t-1} - \mathbb{E}[\beta^* k]) + \tau_1 + \varepsilon_t, \tag{5.1}
\]

where \( \mathbb{E}[\beta^* k] = - (\beta^*)^{-1} \beta^* \theta_0 \) and \( \tau_1 = \beta^* (\alpha^* \beta^*)^{-1} \alpha^* \theta_0 \) are, respectively, the stationary mean of the cointegration relation and the trend in the level of the process, cf. (4.3).

In an ordinary cointegration analysis model (5.1) would serve as a starting point for formulating and testing hypotheses of interest on the parameters. However, as described in Section 4.2, in the present context there are no further testable hypotheses and all there is left to interpret is the model.

5.1.2. Interpretation of the cointegration model

We obtain the following model, where we have used the just-
identified normalization of Section 3.1 by which the first element of \( \beta \) is 1.

\[
\frac{\Delta k^c_t}{\Delta k^g_t} = \begin{pmatrix} -0.084 \\ 0.018 \end{pmatrix} \begin{pmatrix} 1 & -1.235 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k^c_{t-1} \\ k^g_{t-1} \end{pmatrix} + \begin{pmatrix} -1.560 \\ -1.526 \end{pmatrix} + \varepsilon_t
= \begin{pmatrix} -0.084 \\ 0.018 \end{pmatrix} (s_{t-1} - 3.056) + \begin{pmatrix} -1.817 \\ -1.471 \end{pmatrix} + \varepsilon_t,
\]

where

\[
\Delta k^c_t = \hat{\beta}^* k^c_t = k^c_t - 1.235 k^g_t \quad \text{and} \quad \text{drift term} \quad \theta_0 \quad \text{is decomposed in (5.1)}.
\]

Parameter estimates are obtained by reduced rank regression as described in Appendix.

Note that the adjustment coefficients have opposite signs for males and females such that the two indices are either pushed together or pushed apart in response to disequilibrium errors. However, since the female coefficient is much smaller than the male coefficient, the adjustments are primarily taken by the male index, while the female index evolves essentially like a random walk.

It is tempting to interpret the fact that the \( \beta \)-coefficients are of the same magnitude but opposite signs as suggesting that male and female mortality “follow each other closely”, i.e. as approximate coherence. However, this is not a valid conclusion since the (relative) magnitude of the \( \beta \)-coefficient has no significance by itself, but is merely a result of the chosen identification schemes, cf. Section 4.2. The proper conclusion is that forecasted mortalities will be approximately coherent for ages with \( b^c_t \approx 1.25 b^g_t \), and for these ages only. From the right panel of Fig. 3 we see that this relation is satisfied for only a small group of ages around age 25 and age 40.

Regarding the cointegrating relation, it follows from (5.1) that

\[
s_t = (1 + \beta^* \alpha) (s_{t-1} - \mathbb{E}[\beta^* k]) + \mathbb{E}[\beta^* k] + u_t, \tag{5.2}
\]

where \( u_t = \beta^* \varepsilon_t \). Thus the cointegrating relation follows an autoregression with AR-coefficient of \( 1 + \beta^* \alpha \).

We know from Section 4.2 that alternative identification schemes in the underlying Lee–Carter models lead to a new set of adjustment and long-run coefficients of the form \( \tilde{\alpha} = D \alpha \) and \( \tilde{\beta} = D^{-1} \beta \). Observing that \( \beta^* \alpha = \beta \alpha \), it follows that the AR-coefficient is in fact invariant to the identification scheme(s). We can therefore
conclude that the cointegrating relation is always mean-reverting with an estimated AR-coefficient of $1 + \hat{\beta} \hat{\alpha} = 0.89$, implying a strong degree of mean reversion. Due to non-identifiability, no further inference can be drawn about the nature of the joint behaviour under the cointegrated Lee–Carter model.

### 5.2. Cointegration rank and deterministic structure

In contrast to the Lee–Carter model, the CBD model uses two, fully identified factors to describe the mortality evolution of the population under study. In the following we use the CBD model of Section 2.2 as the basis for a four-dimensional cointegration analyses of UK male and female mortality. The higher dimension opens for a richer, yet still interpretable, set of relations while factor identifiability enables the formulation of testable hypotheses.

The CBD model is intended for modelling of pensioners’ mortality only and, consequently, we apply it to ages 60–100, rather than the full age span. Fig. 4 shows the two mortality factors of the CBD model estimated separately for UK males and females over the period 1960–2016. The first factor represents the level of mortality and the second factor represents the slope of the mortality curve. Both factors show a clear trend over time. Not surprisingly, the level is generally declining reflecting an overall decrease in mortality over the past decades for both genders. We note that the profile of the level factor bears some resemblance to that of the mortality index of the Lee–Carter model displayed in Fig. 3. However, in contrast to the mortality index of the Lee–Carter model the level factor of the CBD model represents the absolute level of mortality and it is therefore markedly higher for males than for females. The slope parameter, shown in the right panel of Fig. 4, is generally upward trending indicating greater mortality improvements at younger ages than at older ages. While we observe a somewhat stable difference between male and female levels of mortality, there appears to be no obvious relation between the slope parameters for the two genders.

### 5.2.1. Cointegration rank and deterministic structure

Proceeding as in Section 5.1.1, we first verify that the components of $\hat{k}_t = (\hat{k}_{t,1}, \hat{k}_{t,2}, \hat{k}_{t,3}, \hat{k}_{t,4})$ are $I(1)$-processes (test results not shown).

The next step of the analysis is to determine a suitable VECM for the composite index in which we can subsequently formulate and test specific hypotheses. We employ again the VECM of (4.2) and test jointly for cointegration rank and deterministic structure in this model. Table 4 shows the results of Johansen’s trace test for $\text{rank}(\Pi) \leq r$ against an unrestricted $\Pi$ for the two deterministic structures of relevance, $H_1$ and $H_0^r$. Fig. 5 shows the relation between the various models; note that $H_1(0) = H_0^0(0)$ corresponds to a (multivariate) random walk with drift customarily used for forecasting in the CBD model, cf. Section 2.2.1.

It can be seen from Table 4 that the random walk hypothesis ($\text{rank}(\Pi) = 0$) is rejected in both $H_1(4)$ and $H_0^0(4)$, and we conclude a cointegration rank of at least one. The smallest model of rank 1, $H_1(1)$, is also rejected, while $H_1(2)$ and $H_1(3)$ are both accepted. These two models represent the two smallest acceptable models, cf. Fig. 5. Neither of the models is contained in the other, but $H_1^1(1)$ is arguably the “simpler” model introducing only a single additional trend term while $H_1(2)$ introduces an additional cointegrating relation. Consequently, we adopt $H_1^1(1)$ as our starting model, i.e. the VECM

$$\Delta k_t = \alpha (\beta k_{t-1} + \rho_1 t) + \theta_0 + \varepsilon_t,$$

(5.3)

where $\alpha$ and $\beta$ are four-dimensional vectors.

### 5.2.2. Hypothesis testing and interpretation

Having established that the system has a cointegration rank of one, we can now investigate hypotheses on the nature of the equilibrium correcting relation. We are primarily interested in investigating whether the mortality levels of the two genders enter the cointegrating relation with coefficients of the same magnitude and opposite signs, i.e. whether the distance between
the level parameters is the quantity entering the stable relation. In addition, we are also interested in investigating the degree of dependence between the level and slope parameters which can be formulated as restrictions on the adjustment coefficients. Other hypotheses of interest could be formulated, but we restrict ourselves to these two.

We first formulate the hypothesis of main interest, namely that the first two components of \( \beta \) are of the same magnitude and opposite signs. The last two components of \( \beta \) are left unrestricted. At the same time we would like to test for the absence of a linear drift in the cointegrating relation \( (\rho_1 = 0) \), since this term muddles the interpretation of the stationary relation. Hence, we consider the composite hypothesis

\[
H_0 : (\beta', \rho_1) = (1, -1, \varphi_1, \varphi_2, 0),
\]

where \( \varphi_1, \varphi_2 \in \mathbb{R} \) are unrestricted parameters. The test statistic for this hypothesis is \( \chi^2(2) \)-distributed with a value of 0.005 and a critical value of 5.99 at a 5%-significance level. Hence, we accept hypothesis (5.4) and obtain

\[
\Delta k_T^2 = \begin{pmatrix} -0.105 & 1 \\ 0.002 & 21.66 \\ 0.002 & -12.44 \end{pmatrix} 
\]

and

\[
H_0 : (\alpha', \beta', \rho_1) = (1, -1, \varphi_1, \varphi_2, 0),
\]

where \( \varphi_1, \varphi_2 \in \mathbb{R} \) are unrestricted parameters. The test statistic for this hypothesis is \( \chi^2(5) \)-distributed with a value of 9.54 and a critical value of 11.07 at a 5%-significance level. We conclude that both slope parameters are weak exogenous and are therefore not impacted by the long-run relation. We obtain

\[
\Delta k_T^2 = \begin{pmatrix} 0.089 & 0 \\ 17.78 & 0 \\ 22.12 & 0 \end{pmatrix} 
\]

and

\[
\sum^T_k \alpha_k \sim \mathcal{N}(C, \Sigma),
\]

where \( C = \widehat{k}_{T+h} \) is the last value of the estimated indices.

Eq. (5.9) is a general result valid for all VAR(1)-models with constant drift. This is useful for numerical computations, but the mean and variance structures are not easily discerned. Using the cointegrating relations it is possible to obtain more revealing expressions for the model at hand.

First, let \( B = \alpha (\beta' \alpha)^{-1} \beta' \) and \( C = \alpha (\beta' \alpha)^{-1} \alpha' \), and notice that \( I = C + B \), cf. Chapter 3 of Johansen (1995). Next, observe that \( II_1 \) has eigenvalues 1 and \( \lambda = 1 + \beta' \alpha \) with corresponding eigenspaces \( \text{span}(\beta'_s) \) and \( \text{span}(\alpha) \), respectively. Hence, for \( \nu \in \mathbb{R}^k \) and \( i \geq 0 \),

\[
\Pi^i_1 \nu = \Pi^i_1 (C + B) \nu = II^i_1 (Cv + Bv) = Cv + \lambda^i Bv,
\]

since \( Cv \in \text{span}(\beta'_s) \) and \( Bv \in \text{span}(\alpha) \). In words, \( II_1 \) acts on \( \nu \) by leaving intact its \( \beta'_s \)-component, while shrinking its \( \alpha \)-component by a factor of \( \lambda \).<sup>10</sup>

Using (5.10) and the formula for the \( h \) first terms of a geometric series \( \sum_{i=0}^{h-1} \lambda^i = (1 - \lambda^h)/(1 - \lambda) \), we find the following...
expression for the mean of the forecasting distribution

\[
\Pi_T \varphi_k + \sum_{i=0}^{h-1} \Pi_T \theta_0 = C k_T + \alpha (\beta' \alpha)^{-1} \sum_{i=0}^{T} (I + \alpha \beta') \theta_0 = C k_T + \theta_0 h - \alpha (\beta' \alpha)^{-1} (\beta \alpha)^{-1} \beta \theta_0 = r_0 + \tau_1 h, \text{i.e. (5.11) without the exponentially decaying error-correction term.}
\]

For the variance of the forecasting distribution, it can be shown that

\[
\lim_{h \to \infty} \frac{1}{h} \sum_{i=0}^{h-1} \Pi_T \Sigma \Pi_T' = C \Sigma C',
\]

see pp. 68–69 of Johansen (1995). Thus, asymptotically the variance of the process accumulates linearly, with the variance from the random walk component on \(\text{span}(\beta_k)\) dominating the total variability.

To give a more intuitive understanding of the role of cointegration, we conclude with a numerical example where we compare the (joint) forecast from the cointegration model (5.7) with separate, gender-specific forecasts based on the bivariate random walk model of Section 2.2.1. We compare the forecasted cohort remaining life expectancy as it depends on the projected mortality surface over a long horizon and it is therefore well suited to capture differences in dependency structures over time.

The two models describe and project the estimated CBD-distributions for the difference in life expectancy between men and women under the cointegration model than under the random walk model, as illustrated in the right panel of Fig. 6.

In summary, the cointegration model yields forecasts that are similar to those obtained from the simpler random walk model for each sex, but with a more plausible dependency structure. The cross-gender dependency is achieved by a single cointegrating relation derived from data. The resulting forecasts are well-behaved and empirically justified, but they are not coherent (in the strict mathematical sense). This indicates that the current definition of coherence might be too strict and too specific a requirement, and that other types of “coherence” might be equally good – or even better – when judging the quality of joint forecasts.

6. Concluding remarks

In this paper we have discussed the interlinked concepts of identifiability, coherence and cointegration in the context of multi-population mortality modelling. We have made the point that cointegration has an important role to play as an inferential tool to obtain insights into the joint dynamics of mortality factors. This role goes beyond the typical usage of cointegration as merely a tool to obtain coherent forecasts—defined as forecasts for which the relative age-specific mortality rates converge over time.\(^{13}\)

Since its introduction in Li and Lee (2005), the concept of coherence has served as the gold standard for joint forecasts of related populations, and many models have been designed with the explicit goal of achieving coherence. At first sight, coherence seems like a reasonable property, but on further inspection it

\(^{12}\) When \(k_t\) has the form (5.8) Theorem 3.3 reads (in terms of expected value)

\[
\mathbb{E}[k_{T+h}] = C k_T + \sum_{i=0}^{h-1} \theta_0 + \alpha (\beta' \alpha)^{-1} \sum_{i=0}^{T} (I + \alpha \beta') \beta \theta_0 = C k_T + \theta_0 h - \alpha (\beta' \alpha)^{-1} (\beta \alpha)^{-1} \beta \theta_0 = r_0 + \tau_1 h, \text{i.e. (5.11) without the exponentially decaying error-correction term.}
\]

\(^{13}\) Whether the goal of (statistical) models is to gain insights or to forecast is the object of a long-standing, and still highly relevant, debate, see e.g. Breiman (2001) and Shmueli (2010).
appears somewhat arbitrary and specifically tailored to models of the Lee–Carter type with log-linear modelling of mortality rates. Joint forecasts based on other model types, e.g. logistic, can produce equally plausible dependency structures and thus be equally “coherent”, even if they lack relative convergence. In our view, insisting on coherence is not a suitable starting point, nor a reasonable restriction, for a joint analysis and might in fact produce forecasts that are at odds with historic data.

In this paper we have focused on two types of linear trend models suitable for analysing the period effect of age–period mortality models, and we have shown how to interpret these models by use of the Granger decomposition. Cointegration is a technical field, and the analysis and interpretation of models and hypotheses are not straightforward. In many mortality models, e.g. Lee–Carter type models, the factors(s) are only semi-identifiable in which case additional difficulties (and pitfalls) arise. From a statistical point of view, it is in general meaningless to test, or impose restrictions, on the long-run coefficients in the context of semi-identifiable models. In contrast, fully identified models give access to the full inferential power of cointegration and in our opinion this is a strong argument in favour of these models if the aim is to gain subject matter insights.

Declaration of competing interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements
The authors wish to thank two anonymous referees for valuable input for improving the presentation.

Appendix. Maximum likelihood estimation of the VECM
Consider the VAR(\(k\)) model in VECM form with \(\Pi = \alpha \beta^t\) for \(p \times r\) matrices \(\alpha\) and \(\beta\).

\[
\Delta y_t = \sum_{k=1}^{p-1} \Gamma_k \Delta y_{t-k} + \alpha \beta^t y_{t-1} + \Phi D_t + e_t, \tag{A.1}
\]

where \(e_t \sim N_p(0, \Sigma_e)\) and independent. Based on data for \(t = 1, \ldots, T\), maximum likelihood estimates of the freely varying parameters \((\alpha, \beta, \Gamma_1, \ldots, \Gamma_{r-1}, \Phi, \Sigma_e)\) can be obtained by reduced rank regression. Following Joensuu (2006), we introduce the shorthand notation

\[
Z_{\Delta y} = \Delta y_t, \tag{A.2}
\]

\[
Z_{\Delta y t} = y_{t-1}, \tag{A.3}
\]

\[
Z_{\Delta y t} = [\Delta y_{t-1}, \ldots, \Delta y_{t-k+1}, \Phi], \tag{A.4}
\]

and write (A.1) on the form

\[
Z_{\Delta y} = \alpha \beta Z_{\Delta y t} + \Psi Z_{\Delta y t} + e_t, \tag{A.5}
\]

where \(\Psi = [\Gamma_1, \ldots, \Gamma_{k-1}, \Phi]\). Define the product moment matrices

\[
M_{ij} = T^{-1} \sum_{t=1}^{T} Z_{\Delta y_t} Z_{\Delta y_t}^t, \quad i, j = 0, 1, 2, \tag{A.6}
\]

and the sample-covariance matrices

\[
S_{ij} = M_{ij} - M_{ii}M_{jj}^{-1}M_{jj}, \quad i, j = 0, 1. \tag{A.7}
\]

The maximum likelihood estimator of \(\beta\) is found by solving the eigenvalue problem

\[
\lambda_{S_{11}} - S_{10}S_{00}^{-1}S_{01} = 0, \tag{A.8}
\]

for eigenvalues \(1 > \hat{\lambda}_1 > \cdots > \hat{\lambda}_p > 0\) with associated eigenvectors \(\hat{v}_1, \ldots, \hat{v}_p\) that satisfy

\[
\hat{\lambda}_i S_{11} \hat{v}_i = S_{10}S_{00}^{-1}S_{01}, \quad i = 1, \ldots, p, \tag{A.9}
\]

\[
\hat{\Psi} = I, \tag{A.10}
\]

where \(\hat{\Psi} = (\hat{v}_1, \ldots, \hat{v}_p)\). The cointegrating relations \(\hat{\beta}\) are given by the first \(r\) eigenvectors

\[
\hat{\beta} = (\hat{v}_1, \ldots, \hat{v}_r). \tag{A.11}
\]

The remaining parameters are estimated as

\[
\hat{a}(\beta) = S_{01} \beta (\beta S_{11} \beta)^{-1}, \tag{A.12}
\]

\[
\hat{\Sigma}(\alpha, \beta) = S_{00} - \alpha (\beta S_{11} \beta) \alpha^t, \tag{A.13}
\]

\[
\hat{\phi}(\alpha, \beta) = M_{02}M_{22}^{-1} - \alpha \beta M_{12}M_{22}^{-1}, \tag{A.14}
\]

with \(\beta = \beta\) and \(\alpha = \hat{a}(\beta)\).

Trace statistic for test of cointegration rank
Recall that \(H(r)\) denotes the hypothesis that \(\text{rank}(\Pi) \leq r\), or equivalently that \(\Pi\) has at most \(r\) non-zero eigenvalues. The likelihood ratio test statistic for \(H(r)\) in \((H(p))\), known as the trace statistic, is given by

\[
LR_{\text{trace}}(r) = -T \sum_{i=r+1}^{p} \log(1 - \hat{\lambda}_i), \tag{A.15}
\]

where \(\hat{\lambda}_i\) are the eigenvalues found by solving (A.8). The asymptotic distribution of the trace statistic depends on the deterministic terms. Critical values are tabulated in Section 15.3 of Johansen (1995) for the linear models in Section 3.2. Intuitively, if \(\text{rank}(\Pi) > r\) then \(LR_{\text{trace}}(r)\) will be small, since \(\hat{\lambda}_{r+1}, \ldots, \hat{\lambda}_p\) will all be close to zero. Conversely, the test statistic will be large if \(\text{rank}(\Pi) > r\), since at least \(\hat{\lambda}_{r+1}\) will deviate from zero.

References


