A Step-by-Step Guide to Building Two-Population Stochastic Mortality Models

Johnny Siu-Hang Li, Rui Zhou and Mary Hardy

Abstract

Two-population stochastic mortality models play an crucial role in the securitization of longevity risk. In particular, they allow us to quantify the population basis risk in longevity hedges that are built from broad-based mortality indexes. In this paper, we propose a systematic process for constructing a two-population mortality model for a pair of populations. The process encompasses four steps, namely (1) determining the conditions for biological reasonableness, (2) identifying an appropriate base model specification, (3) choosing a suitable time-series process and correlation structure for projecting period and/or cohort effects into the future, and (4) model evaluation. In our discussion, we not only consider existing models, but also introduce several significant new models, including a two-population version of the celebrated Cairns-Blake-Dowd model. We demonstrate the proposed model-building process using real data from two pairs of populations.

Keywords: Index-based longevity hedges; Population basis risk; Coherent mortality forecasting

1 Introduction

Rapid, unexpected increases in life expectancy pose a serious challenge to the financial health of pension plan sponsors, potentially affecting their ability to compete in their core business. The impact is particularly severe in today’s low yield environment, where asset returns cannot defray the costs associated with increasing life expectancy unless substantially more investment risk is taken. In recent years, there has been a growing awareness of longevity risk among pension plan sponsors, many of which are contemplating risk transfer solutions to offload the longevity risk from their balance sheets.

The capital markets offer a possible channel for longevity risk transfers. Through a portfolio of mortality-linked securities, longevity risk can be transferred from a pension plan to one or more counterparties who are interested in acquiring an exposure to the risk, in return for a risk premium and/or the diversification benefit it offers. Capital market solutions for hedging longevity risk can be broadly divided into two categories: customized and index-based. A customized security is
one that is linked to the mortality experience of the hedger's own population of individuals. One recent example is the £3.2 billion longevity swap agreed between Legal & General and the pension plan of BAE Systems in 2013. The net payments from this swap are determined by the mortality experience of the plan's 31,000 current pensioners. An index-based security, by contrast, is one that is linked to an index reflecting the realized mortality experience of a larger pool of lives, such as a national population. A well-known example is the portfolio of q-forwards that J.P. Morgan transacted with Lucida PLC in 2008. The payout from it is linked to an index that is based on England and Wales (EW) national male mortality over a range of ages.

Index-based securities are of greater appeal to investors and intermediaries, because they are more transparent, easier to analyze and potentially more liquid. Nevertheless, in using an index-based security, a hedger is exposed to population basis risk, which arises from the difference in mortality improvements between its own population and the population to which the security is linked. Hedgers' concerns about population basis risk, which is still partially understood to date, make index-based securities relatively unpopular, thereby hindering the development of a large and liquid market for longevity risk transfers. To overcome this obstacle, we believe that there is a need to further develop two-population stochastic mortality models, which permit market participants to assess with confidence the amount of population basis risk involved in an index-based longevity hedge. This view is also shared by the Life and Longevity Markets Association (LLMA), which has recently established the Longevity Basis Risk Working Group to search for a readily applicable method for quantifying population basis risk.

Two-population stochastic mortality models also facilitate more sophisticated longevity risk transfers that are designed to suit the specific needs of hedgers and/or investors. In 2010, Swiss Re, which has reinsured annuity business in Britain and life business in the US, issued a longevity trend bond (known as the Kortis deal) to mitigate the risk that it is facing. The bond would trigger in the event that there is a large divergence in the mortality improvements experienced between male lives aged 75 to 85 in EW and male lives aged 55 to 65 in the US. Two-population mortality models, which explicitly capture the random deviations between two mortality trends, are by design highly suitable for pricing and analyzing securities like the Kortis deal.

A few researchers have contributed to the subject of two-population stochastic mortality modeling. Li and Lee (2005) developed two extensions of the Lee-Carter model for modeling multiple populations. The two extensions were subsequently used by Li and Hardy (2011) to quantify the population basis risk in an index-based q-forward longevity hedge. Cairns et al. (2011) discussed the core hypotheses and desirable criteria for use in constructing two-population mortality models. Dowd et al. (2011) proposed a two-population gravity model, in which autoregressive processes are used to reduce any increasing spread between each pair of stochastic factors in the model back to a constant level. The gravity model has a close resemblance to the model proposed by Jarner and Kryger (2011), which is designed for modeling two populations of highly different sizes. Yang and Wang (2013) and Zhou et al. (2014) investigated the use of different multivariate stochastic processes for projecting the stochastic factors in a multi-population mortality model. Zhou
et al. (2013) introduced a two-population Lee-Carter model with transitory jump effects. The mentioned models are meticulously constructed so that they will not lead to diverging long-term forecasts, which do not seem to be biologically reasonable.

Most of the existing two-population mortality models are built on either the Lee-Carter (LC) or the Age-Period-Cohort (APC) structure. Nonetheless, these two model structures, which were developed many years ago, do not always provide the best fitting and forecasting results. In an analysis based on data from the US and EW populations, Cairns et al. (2009) found that Model M7 (the Cairns-Blake-Dowd model with cohort and quadratic effects) performs better than the LC and APC structures in terms of the Bayesian Information Criterion and the robustness of parameter estimates. Similar studies conducted by Cairns et al. (2011) and Dowd et al. (2010a,b) also point to the conclusion that the LC and APC structures are not always the most preferred. In addition, the LC and APC structures do not possess such desirable properties as the “new-data-invariant” property (Chan et al., 2014), which is found in the original Cairns-Blake-Dowd model.\(^1\)

Therefore, a natural question to ask is whether we can leverage the recent success in single-population mortality modeling to the development of two-population mortality models. In this paper, we address this question by studying how other prevalent stochastic mortality models, including the Cairns-Blake-Dowd model and its variants, can be generalized to their two-population counterparts. We do not aim to create a specific new two-population mortality model, because the best model structure is likely to be data dependent. Instead, we set out a step-by-step procedure, in which the most appropriate two-population mortality model for a certain data set can be constructed from a toolkit of parameter structures and stochastic processes. All required likelihood functions that are not available in the existing literature are derived. Our idea is in some sense similar to that of Hunt and Blake (2014), who introduced a general procedure for constructing single-population mortality models using a combination of parametric functions and expert judgement.

The difficulty of our work lies in the fact that some of the candidate model structures contain multiple stochastic factors. Hence, if they are expanded to a two-population setting directly, the resulting models would be highly complex, involving a large number of parameters. The structural complexity often implies a lack of robustness and also a slow convergence rate in the estimation process. To alleviate this problem, we implement the “top-down” approach, which was proposed by Cairns (2013) as a concept, to enforce the principle of parsimony. In using this approach, we first fit the full model with all stochastic factors. Then the full model is systematically simplified by, for example, merging certain pairs of stochastic factors and eliminating some parameters in the time-series processes for the stochastic factors. The simplification ends when we arrive at the most parsimonious model specification that gives an adequate fit to the historical data.

\(^1\)The new-data-invariant property refers to the property that when an additional year of mortality data becomes available and the model is updated accordingly, the previously estimated parameters are not affected.
The remainder of this paper is organized as follows. Section 2 describes the data sets that we use in our illustrations. Section 3 provides an overview of the proposed step-by-step procedure for constructing two-population models. Sections 4 to 7 respectively detail each step in the model-building procedure. Finally, Section 8 concludes the paper.

2 Data

We illustrate the proposed model-building procedure with historical mortality data from two pairs of populations, namely (1) the populations of EW males and UK male insured lives, and (2) the populations of EW males and Dutch males. In the former pair, the second population is (approximately) a sub-population of the first population. In the latter pair, the physical locations of the two populations are different.

The data (death and exposure counts) for EW and Dutch populations are obtained from the Human Mortality Database (2012), while that for the UK insured lives are obtained from the Continuous Mortality Investigation (CMI) Bureau of the Institute and Faculty of Actuaries. For both pairs of populations, we use a sample period of 1947 to 2005 and a sample age range of 60 to 84.

3 An Overview of the Model-Building Procedure

Our goal is to set out a framework to generalize the seven stochastic mortality models documented by Cairns et al. (2009) to their two-population versions. To construct a two-population mortality model, we use two identical single-population models, one for each population. The seven candidate model structures are shown in Table 1. In describing the model structures, the following conventions are used:

- $m_{x,t}^{(j)}$ is the population $j$’s central death rate at age $x$ in year $t$;
- $q_{x,t}^{(j)}$ is the probability that an individual in population $j$ aged exactly $x$ at exact time $t$ will die between $t$ and $t + 1$;
- $\beta_i^{(i,j)}$, $i = 1, 2, 3$, are age-specific parameters for population $j$;
- $\kappa_i^{(i,j)}$, $i = 1, 2, 3$, are time-varying stochastic factors for population $j$;
- $\gamma_{t-x}^{(i,j)}$, $i = 3, 4$, are cohort-related stochastic factors for population $j$;
- $n_a$ is the number of ages covered in the sample age range;

\footnote{The Age-Period-Cohort model has been used widely in medical statistics since a long time ago. The paper by Osmond (1985) is an example.}
Model M1: The Lee-Carter Model (Lee and Carter, 1992)
\[
\ln(m_{x,t}^{(j)}) = \beta_{x}^{(1,j)} + \beta_{x}^{(2,j)} \kappa_{t}^{(2,j)}
\]

Model M2: The Renshaw-Haberman Model (Renshaw and Haberman, 2006)
\[
\ln(m_{x,t}^{(j)}) = \beta_{x}^{(1,j)} + \beta_{x}^{(2,j)} \kappa_{t}^{(2,j)} + \beta^{(3,j)} \gamma_{t-x}
\]

Model M3: The Age-Period-Cohort Model (Osmond, 1985)
\[
\ln(m_{x,t}^{(j)}) = \beta_{x}^{(1,j)} + n_{t} \kappa_{t}^{(2,j)} + n_{t} \gamma_{t-x}
\]

Model M5: The Original Cairns-Blake-Dowd (CBD) Model (Cairns et al., 2006)
\[
\ln \left( \frac{q_{x,t}^{(j)}}{1 - q_{x,t}^{(j)}} \right) = \kappa_{t}^{(1,j)} + \kappa_{t}^{(2,j)} (x - \bar{x})
\]

Model M6: The CBD Model with a Cohort Effect (Cairns et al., 2009)
\[
\ln \left( \frac{q_{x,t}^{(j)}}{1 - q_{x,t}^{(j)}} \right) = \kappa_{t}^{(1,j)} + \kappa_{t}^{(2,j)} (x - \bar{x}) + \gamma_{t-x}
\]

Model M7: The CBD Model with Quadratic and Cohort Effects (Cairns et al., 2009)
\[
\ln \left( \frac{q_{x,t}^{(j)}}{1 - q_{x,t}^{(j)}} \right) = \kappa_{t}^{(1,j)} + \kappa_{t}^{(2,j)} (x - \bar{x}) + \kappa_{t}^{(3,j)} ((x - \bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}
\]

Model M8: The CBD Model with an Age-Dependent Cohort Effect (Cairns et al., 2009)
\[
\ln \left( \frac{q_{x,t}^{(j)}}{1 - q_{x,t}^{(j)}} \right) = \kappa_{t}^{(1,j)} + \kappa_{t}^{(2,j)} (x - \bar{x}) + \kappa_{t}^{(3,j)} (x_{c}^{(j)} - x)
\]

Table 1: The seven candidate model structures under consideration.

- $\bar{x}$ is the mean age over the sample age range;
- $\hat{\sigma}_x^2$ is the mean of $(x - \bar{x})^2$ over the sample age range;
- $x_{c}^{(j)}$ is a constant that does not vary with age or time.

The model-building procedure consists of the following four stages:

**Stage I**
In this stage, we identify, for each possible model structure, the conditions needed to ensure that the resulting two-population model does not lead to forecasts that are biologically unreasonable.

**Stage II**
In this stage, we focus on the estimation of the base model structures, i.e., the equations displayed in Table 1. The “top-down” approach is used to simplify each model structure until an optimal balance between goodness-of-fit and parsimony is achieved. The best-performing three base model structures are shortlisted.
Stage III
In this stage, we focus on the estimation of the time-series processes for the stochastic factors, i.e., parameters $\kappa_{t}^{(i,j)}$, $\gamma_{t-x}^{(i,j)}$, in the three shortlisted base model structures. The “top-down” approach is used again to simplify the time-series processes. At the end of this stage, the optimal time-series processes for each of the shortlisted base model structures are identified.

Stage IV
In this stage, we evaluate the shortlisted models in terms of their robustness and forecasting performance. On the basis of the evaluation results, the most appropriate two-population mortality model is selected.

The four stages in the model-building process are explained and illustrated in the next four sections.

4 Stage I: Conditions for Biological Reasonableness

We cannot arbitrarily use two single-population mortality models to obtain a joint forecast for two related populations, because in this way the forecast may turn out to be biologically unreasonable. For instance, if we use two unrelated single-population models to project the future mortality of EW males and UK male insured lives, then the resulting projection may imply that the mortality of the latter population will exceed that of the former population beyond a certain future time point. This outcome is anti-intuitive, because we expect that the insured lives, who are on average wealthier and healthier than the general population, to have consistently lower mortality.

An important part of two-population mortality modeling is to ensure that the resulting forecasts are coherent, a concept proposed by Li and Lee (2005). In a coherent forecast, the future mortality rates of two related populations are projected jointly in such a way that they do not diverge over the long run. The concept of coherent mortality forecasting was further investigated by Cairns et al. (2011), who translated the qualitative expectation into the following mathematical hypothesis: the ratio $m_{x,t}^{(j_1)}/m_{x,t}^{(j_2)}$ do not diverge as $t \to \infty$, for all age $x$ and $j_1 \neq j_2$. They also demonstrated that for a two-population APC model (Model M3) to satisfy this hypothesis, it is sufficient to assume that both $\kappa_{t}^{(2,1)} - \kappa_{t}^{(2,2)}$ and $\gamma_{t-x}^{(3,1)} - \gamma_{t-x}^{(3,2)}$ are mean-reverting.

In the first stage of the model-building procedure, we identify, for the remaining six model structures, the sufficient conditions for the non-divergence hypothesis to hold. Using the reasoning of Cairns et al. (2011), we require the difference between each pair of stochastic factors to be mean-reverting. Also, if the responses to the stochastic factors are modeled by age-specific parameters, then we require both populations to share exactly the same set of age-specific response parameters. The non-divergence conditions for all seven model structures under consideration are presented in Table 2.

Depending on their nature, the non-divergence conditions shown in Table 2 are enforced in the different stages of the model-building procedure. In particular, the conditions involving stochastic
Sufficient conditions for non-divergence

<table>
<thead>
<tr>
<th>Model M1</th>
<th>$\beta_{x}^{(2,1)} = \beta_{x}^{(2,2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_{t}^{(2,1)} - \kappa_{t}^{(2,2)}$ is mean-reverting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model M2</th>
<th>$\beta_{x}^{(1,1)} = \beta_{x}^{(1,2)}$, $i = 2, 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_{t}^{(2,1)} - \kappa_{t}^{(2,2)}$ is mean-reverting</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{t-x}^{(3,1)} - \gamma_{t-x}^{(3,2)}$ is mean-reverting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model M3</th>
<th>$\kappa_{t}^{(2,1)} - \kappa_{t}^{(2,2)}$ is mean-reverting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{t-x}^{(3,1)} - \gamma_{t-x}^{(3,2)}$ is mean-reverting</td>
</tr>
</tbody>
</table>

| Model M4 | $\kappa_{t}^{(1,1)} - \kappa_{t}^{(1,2)}$, $i = 1, 2$, is mean-reverting |

| Model M5 | $\kappa_{t}^{(1,1)} - \kappa_{t}^{(1,2)}$, $i = 1, 2$, is mean-reverting |

<table>
<thead>
<tr>
<th>Model M6</th>
<th>$\kappa_{t}^{(1,1)} - \kappa_{t}^{(1,2)}$, $i = 1, 2$, is mean-reverting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{t-x}^{(3,1)} - \gamma_{t-x}^{(3,2)}$ is mean-reverting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model M7</th>
<th>$\kappa_{t}^{(1,1)} - \kappa_{t}^{(1,2)}$, $i = 1, 2, 3$, is mean-reverting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{t-x}^{(4,1)} - \gamma_{t-x}^{(4,2)}$ is mean-reverting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model M8</th>
<th>$\kappa_{t}^{(1,1)} - \kappa_{t}^{(1,2)}$, $i = 1, 2$, is mean-reverting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{t-x}^{(4,1)} - \gamma_{t-x}^{(4,2)}$ is mean-reverting</td>
</tr>
<tr>
<td></td>
<td>$x_{1}^{(1)} = x_{1}^{(2)}$</td>
</tr>
</tbody>
</table>

Table 2: The non-divergence conditions for the candidate model structures under consideration.

Factors are implemented in Stage III where we estimate time-series processes for period and/or cohort effects, while the other conditions are applied in Stage II where we estimate the base model structure.

5 Stage II: The Base Model Structure

5.1 Estimating the Base Model Structure

We estimate parameters in the base model structures by the method of maximum likelihood. In general, assuming Poisson death counts, the log-likelihood for the seven candidate model structures can be expressed as

$$l = \sum_{j=1}^{2} \sum_{x=x_{0}}^{x_{1}} \sum_{t=t_{0}}^{t_{1}} w_{x,t}^{(j)} \left[ D_{x,t}^{(j)} \ln \left( E_{x,t}^{(j)} m_{x,t}^{(j)} \right) - E_{x,t}^{(j)} m_{x,t}^{(j)} - \ln \left( D_{x,t}^{(j)} \right) \right],$$

where $[x_{0}, x_{1}]$ is the sample age range, $[t_{0}, t_{1}]$ is the sample period, $D_{x,t}^{(j)}$ is the actual number of deaths at age $x$, observed in population $j$ and in year $t$, and $E_{x,t}^{(j)}$ is the corresponding number of persons at risk. The value of $m_{x,t}^{(j)}$ depends on the model parameters via the equations shown in Table 1. For models that are built for $q_{x,t}^{(j)}$, an assumption is needed to connect $q_{x,t}^{(j)}$ and $m_{x,t}^{(j)}$. We use the assumption that the force of mortality is constant between integer ages, which implies
\[ m_{x,t}^{(j)} = -\ln(1 - q_{x,t}^{(j)}) \]

The weight function \( w_{x,t}^{(j)} \) is used to exclude the first and last four birth cohorts in the data sample.\(^3\)

For Models M1 to M3, the non-divergence conditions involving the age-specific response parameters should be incorporated into the log-likelihood. For instance, in the expression of \( l \) for Model M1, we should set parameters \( \beta_{x}^{(2,1)} \) and \( \beta_{x}^{(2,2)} \) to a common value, say \( \beta_{x}^{(2)} \).

Except Model M5, all model structures under consideration have an identifiability problem, which refers to the situation when there exist multiple equivalent parameter sets. Therefore, on top of the non-divergence constraints, we need to impose identifiability constraints to stipulate parameter uniqueness. The identifiability constraints used in this paper are adapted from those used by Cairns et al. (2009) in a single-population setting. The issue about identifiability constraints is discussed in greater depth in the next sub-section.

Parameter estimates are obtained by maximizing \( l \). The maximization is accomplished by an iterative Newton-Raphson method, in which parameters are updated one at a time. The updating of a typical parameter \( \theta \) proceeds according to

\[
u(\theta) = \theta - \frac{\partial l}{\partial \theta} \frac{\partial^2 l}{\partial \theta^2},\]

where \( u(\theta) \) is the updated value of \( \theta \) in the iteration. The necessary identifiability constraints (if any) are applied at the end of each iteration.

### 5.2 Simplifying the Base Model Structure

A full model structure generally contains a large number of parameters. For instance, in the full two-population version of Model M7, there are altogether four pairs of stochastic factors. To make this model more parsimonious, we may try to combine one or more pairs of stochastic factors and see if the simplified model still provides an adequate fit to the historical data.

We propose to use the “top-down” approach to systematically simplify a model structure, until an optimal balance between goodness-of-fit and parsimony is reached. We determine optimality by using the Bayesian Information Criterion (BIC) (Schwarz, 1978):

\[
BIC = -\hat{l} + 0.5k \ln(n),
\]

which indicates a model’s goodness-of-fit (measured by the maximized log-likelihood \( \hat{l} \)), taken into account the effective number of model parameters \( k \) and the effective sample size \( n \).\(^4\) A model with a lower BIC value is preferred.

The process for simplifying a full model structure is summarized as follows:

---

\(^3\)To exclude the first and last four birth cohorts, we set \( w_{x,t}^{(j)} \) to 0 for \( t - x = t_0 - x_1, t_0 - x_1 + 1, t_0 - x_1 + 2, t_0 - x_1 + 3, t_1 - x_0, t_1 - x_0 - 1, t_1 - x_0 - 2, t_1 - x_0 - 3 \), and to 1 otherwise.

\(^4\)The effective number of model parameters is the total number of parameters less the number of identifiability constraints applied to the base model structure. The effective sample size is the total number of data points \( (t_1 - t_0) \times (x_1 - x_0) \) less the number of data points that are given a zero weight.
1. Fit the full model.

2. Fit simplified models with one pair of merged stochastic factors. For instance, for Model M7, this step involves the fitting of four models with $\kappa_{t}^{(1,1)} = \kappa_{t}^{(1,2)}$, $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)}$, $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ and $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$, respectively. Identify which of these mergers improve the BIC.

3. If none of the simplified models in Step 2 yields a better BIC value, then stop the simplification process and use the full model.

4. If only one simplified model in Step 2 yields a better BIC value, then choose that simplified model and stop the simplification process.

5. Otherwise, fit simplified models with two common stochastic factors (chosen from the identified collection in Step 2). For example, if it was found in Step 2 that the models with $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)}$, $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ and $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$ give better BIC values, then fit the following three models in this step:

   - The model with $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)}$ and $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$
   - The model with $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ and $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$
   - The model with $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$ and $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)}$

6. Determine if there is a need to fit simplified models with three (or more) stochastic factors. This step applies only to Models M6, M7 and M8 in which there are three or four pairs of stochastic factors. According to our experience, in most cases there is no need to combine more than two pairs of stochastic factors.

7. Compare the fitted models, and select the model structure that gives the best BIC value.

There are two technical points to note in implementing the simplification process. First, to fit models with merged stochastic factors, the log-likelihood in equation (5.1) should be adapted accordingly. For instance, when fitting Model M1 with a merged period effect, we should set not only $\beta_{t}^{(2,1)} = \beta_{t}^{(2,2)} = \beta_{t}^{(2)}$ (the non-divergence condition) but also $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)} = \kappa_{t}^{(2)}$ (the merged period effect) in the expression for $l$. Second, when one or more pairs of stochastic factors are merged, the identifiability constraints have to be adapted accordingly. The necessary identifiability constraints for each possible model specification are provided in Appendix A.

The simplification process above allows us to fine tune each of the seven model structures presented in Table 1. At the end of Stage II, the three fine-tuned models with the best BIC values are shortlisted.
Table 3: The BIC values of all base model structures estimated in implementing the “top-down” approach. Population \( j = 1 \) is EW males and population \( j = 2 \) is UK male insured lives. The lowest three BIC values are displayed in boldface.

<table>
<thead>
<tr>
<th>Model</th>
<th>Simplification(s)</th>
<th>BIC</th>
<th>Model</th>
<th>Simplification(s)</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>None</td>
<td>19,694</td>
<td>M2</td>
<td>None</td>
<td>16,839</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} )</td>
<td>23,218</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} )</td>
<td>17,958</td>
</tr>
<tr>
<td>M2</td>
<td>None</td>
<td>16,839</td>
<td>M5</td>
<td>None</td>
<td>20,311</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>16,807</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>20,074</td>
</tr>
<tr>
<td></td>
<td>( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>16,798</td>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,784</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,784</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,784</td>
</tr>
<tr>
<td>M3</td>
<td>None</td>
<td>17,583</td>
<td>M6</td>
<td>None</td>
<td>17,739</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,410</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,767</td>
</tr>
<tr>
<td></td>
<td>( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,326</td>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>18,353</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>18,353</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,713</td>
</tr>
<tr>
<td>M5</td>
<td>None</td>
<td>20,311</td>
<td>M7</td>
<td>None</td>
<td>17,603</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>67,074</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,603</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} )</td>
<td>21,118</td>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,532</td>
</tr>
<tr>
<td>M6</td>
<td>None</td>
<td>17,716</td>
<td>M8</td>
<td>None</td>
<td>17,445</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,464</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,747</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,464</td>
<td></td>
<td>( \kappa_t^{(3,1)} = \kappa_t^{(3,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,445</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(3,1)} = \kappa_t^{(3,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,439</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,445</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,510</td>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>67,301</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>62,711</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,903</td>
</tr>
<tr>
<td></td>
<td>( \kappa_t^{(2,1)} = \kappa_t^{(2,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>17,892</td>
<td></td>
<td>( \kappa_t^{(1,1)} = \kappa_t^{(1,2)} ), ( \gamma_{t-x} = \gamma_{t-x} )</td>
<td>44,187</td>
</tr>
</tbody>
</table>

5.3 Illustrations

5.3.1 EW Males and UK Male Insured Lives

In this illustration, population \( j = 1 \) refers to EW males and population \( j = 2 \) refers to UK male insured lives. The simplification process is applied to all seven model structures in Table 1. The BIC values for all models estimated during the simplification process are displayed in Table 3. To further explain the simplification process, we consider Models M2 and M7 as examples.

For Model M2, we find that only the merger of the cohort effects leads to a better BIC value. The simplification process therefore stops at Step 2, and the fine-tuned Model M2 is the one with \( \gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)} \). Figures 1 and 2 depict the parameter estimates of the full Model M2 and the simplified Model M2 with \( \gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)} \), respectively. It is noteworthy that as we combine the cohort effects, the estimates of the period effects (i.e., parameter \( \kappa_t^{(2,2)} \)'s) become significantly different. Specifically, instead of fluctuating around zero, the estimates of the period effects exhibit
Figure 1: Parameter estimates of the full two-population version of Model M2. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

a downward trend, which is more demographically reasonable. This is possibly because the merger of cohort effects allows the estimation to be based on information from both populations, thereby helping the estimation process to better distinguish between cohort and period effects, which are often difficult to separate because they are both related to the passage of time.

For Model M7, we find that all variants fitted in Step 2 produce better BIC values in comparison to the full model. We therefore proceed to Step 3, where six simplified models, each of which contains two combined stochastic factors, are fitted. Among these six models, the following three outperform those in the previous step:

1. the variant with $\gamma_{(4,1)}^{t-x} = \gamma_{(4,2)}^{t-x}$ and $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ ($BIC = 17,491$);
2. the variant with $\kappa_{t}^{(2,1)} = \kappa_{t}^{(2,2)}$ and $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ ($BIC = 17,517$);
3. the variant with $\kappa_{t}^{(1,1)} = \kappa_{t}^{(1,2)}$ and $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$ ($BIC = 17,514$).

We then determine if there is a need to further simplify these three models. To illustrate, let us consider the variant with $\gamma_{(4,1)}^{t-x} = \gamma_{(4,2)}^{t-x}$ and $\kappa_{t}^{(3,1)} = \kappa_{t}^{(3,2)}$. For this variant, the only possible further simplifications are to merge the first and/or second pairs of period effects. However, these further simplifications will not yield better BIC values for sure, because we know from Table 3 that if the cohort effects are merged, then merging the first or second pair of period effects lead to BIC values (62,612 and 18,080, respectively) that are even worse than those obtained in Step 2. Using the same reasoning, we find that there is no need to further simplify the remaining two
Figure 2: Parameter estimates of the two-population version of Model M2 with a merged cohort effect. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

variants. The simplification process terminates here, and the fine-tuned Model M7 is the one with $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$ and $\kappa_t^{(3,1)} = \kappa_t^{(3,2)}$, which gives the best BIC value among all variants we fitted.

Figures 3 and 4 show the parameter estimates of the full and fine-tuned versions Model M7, respectively. As some of the stochastic factors are combined, the trends in the parameter estimates take more reasonable shapes. In the fine-tuned version, the trends in the first and second period effects are steadier, and the pattern of the cohort effects is more in line with that obtained by other researchers, including Cairns et al. (2009) who found that cohort mortality was falling at a faster rate for males born after 1920.

The “top-down” approach allows us to fine tune each of the seven model structures in Table 1. Finally, we shortlist the three fine-tuned model structures with the best BIC values, namely (1) Model M2 with $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$ ($BIC = 16,798$), (2) Model M3 with $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$ ($BIC = 17,326$) and (3) Model M6 with $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$ ($BIC = 17,439$). The stochastic factors contained in the three shortlisted models will be modeled in the next stage of the model-building process.

5.3.2 EW Males and Dutch Males

In this illustration, population $j = 1$ refers to EW males and population $j = 2$ refers to Dutch males. The BIC values for all models estimated during the simplification process are displayed in Table 4.

For this pair of populations, the three shortlisted model structures are (1) the full version
Figure 3: Parameter estimates of the full two-population version of Model M7. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

It is noteworthy that for this pair of populations, the candidate model structures have less room for simplification. This might be attributed to the fact that in comparison to the population of UK insured lives, the Dutch population is less related to the EW population.

6 Stage III: The Stochastic Factors

In this stage, we model the stochastic factors encompassed in the model structures shortlisted in Stage II. We first discuss how stochastic factors related to period effects are modeled, and then we move on to the stochastic factors related to cohort effects. The technical discussions are followed by illustrations based the shortlisted model structures fitted to the two pairs of populations under consideration.

6.1 Modeling Period Effects

6.1.1 The Candidate Processes

As discussed in Section 4, the evolution of period effects must meet certain requirements that ensure the resulting forecasts do not diverge over the long run. In general, we require the difference
Figure 4: Parameter estimates of the two-population version of Model M7 with $\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)} = \gamma_{t-x}^{(3,1)} = \kappa_t^{(3,2)} = \kappa_t^{(3)}$. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

$\kappa_t^{(i,1)} - \kappa_t^{(i,2)}$, where $i = i_0, \ldots, i_1$, to be mean-reverting. The values of $i_0$ and $i_2$ depend on the base model structure: for Models M1, M2 and M3, we have $i_0 = i_1 = 2$; for Models M5, M6 and M8, we have $i_0 = 1$ and $i_1 = 2$; and for Model M7, we have $i_0 = 1$ and $i_1 = 3$.

We consider the following three candidate time-series processes for modeling the evolution of period effects. All three processes guarantee that the aforementioned requirements are satisfied.

- **A Random Walk and a First Order Autoregression (RWAR)**

  This approach is used by Cairns et al. (2011) to demonstrate their modeling framework. In this approach, $\kappa_t^{(i,1)}$ is modeled by a random walk with drift, while the difference $\kappa_t^{(i,1)} - \kappa_t^{(i,2)}$ is modeled by a first order autoregressive process:

  \[
  \Delta \kappa_t^{(i,1)} = \mu^{(i,1)} + Z_t^{(i,1)},
  \]

  \[
  \kappa_t^{(i,1)} - \kappa_t^{(i,2)} = \mu^{(i,2)} + \phi^{(i,2)}(\kappa_{t-1}^{(i,1)} - \kappa_{t-1}^{(i,2)}) + Z_t^{(i,2)},
  \]

  where $\Delta$ is the first difference operator, $\mu^{(i,1)}$, $\mu^{(i,2)}$ and $\phi^{(i,2)}$ are model parameters, and $Z_t^{(i,1)}$ and $Z_t^{(i,2)}$ are random innovations. We require $|\phi^{(i,2)}| < 1$ so that the latter process is stationary, thereby ensuring $\kappa_t^{(i,1)} - \kappa_t^{(i,2)}$ is mean-reverting.

- **A Vector Autoregression (VAR)**
Table 4: The BIC values of all base model structures estimated in implementing the “top-down” approach. Population \( j = 1 \) is EW males and population \( j = 2 \) is Dutch males. The lowest three BIC values are displayed in boldface.

This modeling approach is briefly mentioned by Cairns et al. (2011) and is implemented by Zhou et al. (2014) in their work on two-population Lee-Carter modeling. The VAR for period effects can be written as follows:

\[
\Delta \kappa_t^{(i,j)} = \mu^{(i,j)} + \phi_1^{(i,j)} \Delta \kappa_{t-1}^{(i,1)} + \phi_2^{(i,j)} \Delta \kappa_{t-1}^{(i,2)} + Z_t^{(i,j)},
\]

for \( i = i_0, \ldots, i_1 \) and \( j = 1, 2 \), where \( \mu^{(i,j)} \), \( \phi_1^{(i,j)} \) and \( \phi_2^{(i,j)} \) are model parameters, and \( Z_t^{(i,1)} \) and \( Z_t^{(i,2)} \) are random innovations. Parameters \( \phi_1^{(1,1)} \) and \( \phi_2^{(1,2)} \) capture the auto-correlations of the changes in period effects (i.e., the dependence of \( \Delta \kappa_t^{(i,j_0)} \) and \( \Delta \kappa_t^{(i,j_1)} \) for \( j_0 = j_1 \)), whereas parameters \( \phi_2^{(1,1)} \) and \( \phi_1^{(1,2)} \) capture the cross-correlations (i.e., the dependence of \( \Delta \kappa_t^{(i,j_0)} \) and \( \Delta \kappa_t^{(i,j_1)} \) for \( j_0 \neq j_1 \)).

In estimating the VAR, we need to impose the following constraints on the model parameters:

\[
\frac{\mu^{(i,1)}}{1 - \phi_1^{(i,1)} - \phi_2^{(i,1)}} = \frac{\mu^{(i,2)}}{1 - \phi_1^{(i,2)} - \phi_2^{(i,2)}},
\]

for \( i = i_0, \ldots, i_1 \). These constraints imply that the long-term (unconditional) mean of \( \Delta \kappa_t^{(i,j)} \) (i.e., the change in \( \kappa_t^{(i,j)} \) per unit time) is

\[
M_{VAR}^{(i)} = \frac{\mu^{(i,1)}}{1 - \phi_1^{(i,1)} - \phi_2^{(i,1)}},
\]
which means over the long run, the period effects for the two populations are expected to change at the same rates. The non-divergence conditions are therefore satisfied.

- A Vector Error Correction Model (VECM)

Zhou et al. (2014) also use a VECM to model period effects. The VECM for period effects can be expressed as

\[
\Delta \kappa_{t}^{(i,j)} = \mu^{(i,j)} + \alpha^{(i,j)}(\kappa_{t-1}^{(i,1)} - \kappa_{t-1}^{(i,2)}) + \phi_1^{(i,j)} \Delta \kappa_{t-1}^{(i,1)} + \phi_2^{(i,j)} \Delta \kappa_{t-1}^{(i,2)} + Z_t^{(i,j)},
\]

for \(i = i_0, \ldots, i_1\) and \(j = 1, 2\), where \(\alpha^{(i,j)}\), \(\mu^{(i,j)}\), \(\phi_1^{(i,j)}\) and \(\phi_2^{(i,j)}\) are model parameters, and \(Z_t^{(i,1)}\) and \(Z_t^{(i,2)}\) are random innovations. In comparison to the VAR, the VECM contains additionally an error-correction term, \(\alpha^{(i,j)}(\kappa_{t-1}^{(i,1)} - \kappa_{t-1}^{(i,2)})\), which pulls the period effects to their long-term trend. Under the VECM, the long-term mean of \(\Delta \kappa_t^{(i,j)}\) is given by

\[
M_{VECM}^{(i)} = \frac{\mu^{(i,1)}(1 - \phi_1^{(i,2)} - \phi_2^{(i,2)}) - \mu^{(i,2)}(1 - \phi_1^{(i,1)} - \phi_2^{(i,1)})}{\alpha^{(i,2)}(1 - \phi_1^{(i,1)} - \phi_2^{(i,1)}) - \alpha^{(i,2)}(1 - \phi_1^{(i,2)} - \phi_2^{(i,2)})}.
\]

which implies non-divergence, as the expression is independent of \(j\).

In fitting the three processes, it is assumed that the vector of \(Z_t^{(i,j)}\)'s, where \(i = i_0, \ldots, i_1\) and \(j = 1, 2\), follows a multivariate normal distribution with a zero mean vector and a constant covariance matrix \(V\). It is further assumed that for \(t \neq u\), the vector of \(Z_t^{(i,j)}\)'s and the vector of \(Z_u^{(i,j)}\)'s are independent. On the basis of these distributional assumptions, one can derive the likelihood functions, from which maximum likelihood parameter estimates can be obtained. The likelihood functions for the case when there is only one pair of period effects can be found in the papers by Cairns et al. (2011) and Zhou et al. (2013). However, when there are multiple pairs of period effects, the likelihood functions are more complicated. In Appendix B, we derive the likelihood functions for the three processes in the general case when there are one or more pairs of period effects.

The model structures shortlisted in Stage II may contain one or more merged pairs of period effects. For instance, one of the shortlisted model structures for EW and Dutch males is Model M6 with a merged period effect. The process we use for modeling a merged period effect depends on what process is used for modeling the remaining period effects that are not merged. If a RWAR is used to model the remaining period effects, then a random walk with drift is used to model a merged period effect. If a VAR or VECM is used to model the remaining period effects, then an ARIMA(1,1,0) process is used to model a merged period effect. The processes for all (both merged and not merged) period effects are estimated jointly by maximizing the corresponding likelihood function, which is again derived on the basis of the previously mentioned distributional assumptions. In Appendix C, we explain how a likelihood function should be adapted when the model structure contains a merged period effect.

\[\text{An ARIMA}(1,1,0)\text{ process refers to an autoregressive integrated moving average process with an order of } p = 1, d = 1 \text{ and } q = 0. \text{ An ARIMA}(1,1,0)\text{ contains a first order autoregressive term, which is in line with the VAR and VECM processes.}\]
6.1.2 Simplifying the Processes

As we model more pairs of period effects, the number of parameters in the time-series processes become larger, due partly to the increase in the number of autoregressive parameters, i.e., $\phi^{(i,2)}$'s, $\phi^{(i,j)}_1$'s and $\phi^{(i,j)}_2$'s, and partly to the increase in the dimension of the covariance matrix $V$ for the random innovations. Among the large collection of parameters, some may not be particularly helpful in predicting the evolution of the period effects. It is therefore necessary to simplify the time-series processes, especially when dealing with base model structures such as Model M7 that contain several pairs of period effects.

The “top-down” approach is used here again to systematically simplify the time-series processes for period effects. The simplification procedure is detailed below.

1. Reduce the number of autoregressive parameters:

   (a) Consider the first pair of period effects ($i = i_0$). The autoregressive parameters for this pair of period effects are $\phi^{(i_0,1)}_1$, $\phi^{(i_0,1)}_2$, $\phi^{(i_0,2)}_1$ and $\phi^{(i_0,2)}_2$.

      i. Estimate the process with all four autoregressive parameters.

      ii. Simplify the process by omitting one or both cross-correlation parameters (i.e., $\phi^{(i_0,2)}_2$ and/or $\phi^{(i_0,1)}_1$). Identify the set of cross-correlation parameters that gives the best BIC value. Use this set of cross-correlation parameters for this pair of period effects thereafter.

      iii. Further simplify the process by omitting one or both auto-correlation parameters (i.e., $\phi^{(i_0,1)}_1$ and/or $\phi^{(i_0,2)}_2$). Find the set of auto-correlation parameters that gives the best BIC value. Use this set of auto-correlation parameters for this pair of period effects thereafter.

   (b) Repeat Step (a) for the other pairs of period effects.

2. Simplify the covariance matrix for the random innovations:

   (a) Fit the process with a complete covariance matrix.

   (b) Fit the processes with one covariance parameter (i.e., an off-diagonal element in the covariance matrix $V$) equal to 0. By comparing the resulting BIC values, identify the collection of covariance parameters that can potentially be omitted.

   (c) Fit simplified processes with two covariance parameters (chosen from the collection in Step (b)) equal to 0. This step is necessary only if it was found in Step (b) that two or more covariance parameters can potentially be omitted.

   (d) Fit simplified models with three or more covariance parameters equal to 0, if there is a need to do so.

   (e) Select the simplified covariance matrix that gives the best BIC value.
Note that the autoregressive parameters in a RWAR cannot be reduced, since they are needed to ensure the non-divergence conditions hold. Also, when there is only one pair of period effects, the covariance matrix is not simplified.

If Model M7 is used as the base model structure, then the covariance matrix of the random innovations is particularly large, containing up to 21 distinct elements. It can therefore be very time-consuming to implement the “top-down” approach described above fully. To reduce computational effort, we propose an abridged version of the “top-down” approach, in which the parameters in the covariance matrix are first divided into groups. The simplification procedure is then applied to the groups of parameters rather than the individual parameters.

Let us consider the case when the base model structure is Model M7 without any merged period effects. In this case, the covariance matrix \( V \) has a dimension of \( 6 \times 6 \). To implement the abridged “top-down” approach, we first divide the covariance matrix into nine 2-by-2 matrices, namely \( V_{ij} \) for \( i, j = 1, 2, 3 \), where \( V_{ij} \) denotes the covariance matrix between the \( i \)th pair of period effects and the \( j \)th pair of period effects. The division of \( V \) is illustrated in Figure 5. Then, the following procedure is used to simplify the covariance matrix.

1. Fit the process with a complete covariance matrix.

2. Fit processes with the following simplifications in the covariance matrix:
   - \( V_{12} = 0 \)
   - \( V_{13} = 0 \)
   - \( V_{23} = 0 \)

   By comparing the resulting BIC values, identify the 2-by-2 matrices that can potentially be set to zero.

3. If it was found in Step 2 that two or more 2-by-2 matrices can potentially be set to zero, fit processes with the following simplifications:
   - \( V_{12} = 0 \) and \( V_{13} = 0 \)
   - \( V_{12} = 0 \) and \( V_{23} = 0 \)
   - \( V_{13} = 0 \) and \( V_{23} = 0 \)
   - \( V_{12} = 0, V_{13} = 0, \) and \( V_{23} = 0 \)

4. Select the simplification that gives the best BIC value.

Note that \( V_{11}, V_{22}, \) and \( V_{33} \) cannot be set to zero in the procedure above because they involve the variance of \( Z_t^{(i,j)} \).

In the case when one or more pairs of period effects in Model M7 are merged, then the division of \( V \) is slightly different. For example, if the third pair of period effects are merged, then we divide
Figure 5: The division of the covariance matrix in the implementation of the abridged “top-down” approach for Model M7.

The 5 × 5 covariance matrix of the vector of innovations, \((Z_t^{(1,1)}, Z_t^{(1,2)}, Z_t^{(2,1)}, Z_t^{(2,2)}, Z_t^{(3,3)})\), is divided into the following three groups: \((V(1, 3), V(1, 4), V(2, 3), V(2, 4), (V(1, 5), V(2, 5)))\) and \((V(3, 5), V(4, 5))\), where \(V(i, j)\) denotes the \((i, j)\)th element in the covariance matrix.

6.2 Modeling Cohort Effects

The candidate model structures in Table 1 contain at most one pair of cohort effects. If the cohort effects of the two populations are merged, then the ARIMA process with the best BIC value is used to model the evolution of the merged cohort effect. If the cohort effects are not merged, then the pair of cohort effects is modeled by either one of the following three bivariate process: a VAR, a VECM and a gravity model. The VAR and VECM processes for the cohort effects are the same as those shown in Section 6.1.1, except that \(\gamma_{i-x}\)'s are replaced by \(\gamma_{i-x}'s\). The gravity process, proposed by Dowd et al. (2011), can be expressed as follows:

\[
\begin{align*}
\Delta \gamma^{(i,1)}_c &= \mu^{(i,1)} + \phi^{(i,1)} \Delta \gamma^{(i,1)}_{c-1} + Z^{(i,1)}_c, \\
\Delta \gamma^{(i,2)}_c &= \mu^{(i,2)} + \phi^{(i,2)} (\gamma^{(i,1)}_{c-1} - \gamma^{(i,2)}_{c-1}) + \phi^{(i,2)} \Delta \gamma^{(i,2)}_{c-1} + Z^{(i,2)}_c,
\end{align*}
\]

where \(c = t - x\) denotes the year of birth, \(\mu^{(i,1)}, \mu^{(i,2)}, \phi^{(i,1)}\) and \(\phi^{(i,2)}\) are model parameters, and \(Z^{(1,1)}_c\) and \(Z^{(1,2)}_c\) are the random innovations, which are assumed to follow a bivariate normal distribution with a zero mean vector and a constant covariance matrix. The term \(\phi^{(i,2)} (\gamma^{(i,1)}_{c-1} - \gamma^{(i,2)}_{c-1})\), where \(|\phi^{(i,2)}| < 1\), reduces any increasing spread between the cohort effects of the two populations back to a constant level. The likelihood function for the gravity process is provided in the paper by Dowd et al. (2011). The “top-down” approach is used again to reduce the number of autoregressive parameters (i.e., parameters \(\phi^{(i,1)}\) and \(\phi^{(i,2)}\)) in the three bivariate processes.
<table>
<thead>
<tr>
<th>Simplification</th>
<th>Autoregressive parameters (cross-correlations)</th>
<th>BIC</th>
<th>Autoregressive parameters (auto-correlations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>126.0853</td>
<td>121.8433</td>
</tr>
<tr>
<td>$\phi_2^{(1,1)} = 0$</td>
<td></td>
<td>120.3632</td>
<td>123.4873</td>
</tr>
<tr>
<td>$\phi_1^{(1,2)} = 0$</td>
<td></td>
<td>118.3848</td>
<td>121.5227</td>
</tr>
<tr>
<td>$\phi_2^{(1,1)} = 0$, $\phi_1^{(1,2)} = 0$</td>
<td></td>
<td>124.4615</td>
<td>128.6329</td>
</tr>
</tbody>
</table>

Table 5: The BIC values of all time-series processes estimated in implementing the “top-down” approach for the period effects in Model M2. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives. The lowest BIC value is displayed in boldface.

6.3 Illustrations

6.3.1 EW Males and UK Insured Lives

In Stage II, we shortlisted Model M2 with a merged cohort effect, Model M3 with a merged cohort effect and Model M6 with a merged cohort effect for this pair of populations. In Tables 5 to 7 we show the BIC values involved in the implementation of the “top-down” approach for modeling the period effects in these base model structures.

Let us use the VECM process for the period effects in the third shortlisted model structure to demonstrate how a process is simplified in the “top-down” approach. The relevant BIC values are displayed in the last column of Table 7.

1. Consider the autoregressive parameters that are related to cross-correlations for the first pair of period effects. It is found that the omission of both $\phi_2^{(1,1)}$ and $\phi_1^{(1,2)}$ leads to a better BIC value.

2. Given $\phi_2^{(1,1)} = \phi_1^{(1,2)} = 0$, estimate the processes with one or both of the auto-correlation parameters (for the first pair of period effects) equal zero. It is found that omitting the auto-correlation parameters does not improve the BIC value.

3. Consider the autoregressive parameters that are related to cross-correlations for the second pair of cohort effects. It is found that the omission of both $\phi_2^{(2,1)}$ and $\phi_1^{(2,2)}$ leads to a better BIC value.

4. Given $\phi_2^{(1,1)} = \phi_1^{(1,2)} = \phi_2^{(2,1)} = \phi_1^{(2,2)} = 0$, estimate the processes with one or both of the auto-correlation parameters (for the second pair of period effects) equal zero. It is found
### Table 6: The BIC values of all time-series processes estimated in implementing the “top-down” approach for the period effects in Model M3. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives. The lowest BIC value is displayed in boldface.

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Autoregressive parameters (cross-correlations)</th>
<th>Autoregressive parameters (auto-correlations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>BIC</td>
</tr>
<tr>
<td>None</td>
<td>127.4820</td>
<td>126.2344</td>
</tr>
<tr>
<td>$\phi_{(1,1)} = 0$</td>
<td>123.7855</td>
<td>128.3661</td>
</tr>
<tr>
<td>$\phi_{(1,2)} = 0$</td>
<td>122.2140</td>
<td>122.0635</td>
</tr>
<tr>
<td>$\phi_{2}^{(2,1)} = 0, \phi_{1}^{(1,2)} = 0$</td>
<td>120.4360</td>
<td>126.5091</td>
</tr>
<tr>
<td></td>
<td>122.4628</td>
<td>125.4970</td>
</tr>
<tr>
<td></td>
<td>122.3145</td>
<td>122.0537</td>
</tr>
<tr>
<td></td>
<td>120.2844</td>
<td>124.8948</td>
</tr>
</tbody>
</table>

That omitting $\phi_{2}^{(2,2)}$ can further improve the BIC value.

5. Given $\phi_{2}^{(1,1)} = \phi_{1}^{(1,2)} = \phi_{2}^{(2,1)} = \phi_{2}^{(2,2)} = 0$, fit processes with one element in $V$ equal zero. It is found that setting $V(1,4) = 0$ or $V(2,4) = 0$ improves the BIC value.

6. Fit the process with $\phi_{2}^{(1,1)} = \phi_{1}^{(1,2)} = \phi_{2}^{(2,1)} = \phi_{1}^{(2,2)} = V_{2}^{(1,4)} = V_{2}^{(2,4)} = 0$. It is found that setting both $V(1,4)$ and $V(2,4)$ to zero further improves the BIC value. Because it was previously found that setting any other element in $V$ to zero does not yield a better BIC value, there is no need to further simplify $V$. The simplified process fitted in this step is the final VECM process.

The processes selected for the stochastic factors in the three shortlisted base model structures for EW males and UK male insured lives are summarized as follows:

- **Model M2 with a merged cohort effect:**
  A VAR with $\phi_{2}^{(1,1)} = \phi_{1}^{(1,2)} = 0$ for the period effects and an ARIMA(2,1,0) for the merged cohort effect

- **Model M3 with a merged cohort effect:**
  A VECM with $\phi_{2}^{(1,1)} = \phi_{1}^{(1,2)} = 0$ for the period effects and an ARIMA(1,1,0) for the merged cohort effect

- **Model M6 with a merged cohort effect:**
  A VAR with $\phi_{2}^{(1,1)} = \phi_{1}^{(1,2)} = \phi_{2}^{(2,1)} = \phi_{1}^{(2,2)} = V(1,4) = V(2,4) = 0$ for the period effects and an ARIMA(2,1,0) for the merged cohort effect
### Table 7: The BIC values of all time-series processes estimated in implementing the “top-down” approach for the period effects in Model M6. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives. The lowest BIC value is displayed in boldface.
### 6.3.2 EW Males and Dutch Males

In Table 8 we show the BIC values of some of the estimated time-series processes for EW and Dutch males. For brevity, we only show the best BIC value we obtained for each combination of base model structure and time-series process.\(^6\)

The processes selected for the stochastic factors in the three shortlisted base model structures for EW males and Dutch males are summarized as follows:

- The full version of Model M2:
  
  A VECM with \(\phi_{2}^{(2,1)} = 0\) for the period effects and a VECM with \(\phi_{1}^{(3,1)} = \phi_{2}^{(3,1)} = \phi_{1}^{(3,2)} = \phi_{2}^{(3,2)} = 0\) for the cohort effects

- Model M6 with \(\kappa_{(2,1)}^t = \kappa_{(2,2)}^t\):
  
  A VECM with \(\phi_{2}^{(2,1)} = 0\) for the period effects and a VAR with \(\phi_{1}^{(3,1)} = \phi_{2}^{(3,1)} = \phi_{1}^{(3,2)} = \phi_{2}^{(3,2)} = 0\) for the cohort effects

- Model M8 with \(\kappa_{(2,1)}^t = \kappa_{(2,2)}^t\):
  
  A VECM with \(\phi_{2}^{(2,1)} = 0\) for the period effects and a VAR with \(\phi_{1}^{(3,1)} = \phi_{2}^{(3,1)} = \phi_{2}^{(3,2)} = 0\) for the cohort effects

### 7 Stage IV: Model Evaluation

#### 7.1 Evaluation Criteria

In Stage I, we shortlist several base model structures rather than simply picking the one that gives the best BIC value, because when considered together with the time-series processes, the base model structure with the highest BIC ranking may not necessarily be best choice. It follows that there is a need for an additional step to select a combination of base model structure and

\(^6\)The BIC values shown in Table 8 cannot be compared across different base model structures. This is because the log-likelihoods in this stage are conditional on the estimated period and cohort effects, which are different among different base model structures.
time-series processes for the pair of populations in question. This important step is accomplished in Stage IV of the model-building procedure.

The selection cannot be based on BIC values, because in using a two-stage estimation methodology, we do not have the necessary information from which the BIC value for a combination of base model structure and time-series processes can be calculated. To explain why, let us first define the following notation. We let \( \theta \) be the set of all parameters in the base model structure and the time-series processes, \( \theta_1 \) be the set of all parameters in the base model structure, \( \theta_2 \) be the stochastic factors, and \( D \) be the data that is composed of the historical death and exposure counts. Note that \( \theta_2 \subset \theta_1 \subset \theta \). To calculate the BIC value for the combination of base model structure and time-series processes, we require the maximized value of \( L(\theta|D) \), the likelihood of \( \theta \) given the data \( D \). What we obtain from Stages II and III are the maximized values of \( L_1(\theta_1|D) \), the likelihood of \( \theta_1 \) given \( D \), and \( L_2(\theta_3|\hat{\theta}_2) \), the likelihood of \( \theta_3 \) given \( \hat{\theta}_2 \), the estimated stochastic factors. These two maximized likelihoods, however, are not sufficient to imply what we need. In particular, the product of the maximized values of \( L_1(\theta_1|D) \) and \( L_2(\theta_3|\hat{\theta}_2) \) is not the maximized value of \( L(\theta|D) \).

The selection criteria we use in this stage of the model-building process are robustness and in-sample forecasting performance. To measure robustness, we examine how forecasts will change when different look-back windows are used. To measure in-sample forecasting performance, we first re-estimate the mortality models to data over a restricted sample period, say \([t_0, t^*] \), where \( t_0 < t^* < t_1 \), and then examine the accuracy of the forecasts generated from the re-estimated models over the time period from \( t^* + 1 \) to \( t_1 \). In our work, the quantities being forecasted are cohort mortality rates. However, depending on the ultimate purpose of the mortality model, other quantities, such as annuity rates, life expectancies and cohort survival probabilities, may also be used in this step.

### 7.2 Illustrations

#### 7.2.1 EW Males and UK Male Insured Lives

We re-estimate the three selected combinations of base model structure and time-series processes to two alternative look-back windows: 45 years (1961-2005) and 25 years (1981-2005). We then produce, for each model and look-back window, forecasts of central death rates for the cohort aged 65 in 2006. If a model is robust, then it should generate similar forecasts even if it is fitted to different look-back windows. The results, presented in Figure 6, suggest that for EW males and UK male insured lives, Model M6 with (with a merged cohort effect and a VAR process of the period effects) is the most robust with respect to changes in the data sample period.

Next, we re-fit the three combinations of base model structure and time-series processes to a restricted sample period (1947-1990). We then compare the mean and 90% interval forecasts of the central death rates (for the cohort aged 65 in 1991) generated from the re-fitted models with
Figure 6: Forecasts of the central death rates for the cohort aged 65 in 2006, based on models fitted to data over the periods of 1947-2005, 1961-2005 and 1981-2005. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

the corresponding actual values. The results, which are shown in Figure 7, indicate that Model M6 (with a merged cohort effect and a VAR process of the period effects) produces forecasts that are the most in line with the actual values. Two other restricted sample periods (1947-1985 and 1947-1995) have also been considered, and the results point to the same conclusion.

On the basis of these two analyses, we select Model M6 (with a merged cohort effect and a VAR process of the period effects) for EW males and UK male insured lives.

7.2.2 EW Males and Dutch Males

We repeat the analyses of robustness and in-sample forecasting performance for EW Males and Dutch Males.

Figure 8 shows the forecasts of the central death rates for the cohort aged 65 in 2006, produced by models that are fitted to data over the periods of 1947-2005, 1961-2005 and 1981-2005. The results indicate that Model M8 (with $\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$, a VECM process for the period effects and a VAR process for the cohort effects) is the most robust with respect to changes in the sample period. The other two models produce similar degrees of robustness.

However, the performance of Model M8 in the aspect of in-sample forecasting is not satisfactory. This problem can be seen from Figure 9, in which we compare the actual central death rates (for the cohort aged 65 in 1991) with the corresponding mean and 90% interval forecasts generated from the models that are re-fitted to a restricted sample period of 1947-1990. We observe that for EW males, Model M8 produces a mean forecast that is significantly biased high and a forecast
Figure 7: Mean and 90% interval forecasts of the central death rates for the cohort aged 65 in 1991, based on models fitted to data over the period 1947 to 1990. Population $j = 1$ is EW males and population $j = 2$ is UK male insured lives.

Figure 8: Forecasts of the central death rates for the cohort aged 65 in 2006, based on models fitted to data over the periods of 1947-2005, 1961-2005 and 1981-2005. Population $j = 1$ is EW males and population $j = 2$ is Dutch males.
Figure 9: Mean and 90% interval forecasts of the central death rates for the cohort aged 65 in 1991, based on models fitted to data over the period 1947 to 1990. Population $j = 1$ is EW males and population $j = 2$ is Dutch males.

interval that cannot adequately capture the actual values. Among the three models, Model M6 (with $\kappa^{(2,1)}_t = \kappa^{(2,2)}_t$, a VECM process for the period effects and a VAR process for the cohort effects) produces the best in-sample forecasting performance.

Hence, the conclusion for this pair of populations is not definite. The ultimate choice depends very much on the user’s priority and involves some degrees of subjectivity. For instance, in an application where the model will be constantly re-calibrated, robustness may be the user’s first priority, which means Model M8 should be chosen.

8 Concluding Remarks

One of the biggest challenges in the development of a large and liquid market for longevity risk transfers is the conflicting interests between the buy-side and sell-side of the market. Buyers of longevity risk generally prefer standardized securities that are linked to broad-based mortality indexes, while sellers may have concerns about the population basis risk entailed in using standardized instruments. Two-population stochastic mortality models allow market participants to better estimate population basis risk, thereby narrowing the gap between both sides of the market.

This paper addresses the need for two-population stochastic models by contributing a step-by-step procedure for generalizing existing single-population mortality models to their two-population versions. To prevent the resulting model from being overly complicated, we propose to use a “top-down” approach to systematically simplify the base model structure and the time-series processes involved. The two-population mortality models developed from our proposed procedure can be
used readily with the frameworks of Li and Hardy (2011) and Zhou et al. (2011) to the quantify population basis risk and to analyze longevity risk transfers involving two populations.

We have illustrated the proposed model-building procedure using mortality data from two pairs of populations. For EW males and UK insured lives, the final choice is Model M6 (with a merged cohort effect and a VAR process of the period effects). For EW males and Dutch males, the conclusion is not definite, but Model M8 (with $\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$, a VECM process for the period effects and a VAR process for the cohort effects) appears to be the most appropriate if robustness is the user’s primary concern. The models selected for the two pairs of populations are rather different, indicating that desirable features such as a merged cohort effect are data dependent. It may therefore be more reasonable to consider a model-building procedure that searches for the most suitable model, instead of aiming to develop a universal two-population model that fits all situations.

To focus on the problem of model building, in this paper we do not consider parameter risk, which arises because the true parameter values can never be known. If this risk is ignored, one may potentially under-estimate the actual amount of population basis risk. It is therefore warranted in future research to include parameter uncertainty in the model-building procedure, possibly by using Bayesian approaches (see Cairns et al., 2006, 2011) or bootstrapping methods (see Li, 2010; Renshaw and Haberman, 2008). Another avenue for further research is to study how the model-building procedure can be extended to incorporate mortality jumps, which capture catastrophic mortality events such as a widespread pandemic. The multinomial method recently proposed by Zhou et al. (2013) could be a possible starting point.

Acknowledgments

The authors acknowledge the financial support from the Global Risk Institute, the Center of Actuarial Excellence Program of the Society of Actuaries and the Natural Science and Engineering Research Council of Canada. The authors would also like to express their gratitude to Professor Andrew Cairns and other participants at the Ninth International Longevity Risk and Capital Markets Solutions Conference for their stimulating discussions.

References


### Appendix A

#### Identifiability Constraints for the Base Model Structures

In estimating base model structures, identifiability constraints are needed. The necessary identifiability constraints depend on the model structure as well as the simplification(s) made. Below we provide the necessary identifiability constraints for each possible model specification.

- **Model M1**

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Identifiability Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$\sum_x \beta_x^{(2)} = 1$, $\sum_t \kappa_t^{(2,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$</td>
<td>$\sum_x \beta_x^{(2)} = 1$, $\sum_t \kappa_t^{(2,j)} = 0$</td>
</tr>
</tbody>
</table>

- **Model M2**

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Identifiability Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$\sum_x \beta_x^{(2)} = 1$, $\sum_x \beta_x^{(3)} = 1$, $\sum_t \kappa_t^{(2,j)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$, $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$</td>
<td>$\sum_x \beta_x^{(2)} = 1$, $\sum_x \beta_x^{(3)} = 1$, $\sum_t \kappa_t^{(2,j)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
</tbody>
</table>

- **Model M3**

  - No simplification

  The necessary identifiability constraints for this case are $\sum_t \kappa_t^{(2,j)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3,j)} = 0$, $\min_{\delta(j)} \sum_x (\beta_x^{(1,j)} + \delta(j)(x - \bar{x}) - \beta_x^{(1,j)})^2$, where $\beta_x^{(1,j)} = n_y^{-1} \sum_t \ln n_{x,t}^{(j)}$ and $n_y = t_1 - t_0 + 1$ is the length of the data sample period. The last constraint implies that $\delta(j) = \frac{-\sum_x (\beta_x^{(1,j)} - \beta_x^{(1,j)}(x - \bar{x}))}{\sum_x (x - \bar{x})^2}$.

  To enforce the last constraint, the parameter estimates are revised according to the following formulas at the end of each Newton-Raphson iteration: $\gamma_{t-x}^{(3,j)} = \gamma_{t-x}^{(3,j)} + n_a \delta(j)((t - \bar{t}) - (x - \bar{x}))$, $\kappa_t^{(2,j)} = \kappa_t^{(2,j)} - n_a \delta(j)(t - \bar{t})$, and $\beta_x^{(1,j)} = \beta_x^{(1,j)} + \delta(j)(x - \bar{x})$.  

30
- Simplified by setting $\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$
  The necessary identifiability constraints for this case are $\sum_t \kappa_t^{(2)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3,j)} = 0$ and $\min_\delta \sum_{x,j} (\beta_x^{(1,j)} + \delta(x - \bar{x} - \beta_x^{(1,j)}))^2$. The last constraint implies that $\delta = -\frac{\sum_{x,j} (\beta_x^{(1,j)} - \bar{\beta}_x^{(1,j)}) (x - \bar{x})}{\sum_x (x - \bar{x})^2}$.
  To enforce the last constraint, we revise the parameter estimates according to the following formulas at the end of each Newton-Raphson iteration: $\tilde{\gamma}_{t-x}^{(3,j)} = \gamma_{t-x}^{(3,j)} + n_a \delta((t - \bar{t}) - (x - \bar{x}))$, $\tilde{\kappa}_t^{(2,j)} = \kappa_t^{(2,j)} - n_a \delta(t - \bar{t})$ and $\tilde{\beta}_x^{(1,j)} = \beta_x^{(1,j)} + \delta(x - \bar{x})$.

- Simplified by setting $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$
  The necessary identifiability constraints for this case are $\sum_t \kappa_t^{(2,j)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3)} = 0$ and $\min_\delta \sum_{x,j} (\beta_x^{(1,j)} + \delta(x - \bar{x}) - \bar{\beta}_x^{(1,j)})^2$.
  To enforce the last constraint, we revise the parameter estimates according to the following formulas at the end of each Newton-Raphson iteration: $\gamma_{t-x}^{(3)} = \gamma_{t-x}^{(3)} + n_a \delta((t - \bar{t}) - (x - \bar{x}))$, $\tilde{\kappa}_t^{(2,j)} = \kappa_t^{(2,j)} - n_a \delta(t - \bar{t})$ and $\tilde{\beta}_x^{(1,j)} = \beta_x^{(1,j)} + \delta(x - \bar{x})$.

- Simplified by setting $\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$ and $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$
  The necessary identifiability constraints for this case are $\sum_t \kappa_t^{(2,j)} = 0$, $\sum_{t,x} \gamma_{t-x}^{(3)} = 0$ and $\min_\delta \sum_{x,j} (\beta_x^{(1,j)} + \delta(x - \bar{x}) - \bar{\beta}_x^{(1,j)})^2$.
  To enforce the last constraint, we revise the parameter estimates according to the following formulas at the end of each Newton-Raphson iteration: $\gamma_{t-x}^{(3)} = \gamma_{t-x}^{(3)} + n_a \delta((t - \bar{t}) - (x - \bar{x}))$, $\tilde{\kappa}_t^{(2,j)} = \kappa_t^{(2,j)} - n_a \delta(t - \bar{t})$ and $\tilde{\beta}_x^{(1,j)} = \beta_x^{(1,j)} + \delta(x - \bar{x})$.

- Model M5
  No identifiability constraint is necessary.

- Model M6

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Identifiability Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$\sum_c \gamma_c^{(3,j)} = 0$, $\sum_c c\gamma_c^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(1,1)} = \kappa_t^{(1,2)}$, $\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$, $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$</td>
<td>$\sum_c \gamma_c^{(3,j)} = 0$, $\sum_{c,j} c\gamma_c^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$, $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$</td>
<td>$\sum_c \gamma_c^{(3,j)} = 0$, $\sum_{c,j} c\gamma_c^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(1,1)} = \kappa_t^{(1,2)}$, $\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$</td>
<td>$\sum_c \gamma_c^{(3)} = 0$, $\sum_c c\gamma_c^{(3)} = 0$</td>
</tr>
</tbody>
</table>

- Model M7
### Identifiability Constraints

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Identifiability Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$\sum_c \gamma_c^{(4,j)} = 0$, $\sum_c c \gamma_c^{(4,j)} = 0$, $\sum_c c^2 \gamma_c^{(4,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(1,1)} = \kappa_t^{(1,2)}$</td>
<td>$\sum_{c,j} \gamma_c^{(4,j)} = 0$, $\sum_{c,j} c \gamma_c^{(4,j)} = 0$, $\sum_{c,j} c^2 \gamma_c^{(4,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$</td>
<td>$\sum_{c,j} \gamma_c^{(4,j)} = 0$, $\sum_{c,j} c \gamma_c^{(4,j)} = 0$, $\sum_{c,j} c^2 \gamma_c^{(4,j)} = 0$</td>
</tr>
<tr>
<td>$\gamma_{t-x}^{(4,1)} = \gamma_{t-x}^{(4,2)}$</td>
<td>$\sum_c \gamma_c^{(4)} = 0$, $\sum_c c \gamma_c^{(4)} = 0$, $\sum_c c^2 \gamma_c^{(4)} = 0$</td>
</tr>
</tbody>
</table>

- **Model M8**

<table>
<thead>
<tr>
<th>Simplification</th>
<th>Identifiability Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>$\sum_{t,x,j} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(1,1)} = \kappa_t^{(1,2)}$</td>
<td>$\sum_{t,x,j} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\kappa_t^{(2,1)} = \kappa_t^{(2,2)}$</td>
<td>$\sum_{t,x,j} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
<tr>
<td>$\gamma_{t-x}^{(3,1)} = \gamma_{t-x}^{(3,2)}$</td>
<td>$\sum_{t,x,j} \gamma_{t-x}^{(3,j)} = 0$</td>
</tr>
</tbody>
</table>

In the above, the summations are taken over the entire age range (i.e., $x = x_0, \ldots, x_1$), the entire sample period (i.e., $t = t_0, \ldots, t_1$) and/or the entire range of years of birth (i.e., $c = t_0 - x_1, \ldots, t_1 - x_0$), where appropriate.

### Appendix B

**Deriving the Likelihood Functions for the Period Effect Processes**

Let us first define the following notation:

\[
\mu = \begin{pmatrix}
\mu^{(i_0,1)} \\
\mu^{(i_0,2)} \\
\mu^{(i_0+1,1)} \\
\vdots \\
\mu^{(i_1,1)} \\
\mu^{(i_1,2)}
\end{pmatrix}, \quad Z = \begin{pmatrix}
Z_t^{(i_0,1)} \\
Z_t^{(i_0,2)} \\
Z_t^{(i_0+1,1)} \\
\vdots \\
Z_t^{(i_1,1)} \\
Z_t^{(i_1,2)}
\end{pmatrix}, \quad M_{VAR} = \begin{pmatrix}
M_{VAR}^{(i_0)} \\
M_{VAR}^{(i_0)} \\
M_{VAR}^{(i_0+1)} \\
\vdots \\
M_{VAR}^{(i_1)} \\
M_{VAR}^{(i_1)}
\end{pmatrix}, \quad M_{VECM} = \begin{pmatrix}
M_{VECM}^{(i_0)} \\
M_{VECM}^{(i_0)} \\
M_{VECM}^{(i_0+1)} \\
\vdots \\
M_{VECM}^{(i_1)} \\
M_{VECM}^{(i_1)}
\end{pmatrix},
\]

32
\[
S_t = \begin{pmatrix}
\kappa_t^{(i_0,1)} \\
\kappa_t^{(i_0,1)} - \kappa_t^{(i_0,2)} \\
\kappa_t^{(i_0+1,1)} - \kappa_t^{(i_0+1,2)} \\
\vdots \\
\kappa_t^{(i_1,1)} - \kappa_t^{(i_1,2)} \\
\end{pmatrix}, \quad \phi_{\text{RWAR}} = \begin{pmatrix}
1 & 0 & 0 & 0 & \ldots & 0 & 0 \\
0 & \phi^{(i_0,2)} & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & 1 & 0 & \ldots & 0 & 0 \\
0 & 0 & 0 & \phi^{(i_0+1,2)} & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & 1 & 0 \\
0 & 0 & 0 & 0 & \ldots & 0 & \phi^{(i_1,2)} \\
\end{pmatrix},
\]

\[
\Delta \kappa_t = \begin{pmatrix}
\kappa_t^{(i_0,1)} - \kappa_t^{(i_0-1)} \\
\kappa_t^{(i_0,2)} - \kappa_t^{(i_0-1)} \\
\kappa_t^{(i_0+1,1)} - \kappa_t^{(i_0+1-1)} \\
\kappa_t^{(i_0+1,2)} - \kappa_t^{(i_0+1-1)} \\
\vdots \\
\kappa_t^{(i_1,1)} - \kappa_t^{(i_1-1)} \\
\kappa_t^{(i_1,2)} - \kappa_t^{(i_1-1)} \\
\end{pmatrix}, \quad \phi = \begin{pmatrix}
\phi^{(i_0,1)} \\
\phi^{(i_0,2)} \\
0 & \phi^{(i_0+1,1)} \\
0 & \phi^{(i_0+1,2)} \\
0 & \ldots & \phi^{(i_1,1)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & \phi^{(i_1,1)} \\
0 & 0 & 0 & 0 & \ldots & \phi^{(i_1,2)} \\
\end{pmatrix},
\]

\[
\kappa_t = \begin{pmatrix}
\kappa_t^{(i_0,1)} \\
\kappa_t^{(i_0,2)} \\
\kappa_t^{(i_0+1,1)} \\
\kappa_t^{(i_0+1,2)} \\
\vdots \\
\kappa_t^{(i_1,1)} \\
\kappa_t^{(i_1,2)} \\
\end{pmatrix}, \quad \alpha = \begin{pmatrix}
\alpha^{(i_0,1)} & -\alpha^{(i_0,1)} & 0 & 0 & \ldots & 0 & 0 \\
\alpha^{(i_0,2)} & -\alpha^{(i_0,2)} & 0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \alpha^{(i_0+1,1)} & -\alpha^{(i_0+1,1)} & \ldots & 0 & 0 \\
0 & 0 & \alpha^{(i_0+1,2)} & -\alpha^{(i_0+1,2)} & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & \alpha^{(i_1,1)} & -\alpha^{(i_1,1)} \\
0 & 0 & 0 & 0 & \ldots & \alpha^{(i_1,2)} & -\alpha^{(i_1,2)} \\
\end{pmatrix}.
\]

We use \(f(\cdot|\cdot)\) to denote a generic conditional density function, where the conditioning variables follow the vertical bar, and \(\text{mvnpdf}(X, \nu, \mathcal{V})\) to denote the density function, evaluated at \(X\), for a multivariate normal random vector with a mean vector \(\nu\) and a covariance matrix \(\mathcal{V}\). As previously defined, \(V\) is the covariance matrix of the vector of innovations, \(Z_t\).

The conditional log-likelihood for the RWAR process is given by

\[
\ln(f(S_{t_{i_0}+1}, S_{t_{i_0}+2}, \ldots, S_{t_1}|S_{t_0})) = \sum_{t=t_{i_0}}^{t_{i_1}-1} \ln(f(S_{t+1}|f(S_t))) = \sum_{t=t_{i_0}}^{t_{i_1}-1} \ln(\text{mvnpdf}(S_{t+1}, \mu + \phi_{\text{RWAR}}S_t, V)).
\]
The conditional log-likelihood for the VAR process is given by
\[
\ln(f(\Delta\kappa_{t_0+1}, \Delta\kappa_{t_0+2}, \ldots, \Delta\kappa_{t_1})) = \ln(f(\Delta\kappa_{t_0+1})) + \sum_{t=t_0+1}^{t_1-1} \ln(f(\Delta\kappa_{t+1}|\Delta\kappa_t)) = \ln(\text{mvnpdf}(\Delta\kappa_{t_0+1}, M_{VAR}, V)) + \sum_{t=t_0+1}^{t_1-1} \ln(\text{mvnpdf}(\Delta\kappa_{t+1}, \mu + \phi\Delta\kappa_t, V)).
\]

The conditional log-likelihood for the VECM process is given by
\[
\ln(f(\Delta\kappa_{t_0+1}, \Delta\kappa_{t_0+2}, \ldots, \Delta\kappa_{t_1}|\kappa_{t_0})) = \ln(f(\Delta\kappa_{t_0+1})) + \sum_{t=t_0+1}^{t_1-1} \ln(f(\Delta\kappa_{t+1}|\Delta\kappa_t, \kappa_t)) = \ln(\text{mvnpdf}(\Delta\kappa_{t_0+1}, M_{VECM}, V)) + \sum_{t=t_0+1}^{t_1-1} \ln(\text{mvnpdf}(\Delta\kappa_{t+1}, \mu + \alpha\kappa_t + \phi\Delta\kappa_t, V)).
\]

Appendix C

Adaptations to the Likelihood Functions when There Are Merged Period Effects

Recall that the process we use for modeling a merged period effect depends on what process is used for modeling the remaining period effects that are not merged. If a RWAR is used to model the remaining period effects, then the merged period effect, \(\kappa^{(i)}_t\), is modeled by a random walk with drift:
\[
\Delta\kappa^{(i)}_t = \mu^{(i)} + Z^{(i)}_t,
\]
where \(\mu^{(i)}\) is the drift term and \(Z^{(i)}_t\) is the random innovation, which is assumed to be normally distributed with a zero mean and a constant variance. If a VAR or VECM is used to model the remaining period effects, then the merged period effect is modeled by an ARIMA(1,1,0) process:
\[
\Delta\kappa^{(i)}_t = \mu^{(i)} + \phi^{(i)} \Delta\kappa^{(i)}_{t-1} + Z^{(i)}_t,
\]
where \(\phi^{(i)}\) is an autoregressive parameter. Let \(M^{(i)}\) be the long-term mean of \(\Delta\kappa^{(i)}_t\). If \(\kappa^{(i)}_t\) follows a random walk with drift, then \(M^{(i)} = \mu^{(i)}\). If \(\kappa^{(i)}_t\) follows an ARIMA(1,1,0), then \(M^{(i)} = \frac{\mu^{(i)}}{1-\phi^{(i)}}\).

To illustrate how the likelihood functions should be adapted when there exists a merged period effect, let us suppose that the base model structure is Model M7 with both population sharing the same third period effect. In this case, the log-likelihoods of the period effect processes are the same as those given in Appendix B, except the following changes:
\[
\begin{align*}
\mu &= \left( \begin{array}{c}
\mu^{(1,1)} \\
\mu^{(1,2)} \\
\mu^{(2,1)} \\
\mu^{(2,2)} \\
\mu^{(3)} \end{array} \right), \\
Z &= \left( \begin{array}{c}
Z^{(1,1)}_t \\
Z^{(1,2)}_t \\
Z^{(2,1)}_t \\
Z^{(2,2)}_t \\
Z^{(3)}_t 
\end{array} \right), \\
M_{\text{VAR}} &= \left( \begin{array}{c}
M^{(1)}_{\text{VAR}} \\
M^{(2)}_{\text{VAR}} \\
M^{(3)}_{\text{VAR}} \end{array} \right), \\
M_{\text{VECM}} &= \left( \begin{array}{c}
M^{(1)}_{\text{VECM}} \\
M^{(2)}_{\text{VECM}} \\
M^{(3)}_{\text{VECM}} \end{array} \right), \\
\kappa_t &= \left( \begin{array}{c}
\kappa_{t}^{(1,1)} \\
\kappa_{t}^{(1,2)} \\
\kappa_{t}^{(2,1)} \\
\kappa_{t}^{(2,2)} \\
\kappa_{t}^{(3)} \end{array} \right), \\
S_t &= \left( \begin{array}{c}
\kappa_{t}^{(1,1)} - \kappa_{t-1}^{(1,1)} \\
\kappa_{t}^{(1,2)} - \kappa_{t-1}^{(1,2)} \\
\kappa_{t}^{(2,1)} - \kappa_{t-1}^{(2,1)} \\
\kappa_{t}^{(2,2)} - \kappa_{t-1}^{(2,2)} \\
\kappa_{t}^{(3)} - \kappa_{t-1}^{(3)} \end{array} \right), \\
\phi_{\text{RWAR}} &= \left( \begin{array}{c}
1 \\
0 \\
0 \\
\phi^{(i_0+1,2)} \\
0 \\
0 \\
0 \\
\phi^{(i_0+1,2)} \\
0 \\
0 \\
\phi^{(3)} \end{array} \right), \\
\Delta \kappa_t &= \left( \begin{array}{c}
\phi^{(i_0,1)} \\
\phi^{(i_0,2)} \\
\phi^{(i_0+1,1)} \\
\phi^{(i_0+1,2)} \\
\phi^{(3)} \end{array} \right), \\
\alpha &= \left( \begin{array}{c}
\alpha_{(1,1)} \\
\alpha_{(1,2)} \\
\alpha_{(2,1)} \\
\alpha_{(2,2)} \\
0 \\
0 \\
0 \\
0 \end{array} \right) \end{align*}
\]