Nielsen, B. & Nielsen, J. P. (2014). Identification and forecasting in mortality models. The Scientific World Journal, 2014, 347043 - ?. doi: 10.1155/2014/347043

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 347043, 24 pages
http://dx.doi.org/10.1155/2014/347043

Research Article

Identification and Forecasting in Mortality Models

Bent Nielsen[1,2,3] and Jens P. Nielsen[4]

[1] *Department of Economics, University of Oxford, Oxford OX1 2JD, UK*
[2] *Programme on Economic Modelling, INET, University of Oxford, Oxford OX1 2JD, UK*
[3] *Nuffield College, Oxford OX1 1NF, UK*
[4] *Cass Business School, City University London, 106 Bunhill Row, London EC1Y 8TZ, UK*

Correspondence should be addressed to Bent Nielsen; bent.nielsen@nuffield.ox.ac.uk

Received 22 January 2014; Accepted 17 April 2014; Published 2 June 2014

Academic Editor: Montserrat Guillén

Mortality models often have inbuilt identification issues challenging the statistician. The statistician can choose to work with well-defined freely varying parameters, derived as maximal invariants in this paper, or with ad hoc identified parameters which at first glance seem more intuitive, but which can introduce a number of unnecessary challenges. In this paper we describe the methodological advantages from using the maximal invariant parameterisation and we go through the extra methodological challenges a statistician has to deal with when insisting on working with ad hoc identifications. These challenges are broadly similar in frequentist and in Bayesian setups. We also go through a number of examples from the literature where ad hoc identifications have been preferred in the statistical analyses.

1. Introduction

Mortality models are commonly used in a wide range of fields such as actuarial sciences, epidemiology, and sociology. They are often used in important decisions such as how to deal with unisex legislation in the pension industry; see Ornelas et al. [1] and Jarner and Kryger [2]. However, such models do often have inbuilt identification issues stemming from overparametrisation. While identification issues are omnipresent in statistical modelling, this paper focuses on mortality modelling, where estimated parameters are treated as time series and extrapolated to give forecasts of future mortality. The underlying theme of this paper is to provide strategies of avoiding arbitrariness resulting from the identification process. We suggest two ways forward. First, we can reparametrise the model in terms of a freely varying parameter, which therefore has to be of lower dimension than the original parameter. Secondly, we can work with an identified version of the original parameter as long as we keep track of the consequences of the identification choice. That way we ensure that two researchers making different identification choices get the same statistical inferences and forecasts.

A simple example is the age-period model for an age-period array of mortality rates. It is well-known that the levels of the age- and period-effects cannot be determined from the likelihood representing the overparametrisation of the model. When the estimated age- and period-effects are treated as time series and subjected to plotting and extrapolation, then our approach ensures that the statistical analysis is the same for two researchers identifying the above model in two different ways. Whereas this issue is relatively simple for the age-period model, identification becomes more tricky for complicated models such as the age-period-cohort model and the model of Lee and Carter [3], let alone two-sample situations.

Mortality models are built as a combination of age, period, and cohort-effects, but the likelihood only varies with a surjective function of these time effects. The time effects can be divided into two parts. One part that moves the likelihood function and another part which does not induce variation in the likelihood function. We will argue that all inferences and forecasts should be concerned primarily with the part of the parameter that moves the likelihood function. This does not preclude the researcher from working with the time effects, but it gives some limitations on what can be done.

This is important because the motivation and the intuition of mortality models typically originate in the time effects. For instance, in the context of an age-period-cohort model linear trends cannot be identified so time series plots of the time effects need to be invariant to linear trends and extrapolations of time effects must preserve the arbitrary linear trend in the time effects. This applies regardless of whether the identification issue is dealt with in a frequentist manner or by Bayesian methods.

To formalise the discussion slightly return to the age-period example. Denote the predictor for the age-period data array by μ. The age-period model then determines how the predictor μ varies with a vector θ summarising age and period effects. That vector is split into two components ξ and λ so that the predictor only depends on θ through ξ but not on λ which cannot be identified by statistical analysis. In the age-period example ξ could reflect the contrasts and the overall level of the predictor μ, whereas λ reflects the level of the age effect. The more principled solution is then to work exclusively with ξ and simply consider θ as a motivation rather than the objective of the analysis. Another solution is to ad hoc identify λ based on a notion of mathematical convenience or based on a particular purpose given the substantive context.

Once an ad hoc identification of λ is chosen the identification problem appears to go away, because the likelihood analysis can now go through. The reason is that the variation of θ is now reduced to the variation of ξ precisely because λ is fixed. Suppose two researchers choose the same likelihood and the same parametrisation of ξ but different ad hoc identifications $\lambda^\dagger$ and $\lambda^\ddagger$. Which of their conclusions will be the same and which will be different? As the likelihood only depends on ξ the fits of the two researchers will be identical. But differences might arise if the statistical inference or forecasting or any other statistical analysis involves λ in some way.

Indeed, with many extrapolation methods forecasts will be invariant to the choice of λ. But, there will also be extrapolation methods where this is not the case. Examples arise in the age-period-cohort model, where linear trends have to be handled with care.

We will start by analysing linearly parametrised models at a rather general level. We do this with two aspects in mind. First, we need to step back to a point in the analysis before ad hoc identification is made. Secondly, we also want to avoid the discussion of how to choose ξ and λ, which tend to be specific to the mortality model in question. Working at the general level we can focus on the mappings between different parametrisations and the invariance properties coming from these mappings. It is then seen that the parameter ξ arises as a maximal invariant. The general setting also allows the formulation of a series of results discussing different types of ad hoc identification, first in a frequentist fashion and then in a Bayesian fashion.

Subsequently, we will consider the age-period-cohort model in detail, both for one- and two-sample situations. Using the general results it becomes easier to see that a number of popular methods inadvertently include features that are not invariant to ad hoc identification. These include the "intrinsic estimator" advocated by Yang et al. [4], the "mixed model approach" by Yang and Land [5], the Bayesian approach by Berzuini and Clayton [6], and the two-sample analysis by Riebler and Held [7]. Finally, we consider the nonlinearly parametrised model of Lee and Carter [3]. The nonlinearity gives a further complication since the mapping from the time effects to the mortality predictor is nondifferentiable. As it turns out the mortality predictor varies in a smooth space, so the nondifferentiability is avoided by working directly with the mortality predictor instead of the original time effects. Instead, a Lee-Carter application should consider whether a certain matrix has rank of unity or zero. Apart from that the analysis is similar to that of linearly parametrised models. Likewise a theory is given for two-sample situations.

Throughout the paper our concern rests exclusively with the identification problem and the consequences of ad hoc identification for estimation, plots, inference, and forecasting. In practice, important additional concerns are how to choose appropriate models and forecasting methods. We would like to refer to Girosi and King [8], Pitacco et al. [9] for general discussions of these issues, and also to Kuang et al. [10] and Coelho and Nunes [11] for discussions of forecast methods in the light of structural breaks. Instead, the aim of the paper is to present an overall framework that can help streamlining the identification discussion that has appeared in so many papers in so many fields over so many years.

Section 2 of this paper considers standard linear statistical models, which lend themselves to a relative straightforward analysis based on linear algebra. Any ad hoc identification splits the time effect into two components. The first component is an arbitrary component, which is not needed for the identification of the likelihood. The other component is necessary and sufficient to identify the model and hence sufficient for statistical analysis. In Section 3 it is outlined how to analyze the statistical model when the latter component is ad hoc identified. It is argued that this can cause difficulties for estimation, interpretation, and forecast. In Section 4 it is shown that Bayesian analysis shares the same challenges as the frequentist approach. In Sections 5 and 6 we study the two particular examples: the omnipresent age-period-cohort and Lee-Carter mortality models. All proofs are collected in the Appendix.

2. Statistical Models with Linear Parametrisations

In this section we present the identification problem in a linear framework. The problem is solved by analysing the mapping from the original time effect to the predictor which, in turn, leads to standard statistical analysis. In Section 6 we show how these ideas transfer to a nonlinear context. This contrasts with Section 3 in which we illustrate the analytical challenges and inconveniences arising from ad hoc identification.

In Section 2.1 we present the overparametrized linear model for the mortality predictor. The identification problem

is defined in Section 2.2 via the likelihood. In an overparametrized linear model two different parameters might produce the same likelihood. In Section 2.3 we analyze the mapping from the overparametrised parameter to the predictor. This mapping enables us to split the overparametrised parameter into two. One arbitrary parameter and one parameter identify the model without being overparametrised. This latter parameter is shown to be a maximal invariant parameter. In Section 2.4 it is demonstrated how any statistical analysis can be based on this maximal invariant parameter alone. In particular we comment that visual data representations, hypothesis testing, and forecasting are simple and well defined. This in turn leads to standard statistical analysis.

The analysis of the linearly parametrised involves projections on linear or affine spaces and on their orthogonal complements. It is therefore convenient to introduce the following notation. A matrix m has full column rank if $m'm$ is invertible. In this case the orthogonal complement $m_\perp$ is a matrix so $m_\perp' m = 0$ and $(m, m_\perp)$ is invertible. Thus, when m itself is invertible then $m_\perp$ is the empty matrix. It is not difficult to calculate $m_\perp$ in practise, an explicit construction of $m_\perp$ follows from a singular value decomposition of mm', choosing $m_\perp$ as the eigenvectors associated with the zero eigenvalues. Moreover, let $\overline{m} = m(m'm)^{-1}$ so that $\overline{m}'m$ is the identity matrix, while $\overline{m}_\perp = m_\perp(m_\perp' m_\perp)^{-1}$.

2.1. The Model. Think of the time effect θ as our preferred intuitive, but unidentified parameter, and think of the predictor μ as some function of θ specifying the model at hand. In a Poisson type model, where the mean specifies the distribution, μ could be the log of that mean. Such Poisson models are omnipresent in mortality models. We will often think of θ as containing some time effects. Often forecasting is carried out simply by isolating and extrapolating such a time effect.

Consider a data vector Y of dimension n. This could, for instance, be the vector consisting of the stacked mortality rates for a rectangular age-period array of dimension $I \times J$ in which case $n = IJ$. The statistical model for Y could be a generalized linear model. This involves an appropriately chosen distribution and a link function, which links the expected mortality rate to an n-dimensional predictor, which is denoted by μ. Taken together this defines a likelihood function $\mathsf{L}(\mu; Y)$.

The model for the predictor μ is constructed in terms of, for instance, age, period, and cohort time effects. These time effects are summarized in a vector θ, which is of dimension $q < n$. Therefore μ is a surjective function of θ. For the moment the specification of the predictor is assumed linear so that

$$\mu = D\theta \quad \text{for } \theta \in \Theta = \mathbb{R}^q, \tag{1}$$

for some design matrix $D \in \mathbb{R}^{n\times q}$. We refer to this specification as the mortality model, while the space Θ is the time effect space. The time effect space is chosen as an unrestricted real space in accordance with the starting point of most mortality analyses.

The parameter space for the likelihood function and therefore for the statistical model is given by the range of variation for the predictor μ; that is,

$$M = \left(\mu \in \mathbb{R}^n : \mu = D\theta \text{ for } \theta \in \Theta = \mathbb{R}^q\right). \tag{2}$$

The likelihood function is assumed uniquely identified on this space in the sense that for all pairs of predictors so $\mu^\dagger \neq \mu^\ddagger$ then the likelihood of $\mu^\dagger$, $\mu^\ddagger$ differ; that is,

$$\mathsf{L}\left(\mu^\dagger; Y\right) \neq \mathsf{L}\left(\mu^\ddagger; Y\right), \tag{3}$$

for Y in a set with positive probability.

2.2. The Identification Problem. The identification problem of mortality models arises when the mapping from the time effect space Θ to the parameter space M is surjective but not injective. With a linear parametrisation this arises when the design matrix D has reduced column rank $p < q$ so $D'D$ is singular. In this situation there exists time effects $\theta^\dagger \neq \theta^\ddagger$ with the same likelihood:

$$\mathsf{L}\left(D\theta^\dagger; Y\right) = \mathsf{L}\left(D\theta^\ddagger; Y\right), \tag{4}$$

for all data Y. Then the time effect space Θ is not useful as parameter space for the statistical model.

2.3. Analysing the Mapping $\theta \mapsto \mu$. When analysing the mapping from our intuitively preferred parametrisation θ into the linear predictor μ, we will be able to rewrite θ as a sum of two components: one is a function of the predictor and the other is the arbitrary part varying with θ, but not with the predictor. We provide two methods for analysis.

The first method is to find a basis $X \in \mathbb{R}^{n\times p}$ with full column rank p for the design D. The design matrix of the mortality model can then be expressed as $D = XA'$ for some matrix $A \in \mathbb{R}^{q\times p}$ with full column rank p. Introduce a new p-dimensional parameter:

$$\xi = A'\theta. \tag{5}$$

The parameter space M can then be written more parsimoniously as

$$M = \left(\mu \in \mathbb{R}^n : \mu = X\xi \text{ for } \xi \in \mathbb{R}^p\right). \tag{6}$$

The mapping from ξ to μ is bijective, so the statistical model can just as well be parametrised in terms of $\xi \in \Xi = \mathbb{R}^p$.

Alternatively, the identification problem can be expressed through an invariance argument. This argument relates to the parameterization but resembles the classical invariance argument for reduction of data; see Cox and Hinkley [12, page 157]. With a linear parametrisation the argument involves the orthogonal complement to the matrix A. That is a matrix $A_\perp \in \mathbb{R}^{q\times(q-p)}$ which has the properties that $A_\perp' A = 0$ and that $(A, A_\perp)$ is invertible. The mortality model (1) is defined by the mapping

$$\theta \longmapsto \mu = D\theta = XA'\theta, \tag{7}$$

from $\Theta = \mathbb{R}^q$ to M. This mapping is surjective in that two different values of θ may result in the same μ and therefore the same likelihood. These equivalence classes in the time effect space can be described by the group of transformations

$$g : \theta \longmapsto \theta + A_\perp \zeta, \tag{8}$$

acting on Θ for arbitrary $\zeta \in \mathbb{R}^{q-p}$. Indeed, it holds that θ and $g(\theta)$ will result in the same μ. The mapping (7) is therefore invariant to the group g. We will argue that the parameter $\xi = A'\theta$ is a maximal invariant to the group g acting on Θ, which provides a link with (6). It has to be argued that for any $\theta^\dagger$, $\theta^\ddagger$ so that $\xi^\dagger = A'\theta^\dagger$ equals $\xi^\ddagger = A'\theta^\ddagger$ then $\theta^\ddagger = g(\theta^\dagger)$, see Cox and Hinkley [12, page 159]. For this argument use the orthogonal projection identity to write

$$\theta = A\left(A'A\right)^{-1}\xi + A_\perp\left(A'_\perp A_\perp\right)^{-1}\varphi; \tag{9}$$

for unique $\xi = A'\theta$ and $\varphi = A'_\perp\theta$. Thus, if $A'\theta^\ddagger = A'\theta^\dagger$ then $\theta^\ddagger = g(\theta^\dagger)$ with $\zeta = \varphi^\ddagger - \varphi^\dagger = A'_\perp(\theta^\ddagger - \theta^\dagger)$.

In applications it can be difficult to find a basis X for the design D. It can be easier to find a group g and hence $A_\perp$ and then use this information to construct A and a candidate basis $X = D\overline{A}$, noting that $D = XA'$. This argument leaves it to be proven that X is a basis, or equivalently, that the suggested group g actually describes the equivalence classes of the mapping from θ to μ.

It is useful to note that in the choices of X, A only the spaces spanned by them are unique since $XA' = Xmm^{-1}A'$ for any invertible $m \in \mathbb{R}^{p\times p}$. Likewise, the maximal invariant ξ is only unique up to bijective transformations. This lack of uniqueness has no impact on the analysis of the likelihood albeit it influences interpretations.

2.4. Statistical Analysis Using the Maximal Invariant Parameter. The statistical model parametrised with the maximal invariant parameter ξ can be analysed by standard statistical techniques. This contrasts to a range of problems that arise when working with an ad hoc identified time effect θ. In the following the relatively simple standard statistical analysis of the model parametrised by ξ is discussed with respect to likelihood theory, interpretation, plots, hypothesis testing, forecasting, and Bayesian analysis. In Sections 3 and 4 we give an overview of the much more complicated theory underpinning models parametrised by the ad hoc identified time effect θ. Age-period-cohort examples follow in Section 5.

2.4.1. Exponential Family Theory. Suppose the likelihood is drawn from a generalized linear model based on an exponential family. Then the model is actually a regular exponential family where the maximal invariant parameter ξ is the canonical parameter since it is freely varying in a real space; see Barndorff-Nielsen [13, page 116]. This opens up for a wealth of convenient statistical properties such as a likelihood equation with a simple expression and explicit conditions for a unique solution. In contrast, ad hoc identified parameters are based on an injective mapping of the canonical parameter ξ into θ; see Sections 3.1 and 3.2. It is then more difficult to fully exploit the exponential family theory.

2.4.2. Interpretation and Plots. The maximal invariant parameter ξ varies freely in $\mathbb{R}^p$. It can therefore be interpreted as the parameter of any standard statistical model. Since ξ is freely varying the coordinates of ξ can be interpreted independently. When θ is a collection of time effects then ξ can be organised as a collection of time series. Since the coordinates of ξ are freely varying the time series plots of the components of ξ have the usual interpretation of time series. In contrast, ad hoc identified estimators are constrained to a p-dimensional subspace Θ_λ of $\Theta = \mathbb{R}^q$, which is often affine but can be more complicated. A consequence is that plots are complicated to evaluate; see Section 3.4.1.

2.4.3. Hypothesis Testing. Hypotheses are easily formulated and analysed when using the maximal invariant parametrisation. An affine hypothesis that restricts ξ to vary in a p_H-dimensional affine subspace can be formulated as $H'\xi = \eta$ for known matrices $H \in \mathbb{R}^{p\times(p-p_H)}$, $\eta \in \mathbb{R}^{p-p_H}$. This implies a restriction on the predictor $\mu = X\xi$ of (6). Form the orthogonal complement $H_\perp$ and recall the orthogonal projection identity $I_n = \overline{H}H' + H_\perp\overline{H}'_\perp$ so that $\mu = X\overline{H}H'\xi + XH_\perp\overline{H}'_\perp\xi$. Introduce a p_H-dimensional parameter $\varphi = \overline{H}'_\perp\xi$, a design matrix $X_H = XH_\perp$, and an offset $Z_H = X\overline{H}\eta$. The restricted parameter space is

$$M_H = \left(\mu \in \mathbb{R}^n : \mu = X_H\varphi + Z_H \text{ for } \varphi \in \mathbb{R}^{p_H}\right). \tag{10}$$

In an exponential family context both the unrestricted model and the restricted model form regular exponential families. A variety of nice properties then follow for the estimators and the test statistics from the exponential family theory. Examples are given in Sections 5.3 and 5.5.3. In contrast, the hypothesis derived from restrictions on ad hoc identified parameters and the resulting degrees of freedom are complicated to analyse; see Section 3.4.2.

2.4.4. Forecasting. Most often the objective of a mortality study is to forecast the future mortality. In the linear context, $\mu = X\xi$, this is done by extending the design X and by extrapolating ξ.

It is usually easy to extend the design X into the forecast horizon. This involves the construction of a triangular block matrix with an appropriate number of extra rows corresponding to the data over the forecast horizon as well as extra columns representing the extra parameters that would be needed:

$$X^h = \begin{pmatrix} X & 0 \\ X_1^h & X_2^h \end{pmatrix}. \tag{11}$$

Extrapolating ξ into a vector $\widetilde{\xi}$ then gives the forecast

$$\widetilde{\mu} = \left(X_1^h, X_2^h\right)\begin{pmatrix}\xi \\ \widetilde{\xi}\end{pmatrix}. \tag{12}$$

The extrapolation of the parameter ξ can be done as follows. The estimated parameter, or part of it, can be thought of as a time series. Any forecast techniques from the time series literature applied directly to ξ can be used, subject to the usual contextual considerations.

Ad hoc identified time effects can be extrapolated in a similar way; see Section 3.4.3. This may, however, result in avoidable arbitrary effects in the forecast. Necessary and sufficient conditions for this eventuality are given for age-period-cohort models in Section 5.4.3. The practical examples are mainly Bayesian in nature and are discussed next.

2.4.5. Bayesian Analysis. The introduction of the canonical parameter shows that the likelihood, in Bayesian notation, is of the form $p(y \mid \theta) = p(y \mid \xi)$ where ξ is freely varying. A purist Bayesian analysis can simply introduce a prior on the canonical parameter, $p(\xi)$. This is updated in a straight forward way, resulting in the posterior $p(\xi \mid y) = p(y \mid \xi)p(\xi)/p(y)$.

In contrast, introducing a prior on ad hoc identified parameters gives various difficulties. Only parts of the prior are updated by the likelihood, so that it becomes unclear which information arises from the data and which information arises from the ad hoc identification. Moreover, avoidable arbitrariness is introduced in the forecast; see Section 4. Introduction of hyperparameters exacerbates the issue. Examples are given in Sections 5.4.4, 5.5.2, and 6.1.6.

3. Working with the Time Effects

In Section 2 we considered the situations where estimation, hypothesis testing a hypothesis, or forecasting is carried out using the canonical parameter. However, there might be situations, where the original time effect parametrisation is preferred, perhaps because it is felt that this parametrisation is particularly helpful in guiding the intuition. This requires ad hoc identification of the time effect . In this section we will guide the considerations a statistician that has to go through when insisting on an analysis based on some nonunique parametrisations. As in Section 2 we focus on linearly parametrised models. Specific examples follow in Sections 5 and 6.

In Section 3.1 ad hoc identification is defined. As an example we consider a least squares estimation problem with collinear regressors in Section 3.2. For the age-period-cohort model reviewed in Section 5 it is common to ad hoc identify in two steps: first identifying levels then the linear trends. We consider such two-step ad hoc identification in Section 3.3. The consequence of ad hoc identification is considered in Section 3.4. Indeed, when forecasting the time effect, we do not want the forecast to depend on the identification scheme. The same applies to graphical visualisation of our data, where the eye may extract patterns that depend on the identification scheme. Likewise, confusion may arise when formulating a hypothesis directly on the time effect parameters.

3.1. Ad Hoc Identification. In this section the time effect parametrisation is considered. An identification scheme has to be introduced when working with the time effects. This may rest on mathematical convenience or it may be chosen for a particular purpose given the substantive context. We therefore call it ad hoc identification. Here we consider a simple identification scheme but turn to a more common two-step identification scheme in Section 3.3.

Once the canonical parameter ξ has been estimated there is often a wish to return to the original time effect θ. The two are linked through the surjective mapping

$$\theta \longmapsto \xi = A'\theta, \tag{13}$$

from $\Theta = \mathbb{R}^q$ to $\Xi = \mathbb{R}^p$. Indeed, since ξ is constructed as a function of θ the notation for ξ is often chosen to reflect θ. The canonical parameter ξ does, however, only give partial information about θ. The remaining part, say λ, of θ will have to be chosen by the researcher and combined with ξ.

A linear ad hoc identification of θ comes about by the researcher choosing a constraint

$$L'\theta = \lambda \tag{14}$$

for some known $\lambda \in \mathbb{R}^{q-p}$ and some matrix $L \in \mathbb{R}^{q\times(q-p)}$ chosen so the square matrix (A, L) is invertible. The time effect space Θ is now reduced to an affine subspace

$$\Theta_\lambda = \left(\theta_\lambda \in \Theta : L'\theta_\lambda = \lambda\right). \tag{15}$$

Given θ we can find ξ, λ through (13) and (14) as $(\xi', \lambda')' = (A, L)'\theta$. At the same time, given values of ξ, λ and the invertibility of (A, L), the ad hoc identified time effect is found through

$$\theta_\lambda = \begin{pmatrix} A' \\ L' \end{pmatrix}^{-1} \begin{pmatrix} \xi \\ \lambda \end{pmatrix} = L_\perp\left(A'L_\perp\right)^{-1}\xi + A_\perp\left(L'A_\perp\right)^{-1}\lambda. \tag{16}$$

In this notation a subindex λ is introduced to avoid confusion with the time effect θ in the original mortality model. Indeed, there are now four different parameters in play, namely, the original time effect $\theta \in \Theta$, the predictor $\mu \in M$, the maximal invariant parameter $\xi \in \Xi$ and the ad hoc identified time effect $\theta_\lambda \in \Theta_\lambda$, each of which has a different interpretation. The mapping from θ to each of μ, ξ, and θ_λ is surjective, while there are bijective mappings between the latter three. The interpretations of the time effect θ and the canonical parameter ξ will inevitably be different. For a start they have different dimensions. Endowing the spaces with Euclidean norms shows that distances in the two spaces Θ and Ξ will be judged differently. The time effect θ and the ad hoc identified time effect θ_λ will similarly have different interpretations. Although they have the same dimensions the Euclidean norms on Θ and Θ_λ will be rather different. Confusion may arise in the interpretation of a mortality analysis if there is no clear distinction between θ and θ_λ. In addition an unnecessary arbitrariness may arise when making inference on θ_λ or extrapolating $\theta \in \Theta_\lambda$. We will return to these issues in Section 3.4.

It is perhaps interesting to note that despite the linear parametrisation the ad hoc identification need not be done

in a linear fashion as in (14). Indeed it is common for Poisson models with a log link to ad hoc identify θ through the original multiplicative scale. That means that the ad hoc identification is done nonlinearly through

$$\xi = A'\theta_\lambda, \quad \lambda = f(\theta_\lambda). \tag{17}$$

The fit of the model is unaffected by the ad hoc identification. Indeed the fit is measured in terms of the estimate of the predictor $\mu = D\theta_\lambda$ where $D = XA'$. Since the identification is made so $\xi = A'\theta_\lambda$; the estimated predictor reduces to

$$\widehat{\mu} = D\widehat{\theta}_\lambda = XA'\widehat{\theta}_\lambda = X\widehat{\xi}, \tag{18}$$

regardless of the choice of ad hoc identification.

3.2. A Least Squares Example. As an illustration of estimation in the presence of ad hoc identification consider a normal likelihood. Different, but equivalent, expressions can be found depending on the parametrisation. The likelihood of the predictor μ is

$$L\left(\mu, \sigma^2; Y\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(Y-\mu)'(Y-\mu)\right\} \quad \text{for } \mu \in M, \quad \sigma^2 > 0. \tag{19}$$

Rewriting it in terms of the canonical parameter it is

$$L\left(\xi, \sigma^2; Y\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(Y-X\xi)'(Y-X\xi)\right\} \quad \text{for } \xi \in \Xi = \mathbb{R}^p, \quad \sigma^2 > 0, \tag{20}$$

while introducing the time effect parameter gives

$$L\left(\theta, \sigma^2; Y\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(Y-XA'\theta\right)'\left(Y-XA'\theta\right)\right\} \quad \text{for } \theta \in \Theta = \mathbb{R}^q, \quad \sigma^2 > 0. \tag{21}$$

The likelihood (20) of the canonical parameter ξ can be analysed by the least squares method since the design X has full column rank. The maximum likelihood estimator for ξ and the predictor for the data are

$$\widehat{\xi} = \left(X'X\right)^{-1}X'Y, \qquad \widehat{Y} = X\widehat{\xi} = X\left(X'X\right)^{-1}X'Y. \tag{22}$$

Along with the residual variance this is all the information that is given by the likelihood.

The likelihood (21) of the time effect θ only depends on θ through $\xi = A'\theta$. The lack of identification means that the maximum likelihood estimator for θ has an arbitrary element, so that it is a set valued estimator. Based on (16) this can be expressed by

$$\widehat{\Theta}_\lambda = L_\perp\left(A'L_\perp\right)^{-1}\widehat{\xi} + A_\perp\left(L'A_\perp\right)^{-1}\lambda \quad \text{where } \widehat{\theta}_\lambda \in \Theta_\lambda \subset \Theta, \tag{23}$$

for any L so (A, L) is invertible and for any $\lambda \in \mathbb{R}^{q-p}$. The fit, however, remains the same and (18) becomes

$$\widehat{\mu} = D\widehat{\theta}_\lambda = XA'\left\{L_\perp\left(A'L_\perp\right)^{-1}\widehat{\xi} + A_\perp\left(L'A_\perp\right)^{-1}\lambda\right\} = X\widehat{\xi} = \widehat{Y}. \tag{24}$$

In order to compute actual estimates then L, λ have to be chosen, which amounts to ad hoc identification. For instance, with the ad hoc identifying restrictions $L = A_\perp$ and $\lambda = 0$ then $\widehat{\theta}_\lambda$ can be thought of as the least squares estimator of Y on D using the Moore-Penrose generalised inverse for the singular matrix $D'D$; see Searle [14, page 212]. See Section 5.4.1 for an example.

3.3. Step-Wise Identification. It is common to ad hoc identify parameter in a step-wise fashion. In the first step the time effect parameter is only partially constrained. The full identification then follows in a second step. An example is given in Section 5.4.1 for an age-period-cohort model in which the levels of the time effects are constrained in the first step leaving the ad hoc identification of the linear trends to the second step.

The first step constraints are affine of the type

$$C'\theta_C = \psi, \tag{25}$$

for known matrices $C \in \mathbb{R}^{q\times(q-q_C)}$, $\psi \in \mathbb{R}^{q-q_C}$. The constrained time effect space is then

$$\Theta_C = \left(\theta_C \in \Theta : C'\theta_C = \psi\right). \tag{26}$$

Thereby the q-dimensional time effect space Θ is reduced to a q_C-dimensional variation. The properties of this partially ad hoc identified parameter space depends on the rank of the matrix (A, C). If the number of constraints, $q - q_C$, is at most equal to the number of unidentified components $q - p$, it is possible that (A, C) has full column rank. In that case the constraint implies a partial ad hoc identification without constraining the parameter space M of the statistical model. This is shown in Theorem 1; see also Section 5.4.1 for an example, while the proof is given in the Appendix. When (A, C) has reduced rank the parameter space M is also constrained; see Section 3.4.2 for a discussion.

Theorem 1. *Suppose (A, C) has full column rank. Then the matrix $m = A'_\perp C \in \mathbb{R}^{(q-p)\times(q-q_C)}$ has full column rank and the constraint (25) does not constrain the canonical parameter ξ and the predictor μ. Hence, the predictor space remains of the form (2). The equivalence classes in Θ_C under the mapping $\theta \mapsto \mu = XA'\theta$ are given by the group*

$$g_C : \theta \longmapsto \theta + A_\perp m_\perp \zeta, \tag{27}$$

for arbitrary $\zeta \in \mathbb{R}^{q_C-p}$ where $m_\perp \in \mathbb{R}^{(q-p)\times(q_C-p)}$ is the orthogonal complement of m. The maximal invariant remains $\xi = A'\theta$.

The partial ad hoc identification by (25) implies that any time series analysis of the time effects has to happen

relative to the constrained space Θ_C rather than the space Θ. This is awkward as discussed in Section 3.4 below. It is also considerably more complicated than working with the freely varying canonical parameter ξ; see Section 2.4.2.

3.4. Consequences of Ad Hoc Identification. In the following we will look closer at the consequences of working with the ad hoc identified time effect parameter θ in the context of a linear mortality model of the form $\mu = D\theta$. We consider the consequences for plotting, hypothesis testing, and forecasting.

3.4.1. Plots of Time Effects. In the mortality model (1) the time effect θ is the concatenation of age, period, and cohort effects. It seems natural to think of these individual time effects as time series and to plot them against time. As the time effect θ varies in the unrestricted space $\Theta = \mathbb{R}^q$ this maps the q-vector into unrestricted time series.

Estimates of the time effects are constructed by combining an estimate of ξ with an ad hoc chosen value for $\lambda = L'\theta$, see (14). The resulting estimate $\widehat{\theta}_\lambda$ is therefore constrained to the space $\Theta_\lambda \subset \Theta$. The interpretation of the estimate $\widehat{\theta}_\lambda$ is therefore different from the interpretation of the original time effect θ. Distances on the spaces Θ and Θ_λ are judged differently and the variability of $\widehat{\theta}_\lambda$ is deduced exclusively from $\widehat{\xi}$ through (16). The time series components of $\widehat{\theta}_\lambda$ are now restricted through $\lambda = L'\theta_\lambda$. Plots of the $\widehat{\theta}_\lambda$-time series are therefore interpreted differently from the imagined plots of the original θ-time series and from the plots of the maximal invariant parameter ξ discussed in Section 2.4.2. Indeed, if one were to analyse the estimated $\widehat{\theta}_\lambda$-time series statistically the linear constraint should be taken into account. This is a bit complicated as illustrated below, but it is the consequence of working with the ad hoc identified parameter θ_λ rather than the canonical parameter ξ.

Attempts to give intrinsic meaning to λ will be specific to the index set for the data set at hand. For instance, the requirement that the age effect should be zero on average does not carry over when looking at a subsample or when forecasting. It is not obvious that such an ad hoc identification is any more or less arbitrary than saying that, for instance, the first or the last age effect should have a particular value.

Adding confidence bands to a plot of $\widehat{\theta}_\lambda$ is in itself not difficult. If $\widehat{\xi}$ is asymptotically normal with mean ξ and variance Σ, then $\widehat{\theta}_\lambda$ is asymptotically normal with mean θ_λ and variance $L_\perp(A'L_\perp)^{-1}\Sigma(L'_\perp A)^{-1}L'_\perp$. This is a normal distribution on the space Θ_λ. The interpretation of these standard errors will therefore be similar to that of $\widehat{\theta}_\lambda$ itself.

Finally, it may be of interest to analyse the estimated $\widehat{\theta}_\lambda$-time series statistically. Denote this time series by x_λ. Its sample space is now Θ_λ. A statistical model on Θ_λ can be built as follows. The starting point could be a time series model for unrestricted variables x on the sample space Θ. This gives a joint density for $x \in \Theta$, which can be reduced by marginalisation to a density for $x_\lambda \in \Theta_\lambda$. Whether one is working with the unrestricted model for $x \in \Theta$ or the restricted model for $x_\lambda \in \Theta_\lambda$ inferences that are invariant to g must be based on those statistics of x or x_λ that are invariant to g. Thus, inferences must be based on the maximal invariant under g. For a general overview of invariant reduction see Cox and Hinkley [12, page 175f], whereas Nielsen [15] gives the argument in some detail for an autoregression with a linear trend.

3.4.2. Hypothesis Testing. Having formulated the model in terms of time effects it may be of interest to test the hypothesis that one of these time effects is absent. No identification issues arise when the hypothesis is formulated as a restriction on the canonical parameter ξ as discussed in Section 2.4.3. But one has to be careful when formulating hypotheses in terms of the original time effect. See Sections 5.4.5, 5.5.3, and 5.5.4 for examples.

Affine hypotheses on the time effect are of the form

$$R'\theta_R = \rho, \tag{28}$$

for known matrices $R \in \mathbb{R}^{q\times(q-q_R)}$, $\rho \in \mathbb{R}^{q-q_R}$. The constrained time effect space is then

$$\Theta_R = \left(\theta_R \in \mathbb{R}^q : R'\theta_R = \rho\right). \tag{29}$$

To see how the restriction (28) restricts the predictor space $M \subset \mathbb{R}^n$ recall that the predictor μ only depends on θ through $\xi = A'\theta$. Thus, the analysis of the restriction (28) depends on the interplay between the matrices A, R. Theorem A.3 in Appendix A.3 gives a general result to that effect. It shows that the hypothesis (28) restricts the predictor space M to a p_R-dimensional affine subspace of $\mathbb{R}^n$ in so far as it restricts the canonical parameter ξ. In particular, the degrees of freedom of the hypothesis, $p - p_R$, may in general be different from the dimension reduction of the time effect parameter, $q - q_R$. When this is the case the restriction (28) has an element of ad hoc identifying the time effect.

3.4.3. Forecasts. Forecasts can be made by extrapolating the ad hoc identified time effects θ_λ. Two researchers choosing different ad hoc identification schemes, but otherwise making the same analysis, may make different forecasts. This can be avoided if the extrapolation method is chosen with some care.

Following the linear approach outlined in Section 2.4.4 the predictor $\mu = D\theta = XA'\theta$ is forecasted by extending the design D into

$$D^h = \begin{pmatrix} D & 0 \\ D_1^h & D_2^h \end{pmatrix}. \tag{30}$$

Extrapolating the ad hoc identified θ_λ into a vector $(\theta'_\lambda, \widetilde{\theta}'_\lambda)'$ then gives the forecast

$$\widetilde{\mu} = \left(D_1^h, D_2^h\right)\begin{pmatrix}\theta_\lambda \\ \widetilde{\theta}_\lambda\end{pmatrix} = D_1^h\theta_\lambda + D_2^h\widetilde{\theta}_\lambda. \tag{31}$$

Often both components $D_1^h\theta_\lambda$ and $D_2^h\widetilde{\theta}_\lambda$ depend on the ad hoc identification. Nonetheless, these dependencies of

ad hoc identification may cancel each other so that the overall forecast $\tilde{\mu}$ is invariant to the ad hoc identification. Such invariance would seem desirable in most applications unless there is strong substantial reason for the ad hoc identification scheme. Necessary and sufficient conditions for invariance are presented for the age-period-cohort model in Section 5.4.3 and for a nonlinear model in Section 6.1.5.

In contrast, these considerations are redundant when working with the canonical parameter, ξ; see Section 2.4.4.

4. Bayesian Models and Random Effects Models

Mortality analysis is often carried out using either Bayesian methods or random effects methods. The mortality model is then altered through the introduction of a prior distribution on the parameters. One might think that the identification problems become less of an issue or even disappear. This is not the case since the Bayesian method and the random effects method is based on the mortality likelihood which only depends on the time effect θ through the maximal invariant parameter ξ. Thus, the identification challenges remain. The issue is that a prior on the unidentified part, say λ, of the time effect amounts to an ad hoc identification. Indeed, the conditional prior of λ given ξ is not updated by the mortality likelihood. A main difference is that a maximum likelihood analysis of the original mortality likelihood usually prompts the researcher when there is an identification issue, whereas both Bayesian methods and random effects methods allow computations to go through despite an identification issue.

In Section 4.1 it is seen that introduction of a conditional prior on λ given ξ is the Bayesian analogue of ad hoc identification. This leads to the same type of forecasting challenges as in the frequentist settings as is seen in Section 4.2. In Section 4.3 we show how the Bayesian identification issues transfer to random effects models.

4.1. Bayesian Estimation. For Bayesian and random effects models we formulate a likelihood and a prior. Thus, consider a likelihood $p(y \mid \theta) = L(\theta; y)$. Replacing θ by ξ, λ the identification problem implies that

$$p(y \mid \xi, \lambda) = p(y \mid \xi) \quad \text{for all outcomes } y. \tag{32}$$

The prior on θ is factorised as $p(\theta) = p(\xi, \lambda) = p(\xi)p(\lambda \mid \xi)$. In the case of Bayesian estimation the following result emerges.

Theorem 2. *Suppose the likelihood satisfies* (32). *Then*

(i) *the predictive distribution does not depend on the conditional prior for λ:*

$$p(y) = \int p(y \mid \xi)\, p(\xi)\, d\xi; \tag{33}$$

(ii) *the posterior satisfies*

$$p(\xi \mid y) = \frac{p(y \mid \xi)\, p(\xi)}{p(y)}, \qquad p(\lambda \mid \xi, y) = p(\lambda \mid \xi); \tag{34}$$

(iii) *the posterior means satisfy*

$$E(\xi \mid y) = \int \xi p(\xi \mid y)\, d\xi,$$
$$E(\lambda \mid \xi, y) = E(\lambda \mid \xi), \tag{35}$$
$$E(\lambda \mid y) = \int E(\lambda \mid \xi)\, p(\xi \mid y)\, d\xi.$$

Theorem 2 shows that it suffices to give a prior to ξ and ignore λ as advocated in Section 2.4.5. Indeed the conditional prior for λ given ξ is not updated. Theorem 2 appears to be well-known; see Poirier [16, Proposition 2] or Smith [17, Section B].

Due to Theorem 2 the Bayesian analyst faces the complications outlined in Section 3.4. Indeed, suppose that two Bayesian researchers choose the same likelihood $p(x \mid \xi, \lambda) = p(x \mid \xi)$ and the same prior $p(\xi)$ for ξ, but different conditional priors for λ given ξ. Their marginal distributions for the data are identical, but any inferences regarding interpretation or forecasting will differ in so far as they involve the unidentified parameter λ. A Bayesian researcher should therefore be cautious with inference related to λ. There will of course be situations where the prior knowledge of λ given ξ is found to be of substantive relevance. In such situations it seems more fruitful to change the likelihood to include that information.

4.2. Forecasting. Bayesian forecasts involve integrating an extrapolative distribution. This can be done in two ways, either working exclusively with the identified, maximal invariant parameter ξ as in Section 2.4.4, or working with the time effect $\theta = (\xi, \lambda)$ as in Section 3.4.3.

4.2.1. Forecasting Using the Maximal Invariant Parameter. Consider first the case where only the maximal invariant parameter ξ is used. In that case the forecast is computed by sampling from the posterior $p(\xi \mid y)$ and then extrapolating $\tilde{\mu}$ using the sampled value ξ using some extrapolative methods, say $p(\tilde{\mu} \mid \xi, y)$. In combination this gives the forecast

$$p(\tilde{\mu} \mid y) = \int p(\tilde{\mu} \mid \xi, y)\, p(\xi \mid y)\, d\xi. \tag{36}$$

4.2.2. Forecasting Using the Ad Hoc Identified Time Effect. Consider now forecasts involving the full time effect $\theta = (\xi, \lambda)$. Theorem 2(ii) shows that the posterior satisfies $p(\theta \mid y) = p(\xi \mid y)p(\lambda \mid \xi)$. The distribution forecast with extrapolation $p(\tilde{\mu} \mid \xi, \lambda, y)$ is then

$$p(\tilde{\mu} \mid y) = \iint p(\tilde{\mu} \mid \xi, \lambda, y)\, p(\xi \mid y)\, p(\lambda \mid \xi)\, d\lambda\, d\xi. \tag{37}$$

The concern is now as follows. Suppose a second researcher chooses the same extrapolative method, likelihood, and prior for ξ, but different conditional priors $p^{\dagger}(\lambda \mid \xi)$. In general, this will result in a different distribution forecast:

$$p^{\dagger}(\tilde{\mu} \mid y) = \iint p(\tilde{\mu} \mid \xi, \lambda, y)\, p(\xi \mid y)\, p^{\dagger}(\lambda \mid \xi)\, d\lambda\, d\xi. \tag{38}$$

The question is then under which conditions will $p(\tilde{\mu} \mid y) = p^{\dagger}(\tilde{\mu} \mid y)$ so that the distribution forecasts are invariant to the choice of conditional prior for λ given ξ? A sufficient condition is that the extrapolation method does not depend on λ so

$$p(\tilde{\mu} \mid \xi, \lambda, y) = p(\tilde{\mu} \mid \xi, y). \tag{39}$$

Condition (39) could alternatively be expressed as requiring that the forecast $p(\tilde{\mu} \mid \theta, y) = p(\tilde{\mu} \mid \xi, \lambda, y)$ is invariant to the group g acting on the time effect space Θ so that $p(\tilde{\mu} \mid \xi, \lambda, y) = p\{\tilde{\mu} \mid \xi, g(\lambda), y\}$.

Theorem 3. *Suppose that the likelihood satisfies* (32) *and the priors are probabilities. If the extrapolative distribution does not depend on* λ *so* (39) *holds; then the forecast distribution* $p(\tilde{\mu} \mid y)$ *computed in* (37) *is invariant to the choice of conditional prior for* λ *given* ξ. *The forecast then reduces to* (36).

To summarise, the identification issues surrounding Bayesian analysis are similar to those outlined in the previous sections. Examples of the problems that can arise are discussed in Sections 5.4.4, 5.5.2, and 6.1.6. There are two solutions to the identification problem. The first is only to formulate a prior on ξ; see Section 2.4.5. Incidentally, this is what Bernardo and Smith [18, page 218] do in their discussion of the two-way analysis of variance, albeit without linking it to the considerations of Smith [17]. The prior $p(\xi)$ can of course be constructed by formulating a prior on θ and then reduce it to a prior on ξ by marginalisation so $p(\xi) = \int p(\xi, \lambda) d\lambda$. The other solution is to work with a prior on θ but avoid those parts of the posterior that depend on λ.

4.3. Random Effects Models. It is common to combine mortality models with a random effects approach, which effectively forms a new model. An example is given in Section 5.4.6. We consider the consequence of the lack of identification.

The random effect models are typically constructed as follows. Suppose the density of the data y given the time effects $\theta = (\xi, \lambda)$ is of the form $p(y \mid \xi, \lambda) = p(y \mid \xi)$ as before; see (32). A prior $p(\theta \mid \psi)$ is chosen that now depends on a parameter ψ. The prior can be decomposed as $p(\theta \mid \psi) = p(\xi \mid \psi) p(\lambda \mid \xi, \psi)$. Theorem 2 implies that the density of the data y given ψ is

$$p(y \mid \psi) = \int p(y \mid \xi)\, p(\xi \mid \psi)\, d\xi. \tag{40}$$

This in turn is used to form the random effects likelihood of ψ as

$$L_{\mathrm{RE}}(\psi \mid y) = p(y \mid \psi). \tag{41}$$

This, effectively, defines a new model. The random effects likelihood only depends on the prior $p(\theta \mid \psi)$ through $p(\xi \mid \psi)$. Two researchers choosing the same prior $p(\xi \mid \psi)$ but different conditional priors $p(\lambda \mid \xi, \psi)$ will then get the same random effects likelihood and the same maximum likelihood estimator $\widehat{\psi}$.

In mortality modelling it is common to go one step further and estimate the time effects θ through the mean of the posterior $p(\theta \mid y, \psi)$ evaluated at $\psi = \widehat{\psi}$. Then the identification problem may show up. Theorem 2 shows that

$$p(\xi \mid \widehat{\psi}, y) = \frac{p(y \mid \xi)\, p(\xi \mid \widehat{\psi})}{p(y \mid \widehat{\psi})}, \qquad p(\lambda \mid \xi, \widehat{\psi}, y) = p(\lambda \mid \xi, \widehat{\psi}), \tag{42}$$

so that the prior for ξ is updated, while the conditional posterior for λ given ξ is not updated by the data. Thus, in general the estimate for θ is based, in part, on a prior which is not updated by the data.

5. Age-Period-Cohort Models

We will now apply the theoretical considerations to analyse the age-period-cohort model. The methodological literature on this model is large and the consequences of the above theory are wide ranging.

In Section 5.1 we present the age-period-cohort model along with the maximal invariant parameter. This maximal invariant parameter is also called the canonical parameter because the age-period-cohort model is usually implemented as an exponential family; see Section 2.4.1. When formulating the model we choose a notation matching the age-period-cohort literature rather than the reserving literature. At the same time the exposition takes it starting point in Kuang et al. [19], but the notation deviates.

The implementation of the canonical parameter depends on the type of data array. In Section 5.2 design matrices are given for age-cohort, age-period, and period-cohort data arrays. While they illustrate interesting differences in the structure for these data arrays, they also provide the basis for an immediate implementation via any generalised linear model software. The age-cohort model is expressed as a hypothesis of the age-period-cohort model in Section 5.3. Time effects and forecasting are considered in Section 5.4, while the two-sample age-period-cohort model is discussed in Section 5.5.

5.1. The Model and the Canonical Parameter. Here the age-period model is set up and a quite general identification result is reported.

Consider data Y_{ij} indexed by $(i, j) \in \mathcal{I}$ where i is the age and j is the period. The index set may be a rectangle given by $i = 1, \ldots, I$ and $j = 1, \ldots, J$ so that the cohort $k = I - i + j$ runs from 1 to $K = I + J - 1$. More generally, the index set could be a generalized trapezoid where two corners are cut off the rectangle so that the cohort k runs from $1 + h_1$ to $I + J - 1 - h_2$ for some $h_1, h_2 \geq 0$. The class of generalized trapezoids includes the three types of Lexis diagrams discussed by Keiding [20]. We will return to those special cases below.

The statistical model is defined by the assumption that the variables Y_{ij} are independent with an exponential family distribution with predictor μ_{ij} given by

$$\mu_{ij} = \alpha_i + \beta_j + \gamma_k + \delta \quad \text{for } i, j \in \mathcal{I}. \tag{43}$$

The time effect $\theta = (\alpha_1 \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{h_1+1}, \ldots, \gamma_{I+J-1-h_2}, \delta)'$ now varies in some time effect space $\Theta \in \mathbb{R}^q$ where $q = I + J + K + 1 - h_1 - h_2$.

The model (43) is of the form (1) discussed in Section 2. Specifically, the predictors μ_{ij} can be stacked in a vector μ of dimension $n = \dim \mathcal{I}$ and written as $\mu = D\theta$. Thus, the parameter space for the model is of the form $M = (\mu \in \mathbb{R}^n : \mu = D\theta \text{ for } \theta \in \Theta)$ as outlined in (2). The mapping $\theta \mapsto \mu$ from Θ to M is surjective and the equivalence classes in the time effect space can be described by a group of transformations that are discussed in (8). This group can be represented as

$$g : \begin{pmatrix} \alpha_i \\ \beta_j \\ \gamma_k \\ \delta \end{pmatrix} \longmapsto \begin{pmatrix} \alpha_i + a + (i-1)\,d \\ \beta_j + b - (j-1)\,d \\ \gamma_k + c + (k-1)\,d \\ \delta - a - b - c - (I-1)\,d \end{pmatrix} \quad \text{for } \theta \in \Theta, \tag{44}$$

for any a, b, c, and d. This is of the form (8) with $\zeta = (a, b, c, d)'$ although the definition of the matrix A depends on the structure of the index set $\mathcal{I}$.

A first clue for the canonical parametrisation is given by Fienberg and Mason [21] and Clayton and Schiffler [22] who pointed out that, on the multiplicative scale, ratios of relative risks are invariant. On the additive scale this amounts to looking at second differences, such as $\Delta^2\alpha_i = \alpha_i - 2\alpha_{i-1} + \alpha_{i-2}$. A graphical illustration of the double differences is given in Figure 1 (graphics were done using R 3.0.2, see R Development Core Team [23]), which is taken from Miranda et al. [24]. Panel (a) illustrates the interpretations of the formula for $\Delta^2\alpha_i$ as follows. Consider the 1970 and 1971 cohorts. In 2010 these have ages 40 and 39, while in 2011 these have ages 41 and 40. Thus, $\Delta^2\alpha_{41}$ represents the increase in mortality from ages 40 to 41 in 2011 relative to the increase from ages 39 to age 40 in 2010. An equivalent interpretation is that which represents the increase in mortality from ages 40 to 41 for the 1970 cohort relative to the increase from ages 39 to 40 for the 1971 cohort. In a similar way panels (b) and (c) illustrate the formulas for $\Delta^2\beta_{2012}$ and $\Delta^2\gamma_{1972}$.

Kuang et al. [19] introduces a parameter formed by these second differences as well as three entries of the predictor; that is,

$$\xi = \left(\mu_{i_1 j_1}, \mu_{i_2 j_2}, \mu_{i_3 j_3}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I, \Delta^2\beta_3, \ldots, \Delta^2\beta_J, \Delta^2\gamma_{h_1+3}, \ldots, \Delta^2\gamma_{K-h_2}\right). \tag{45}$$

The parameter ξ varies in the space $\Xi = \mathbb{R}^p$ where $p = q - 4$. If the three points $\mu_{i_1 j_1}$, $\mu_{i_2 j_2}$, and $\mu_{i_3 j_3}$ are chosen not to be linearly related then they define the levels and the linear trends in the predictor. The formal condition is that a certain determinant defined from the indices is nonzero; that is,

$$i_2 j_3 - i_3 j_2 + i_3 j_1 - i_1 j_3 + i_1 k_2 - i_2 k_1 \neq 0. \tag{46}$$

Theorem 4 (see [19], [25, Corollary 2]). *Let μ satisfy* (43). *If the condition* (46) *is satisfied then the parameter ξ of* (45) *satisfies the following:*

(i) *ξ is a function of θ which is invariant to the group g in* (44)*;*

(ii) *μ is a function of ξ;*

(iii) *the parametrisation of μ by ξ is exactly identified in that $\xi^\dagger \neq \xi^\ddagger \Rightarrow \mu(\xi^\dagger) \neq \mu(\xi^\ddagger)$.*

Theorem 4 therefore shows that ξ varies freely in $\Xi = \mathbb{R}^p$. Moreover, ξ is a maximal invariant of the mapping m from θ to μ under the transformations g. It should be noted that the choice of maximal invariant is not unique. Indeed, any bijective mapping of ξ can serve as maximal invariant. The choice of ξ is convenient since it becomes the canonical parameter in generalized linear models of the exponential family type.

In itself this theorem does not tell how to express the predictor μ in terms of the canonical parameter ξ. The link depends on the structure of the index set $\mathcal{I}$. The above mentioned paper gives implicit expressions for generalized trapezoid index sets. In the following we report explicit expressions for the 3 principal Lexis diagrams.

5.2. Design Matrices for Lexis Diagrams. The link between the canonical parameter ξ and the predictor μ is analysed for the 3 principal Lexis diagrams. We start with age-cohort data arrays, which were the focus of attention in Kuang et al. [19]. Such arrays are easiest to analyse because all three time scales increase from the point where $i = j = k = 1$. As a consequence the results are relatively easier for these arrays.

5.2.1. Age-Cohort Data Arrays. Age-cohort data arrays are rectangular in the age and cohort indices and given by

$$\mathcal{I}_{\text{ac}} = \{(i, k) : i = 1, \ldots, I,\ k = 1, \ldots, K\}. \tag{47}$$

Consequently, the period index $j = i + k - 1$ varies over $j = 1, \ldots, J = I + K - 1$. Keiding [20] refers to this Lexis diagram as the first principal set of death.

Age-cohort arrays are in particular used for reserving in general insurance. In that situation, only the triangle $1 \leq i, j, k \leq I$ is observed. The issue is to forecast the other triangle in the square $1 \leq i, k \leq I$. In the reserving literature these triangles are referred to as the upper and lower triangles, since the cohort axis has reverse order. The two-factor age-cohort model for triangular age-cohort arrays is known as the chain-ladder model; see England and Verrall [26] for an overview. Zehnwirth [27] introduced an age-period-cohort model for such triangular arrays. The identification issue is analysed in detail in Kuang et al. [19, 25]. Subsequently, Kuang et al. [28] analysed the Poisson likelihood, while Kuang et al. [10] give an empirical analysis focusing on forecasting.

The age-period-cohort model for the age-cohort arrays is parametrised by

$$\mu_{ik} = \alpha_i + \beta_{i+k-1} + \gamma_k + \delta \quad \text{for } i, k \in \mathcal{I}_{\text{ac}}. \tag{48}$$

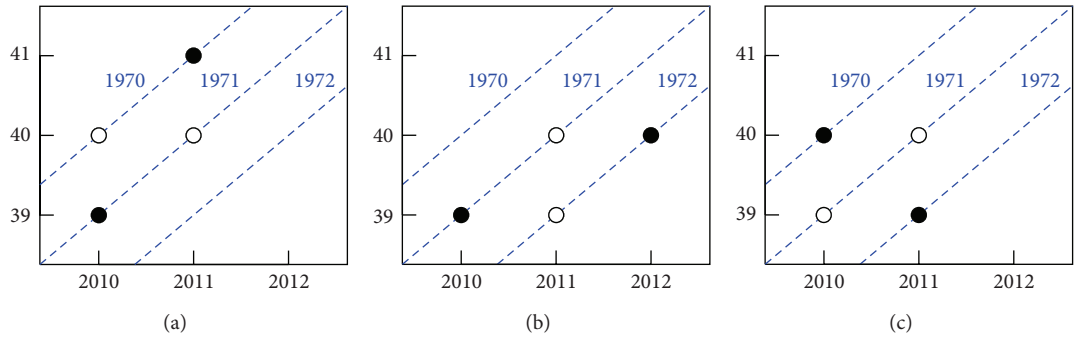

Figure 1: Illustration of interpretation of $\Delta^2\alpha_{41}$, $\Delta^2\beta_{2012}$, and $\Delta^2\gamma_{1972}$.

The time effect $\theta = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_1, \ldots, \gamma_K, \delta)'$ now varies in $\Theta = \mathbb{R}^{2(I+K)}$.

The design matrix linking the canonical parameter ξ in (45) and the predictor μ is essentially an identity linking the two parameters. A natural choice of the three levels points to the predictors that are μ_{11}, μ_{12}, and μ_{21}. We then get the representation

$$\mu_{ik} = \mu_{11} + (i-1)(\mu_{21} - \mu_{11}) + (k-1)(\mu_{12} - \mu_{11}) + \sum_{\ell=3}^{i}\sum_{h=3}^{\ell}\Delta^2\alpha_h + \sum_{\ell=3}^{j}\sum_{h=3}^{\ell}\Delta^2\beta_h + \sum_{\ell=3}^{k}\sum_{h=3}^{\ell}\Delta^2\gamma_h, \tag{49}$$

with the convention that empty sums are zero, and recalling that second differences are defined as $\Delta^2\alpha_i = \alpha_i - 2\alpha_{i-1} + \alpha_{i-2}$ so that $\sum_{h=3}^{i}\Delta^2\alpha_h = \Delta\alpha_i - \Delta\alpha_2$ and $\sum_{\ell=3}^{i}\sum_{h=3}^{\ell}\Delta^2\alpha_h = \alpha_i - \alpha_1 - (i-1)\Delta\alpha_2$.

The identity (49) is crucial to the understanding of the age-period-cohort model. It shows that the predictor has a single level expressed as μ_{11}, which in turn satisfies $\mu_{11} = \alpha_1 + \beta_1 + \gamma_1 + \delta$. The level μ_{11} is therefore estimable, but the individual levels α_1, β_1, γ_1, and δ are not identifiable from the model. Further, the model has two linear trends, here expressed with slopes $\mu_{21} - \mu_{11}$ and $\mu_{12} - \mu_{11}$ in terms of the age and cohort indices. These slopes can be expressed as $\mu_{21} - \mu_{11} = \Delta\alpha_2 + \Delta\beta_2$ and $\mu_{12} - \mu_{11} = \Delta\beta_2 + \Delta\gamma_2$. They are estimable, but the individual slopes $\Delta\alpha_2$, $\Delta\beta_2$, and $\Delta\gamma_2$ are not identifiable.

The design matrix now follows from the identity (49) so that the predictor satisfies $\mu = X\xi$, where

$$\xi = \left(\mu_{11}, \mu_{21} - \mu_{11}, \mu_{12} - \mu_{11}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I, \Delta^2\beta_3, \ldots, \Delta^2\beta_J, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K\right)', \tag{50}$$

$$X_{ik} = \{1, (i-1), (k-1), h(i,3), \ldots, h(i,I), h(j,3), \ldots, h(j,J), h(k,3), \ldots, h(k,K)\}', \tag{51}$$

where $\xi \in \mathbb{R}^p$, where $p = 2(I + K - 2)$ and $h(t,s) = \max(t - s + 1, 0)$.

The identification relies on Theorem 4, which can be specialised to age-cohort arrays as follows.

Theorem 5 (see [19, Theorem 1]). *Let μ satisfy* (48). *The parameter ξ of* (50) *satisfies the following:*

(i) *ξ is a function of θ which is invariant to the group g in* (44);

(ii) *μ is a function of ξ, because of* (49);

(iii) *the parametrisation of μ by ξ is exactly identified in that $\xi^\dagger \neq \xi^\ddagger \Rightarrow \mu(\xi^\dagger) \neq \mu(\xi^\ddagger)$.*

Theorem 5 in turn implies that the parameter ξ varies freely in $\Xi = \mathbb{R}^p$, while the design matrix X given by (51) has full column rank. Originally, the more general Theorem 4 was proved as a corollary to Theorem 5.

5.2.2. Age-Period Arrays. An age-period data array is rectangular in the age and cohort indices and given by

$$\mathcal{I}_{\text{ap}} = \{(i,j) : i = 1, \ldots, I, \ \ j = 1, \ldots, J\}. \tag{52}$$

Consequently, the cohort index $k = j - i + I$ varies over $k = 1, \ldots, K = I + J - 1$. Keiding [20] refers to this Lexis diagram as the third principal set of death.

Age-period arrays are commonly used in epidemiology, in mortality analysis, and in sociology. The analysis of identification issue is largely similar to that of age-cohort arrays. However, the representation of the predictor μ in terms of ξ differs in an intriguing way, because the third time index, the cohort k, is the difference of the other two indices.

The age-period-cohort model for the age-period arrays is parametrised by

$$\mu_{ij} = \alpha_i + \beta_j + \gamma_{j-i+I} + \delta \quad \text{for } i, j \in \mathcal{I}_{\text{ap}}. \tag{53}$$

The time effect $\theta = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_1, \ldots, \gamma_K, \delta)'$ now varies in $\Theta = \mathbb{R}^{2(I+J)}$. A representation of the predictor μ in terms of the canonical parameter ξ is now

$$\mu_{ij} = \mu_{I1} + (i-I)(\mu_{I1} - \mu_{I-1,1}) + (j-1)(\mu_{I2} - \mu_{I1}) + \sum_{\ell=i}^{I-2}\sum_{h=\ell}^{I-2}\Delta^2\alpha_h + \sum_{\ell=3}^{j}\sum_{h=3}^{\ell}\Delta^2\beta_h + \sum_{\ell=3}^{j-i+I}\sum_{h=3}^{\ell}\Delta^2\gamma_{h+2}. \tag{54}$$

The representation (54) differs from that of (49) in a subtle way. The three reference points for the levels of the predictor are chosen in the corner $i = I$, $j = 1$. From this corner period and cohort indices increase, while age decreases. Hence, the age double differences $\Delta^2\alpha_i$ are now cumulated backwards. This phenomenon arises because the cohort index is the difference of the principal indices of age and period, whereas for the age-cohort array the period index is the sum of the principal indices of age and cohort. The predictor is now $\mu = X\xi$ where, with $h(t,s) = \max(t - s + 1, 0)$,

$$\begin{aligned}\xi = \big(&\mu_{I1}, \mu_{I1} - \mu_{I-1,1}, \mu_{I2} - \mu_{I1}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I,\\ &\Delta^2\beta_3, \ldots, \Delta^2\beta_J, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K\big)',\end{aligned} \tag{55}$$

$$\begin{aligned}X_{ij} = \big\{&1, i - I, j - 1, h(1,i), \ldots, h(I-2,i),\\ &(j,3), \ldots, h(j,J), h(j-i+I,3), \ldots,\\ &h(j-i+I,K)\big\}'.\end{aligned} \tag{56}$$

The identification relies on Theorem 4. It is specialised to age-period arrays as follows.

Theorem 6 (see [24, Theorem 4.1]). *Let μ satisfy* (53). *The parameter ξ of* (55) *satisfies the following:*

(i) *ξ is a function of θ which is invariant to the group g in* (44)*;*

(ii) *μ is a function of ξ, because of* (54)*;*

(iii) *the parametrisation of μ by ξ is exactly identified in that $\xi^\dagger \neq \xi^\ddagger \Rightarrow \mu(\xi^\dagger) \neq \mu(\xi^\ddagger)$.*

The group of transformations in (44) can be specialised as

$$g : \begin{pmatrix}\alpha_i\\ \beta_j\\ \gamma_{i-i+I}\\ \delta\end{pmatrix} \longmapsto \begin{pmatrix}\alpha_i + a + di\\ \beta_j + b - dj\\ \gamma_{j-i+I} + c + d(j - i + I)\\ \delta - a - b - c - dI\end{pmatrix} \quad \text{for } \theta \in \Theta = \mathbb{R}^{2(I+J)}; \tag{57}$$

see, for instance, Carstensen [29]. This is of the form (8) with $\zeta = (a,b,c,d)'$ and

$$A'_\perp = \begin{pmatrix} 1 & 1 & \cdots & 1 & & & & & & & & -1\\ & & & & 1 & 1 & \cdots & 1 & & & & -1\\ & & & & & & & & 1 & 1 & \cdots & 1 & -1\\ 1 & 2 & \cdots & I & -1 & -2 & \cdots & -J & 1 & 2 & \cdots & K & -I \end{pmatrix}. \tag{58}$$

5.2.3. Period-Cohort Arrays. An age-period data arrays is rectangular in the age and cohort indices and given by

$$\mathcal{I}_{\text{pc}} = \{(j,k) : j = 1, \ldots, J,\ k = 1, \ldots, K\}. \tag{59}$$

Consequently, the age index $i = j - k + K$ varies over $i = 1, \ldots, I = J + K - 1$. Keiding [20] refers to this Lexis diagram as the second principal set of death. Age-period arrays are commonly used in prospective cohort studies in epidemiology and in sociology. The analysis is similar to that of age-period arrays when swapping the role of age and cohort.

The age-period-cohort model for the age-cohort arrays is parametrised by

$$\mu_{jk} = \alpha_{j-k+1} + \beta_j + \gamma_k + \delta \quad \text{for } j,k \in \mathcal{I}_{\text{ap}}. \tag{60}$$

The time effect $\theta = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_1, \ldots, \gamma_K, \delta)'$ now varies in $\Theta = \mathbb{R}^{2(J+K)}$. A representation of the predictor μ in terms of the canonical parameter ξ is now

$$\begin{aligned}\mu_{jk} = {}&\mu_{1K} + (j-1)(\mu_{2K} - \mu_{1K}) + (k - K)(\mu_{1K} - \mu_{1,K-1})\\ &+ \sum_{\ell=3}^{j-k+1}\sum_{h=3}^{\ell}\Delta^2\alpha_h + \sum_{\ell=3}^{j}\sum_{h=3}^{\ell}\Delta^2\beta_h\\ &+ \sum_{\ell=i}^{K-2}\sum_{h=\ell}^{K-2}\Delta^2\gamma_{h+2}.\end{aligned} \tag{61}$$

Thus, the canonical parameter and the design matrix are given by

$$\begin{aligned}\xi = \big(&\mu_{1K}, \mu_{2K} - \mu_{1K}, \mu_{1K} - \mu_{1,K-1}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I,\\ &\Delta^2\beta_3, \ldots, \Delta^2\beta_J, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K\big)',\end{aligned} \tag{62}$$

$$\begin{aligned}X_{jk} = \big\{&1, j - 1, k - K, h(j-k+1,3), \ldots, h(j-k+1,I),\\ &h(j,3), \ldots, h(j,J), h(1,k), \ldots, h(K-2,k)\big\}'.\end{aligned} \tag{63}$$

In parallel with Theorem 6 we then have the following identification result.

Theorem 7. *Let μ satisfy* (60). *The parameter ξ of* (62) *satisfies the following:*

(i) *ξ is a function of θ which is invariant to the group g in* (44)*;*

(ii) *μ is a function of ξ, because of* (61)*;*

(iii) *the parametrisation of μ by ξ is exactly identified in that $\xi^\dagger \neq \xi^\ddagger \Rightarrow \mu(\xi^\dagger) \neq \mu(\xi^\ddagger)$.*

5.3. Expressing the Age-Cohort Model as a Hypothesis. It is often of interest to test the absence of the period effect. An application to analysing asbestos related mortality can be found in Miranda et al. [24].

The hypothesis is that $\beta_1 = \cdots = \beta_J$, when expressed in terms of the time effect parameters. The restricted model is given by, with $k = j - i + I$,

$$\mu^{\text{ac}}_{ij} = \alpha_i + \gamma_{j-i+I} + \delta \quad \text{for } i,j \in \mathcal{I}_{\text{ap}}. \tag{64}$$

The identification problem simplifies to a question of determining the levels of α_i and γ_k. Therefore the (log) relative risk parameters $\Delta\alpha_i$ are identified as pointed out by Clayton

and Schifflers [30]. In this model the cohort index is present and keeps the difference of the principal age and period indices. Therefore the representation of the predictor involves backward cumulated age differences as before but with a subtle change of sign, so that (54) reduces to

$$\mu_{ij}^{\mathrm{ac}} = \mu_{I1} - \sum_{\ell=i}^{I-1} \Delta\alpha_{\ell+1} + \sum_{\ell=2}^{k} \Delta\gamma_{\ell}. \tag{65}$$

As a consequence the canonical parameter and the design reduce to $\mu_{ij}^{\mathrm{ac}} = X_{ij}^{\mathrm{ac}} \xi_{ij}^{\mathrm{ac}}$, where

$$X_{ij}^{\mathrm{ac}} = \left\{1, -1_{(1\geq i)}, \ldots, -1_{(I-1\geq i)}, 1_{(k\geq 2)}, \ldots, 1_{(k\geq K)}\right\}, \tag{66}$$

$$\xi^{\mathrm{ac}} = \left(\mu_{I1}, \Delta\alpha_2, \ldots, \Delta\alpha_I, \Delta\gamma_2, \ldots, \Delta\gamma_K\right)'. \tag{67}$$

Miranda et al. [24, Theorem 4.2] establish an identification result similar to Theorem 6.

The age-cohort model can also be formulated as a hypothesis on the maximal invariant ξ in the age-period-cohort model following Section 2.4.3. The period effects $\Delta^2\beta_j$ are set to zero through $H'\xi = 0$, where $H' = (0, I_{J-2}, 0)$. Applying this to the expression for ξ in (55) gives

$$\begin{aligned}\xi_H = \overline{H}_{\perp}'\xi = \big(&\mu_{I1}, \Delta\alpha_I - \Delta\gamma_2, \Delta\gamma_2, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I,\\ &\Delta^2\gamma_3, \ldots, \Delta^2\gamma_K\big),\end{aligned} \tag{68}$$

since in the absence of period effects; then $\mu_{I1} - \mu_{I-1,1} = \Delta\alpha_I - \Delta\gamma_2$ and $\mu_{I2} - \mu_{I1} = \Delta\gamma_2$. The double differences cumulate to first differences through $\sum_{i=3}^{I} \Delta^2\alpha_i = \Delta\alpha_I - \Delta\alpha_2$, so the above expression ξ_H is seen to be a linear transformation of ξ^{ac} in (67). In other words the age-cohort model arises from the age-period-cohort model by restricting the maximal invariant parameter.

5.4. Working with the Time Effect. There is a large literature seeking to identify the original time effects α_i, β_j, and γ_k of the age-period-cohort model from the predictor. Here we look closer at some of those ad hoc identification proposals.

5.4.1. Ad Hoc Identification of Levels. For the age-period-cohort model it is popular to impose ad hoc identifications in two steps of the type discussed in Section 3.3. Here the first step is concerned with the level of the time effects and the second step is concerned with the linear trend. Examples are given in Sections 5.4.2 and 5.5.4.

A common first step ad hoc identification is to require that

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{k=1}^{K} \gamma_k = 0. \tag{69}$$

This ad hoc identification is specific to the chosen data range. For instance, the constraint $\sum_{i=1}^{I} \alpha_i = 0$ is not easily transferable to a different data set drawn from the same population but with a different set of age groups. This aspect would have to be kept in mind if a substantive motivation was to be found for this constraint. Other ad hoc identification schemes such as $\alpha_I = \beta_J = \gamma_K = 0$ have similar problems.

The constraint (69) is a special case of affine constraints of the form $C'\theta_C = \psi$ discussed in Section 3.3. The involved dimensions are $q = 2(I + J)$ and $p = q - 4$, while the number of constrains is $q - q_C = 3$. The matrix $C' \in \mathbb{R}^{(q-q_C)\times q}$ is given by the top left $\{3 \times (q-1)\}$-block of $A_{\perp}'$ in (58) padded with a column of zeros, while $\psi \in \mathbb{R}^3$ is given by $\psi = 0$. Theorem 1 shows that $m = A_{\perp}'C \in \mathbb{R}^{(q-p)\times(q-q_c)}$ has full rank. Indeed, m and its orthogonal complement are given by

$$m = A_{\perp}'C = \begin{pmatrix} I & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & K \\ I\breve{i} & -J\breve{j} & K\breve{k} \end{pmatrix}, \qquad m_{\perp} = \begin{pmatrix} -\breve{i} \\ \breve{j} \\ -\breve{k} \\ 1 \end{pmatrix}, \tag{70}$$

where, for instance, $\breve{i} = I^{-1}\sum_{i=1}^{I} i = (I+1)/2$. Thus, the constrained group of equivalence classes (27) is

$$g_C : \begin{pmatrix} \alpha_i \\ \beta_j \\ \gamma_k \\ \delta \end{pmatrix} \longmapsto \begin{Bmatrix} \alpha_i + d\left(i - \breve{i}\right) \\ \beta_j - d\left(j - \breve{j}\right) \\ \gamma_k + d\left(k - \breve{k}\right) \\ \delta \end{Bmatrix} \quad \text{for } \theta \in \Theta_C. \tag{71}$$

5.4.2. Ad Hoc Identification of Slopes: The "Intrinsic" Estimator. The "intrinsic" estimator is a popular estimator in the sociology literature; see Yang et al. [4] and see also O'Brien [31, 32] and Fu et al. [33] for a recent discussion of its merits. It has its roots in a suggestion by Kupper et al. [34], with an early critique given by Holford [35].

The "intrinsic" estimator is defined in two steps. In the first step, the levels are identified by the ad hoc constraint (69). Three of the θ-coordinates are then dropped; that is α_I, β_J, and γ_K are dropped. In a second step the linear trend is ad hoc identified using a Moore-Penrose inverse as in (23).

We can analyse these steps using the developed framework. The first step identifies the levels by the ad hoc constraint (69), which is a constraint of the form $C'\theta = 0$ for the C discussed in Section 5.4.1. This θ is defined on Θ_C which is a linear subspace with a dimension deficiency of 3. Introduce a selection matrix $S_{\perp} \in \mathbb{R}^{q\times(q-3)}$ that selects all coordinates of θ except α_I, β_J, and γ_K. Thus $S_{\perp}$ arises as a q-dimensional with 3 columns deleted corresponding to α_I, β_J, and γ_K. This is chosen so that $(C, S_{\perp})$ is invertible. Then $S_{\perp}'\theta$ is freely varying in that $S_{\perp}'\Theta_C = \mathbb{R}^{q-3}$. The skew projection identity $I_q = S(C'S)^{-1}C' + C_{\perp}(S_{\perp}'C_{\perp})^{-1}S_{\perp}'$ and the constraint $C'\theta = 0$ then implies that $\theta = C_S\vartheta$ where $C_{S\perp} = C_{\perp}(S_{\perp}'C_{\perp})^{-1}$ and $\vartheta = S_{\perp}'\theta \in \mathbb{R}^{q-3}$. Note that while $C_{S\perp}$ depends on $S_{\perp}$ and $C_{\perp}$, it does not depend on the normalisation of $C_{\perp}$, since we can replace $C_{\perp}$ by $C_{\perp}m$ for arbitrary invertible matrices $m \in \mathbb{R}^{(q-3)\times(q-3)}$. This implies that C_S is a function of $S_{\perp}$ and C. The predictor μ is now parametrised by $\mu = XA'\theta = XA_C'\vartheta$ with $A_C' = A'C_S$. This corresponds to equation 5 of Yang et al. [4] who use the notation X and b for XA_C' and ϑ, respectively.

In the second step the linear trend is ad hoc identified through a time effect parameter of the form (23) with A, θ replaced by A_C, ϑ so that $\theta_{\text{ad.hoc}} = C_\perp \vartheta_{\text{ad.hoc}}$ where $\vartheta_{\text{ad.hoc}} = L_\perp (A_C' L_\perp)^{-1} \xi + (A_C)_\perp \{L' (A_C)_\perp\}^{-1} \lambda$ for some scalar λ and some matrix $L_\perp \in \mathbb{R}^{(q-3)\times p}$.

The "intrinsic" estimator is ad hoc identified through the choices $\lambda = 0$ and $L_\perp = A_C$, while C is chosen by (69). It therefore estimates an "intrinsic" parameter:

$$\theta_{\text{intrinsic}} = C_{S\perp} C_{S\perp}' A \left(A' C_{S\perp} C_{S\perp}' A \right)^{-1} \xi, \tag{72}$$

which depends on the choices of $S_\perp'$, C, and $A_\perp$. However, since we can replace $C_\perp$ by $C_\perp m$ for arbitrary invertible matrices $m \in \mathbb{R}^{(q-3)\times(q-3)}$ without changing $\theta_{\text{intrinsic}}$ the expression $\theta_{\text{intrinsic}}$ does not depend on the normalisation of $C_\perp$. The "intrinsic" parameter satisfies the following result.

Theorem 8. *The "intrinsic" parameter is an injective mapping of the canonical parameter $\xi \in \mathbb{R}^p$ into a $p = q-4$ dimensional linear subspace $\Theta_{\text{intrinsic}}$ of $\Theta = \mathbb{R}^q$. The "intrinsic" time effect space is a p-dimensional linear subspace of $\mathbb{R}^q$ of the form*

$$\begin{aligned} &\Theta_{\text{intrinsic}} \\ &= \left\{ \theta \in \mathbb{R}^q : \theta = C_{S\perp} C_{S\perp}' A \left(A' C_{S\perp} C_{S\perp}' A \right)^{-1} \xi \text{ for } \xi \in \mathbb{R}^p \right\}, \\ &= \left\{ \theta \in \mathbb{R}^q : C'\theta = 0,\ w' \left(C_\perp' S_\perp \right) \left(S_\perp' C_\perp \right) \overline{C}_\perp' \theta = 0 \right\}, \end{aligned} \tag{73}$$

where $w \in \mathbb{R}^{q-3}$ is uniquely defined up to a scale by $w' C_\perp' A = 0$.

Theorem 8 implies that the "intrinsic" parameter should be interpreted as an object varying in the linear subspace $\Theta_{\text{intrinsic}}$ rather than in the unrestricted time effect space $\Theta = \mathbb{R}^q$. As outlined in Section 3.4 this has consequences for the interpretation of plots of the time effects, hypothesis testing, and forecasts. A consequence of this argument is that different choices of $C, S_\perp, L$, and λ would lead to other ad hoc identified parameters varying in other affine subspaces of Θ. In other words, the "intrinsic" estimator carries the cost of working with the somewhat complicated linear subspace $\Theta_{\text{intrinsic}}$. This effort may be worthwhile if the particular choice of C, $S_\perp$, L, and λ can be made on substantive grounds.

5.4.3. Forecasting. Forecasting of future mortality rates involves an extrapolation of the time parameters. In Section 2.4.4 it was argued that ad hoc identification may introduce an undesirable arbitrariness in the forecast. When working exclusively with the canonical parameter ξ this arbitrariness is avoided. It is, however, also possible to work with ad hoc identified time effects under specific circumstances that we characterise here for age-period arrays. This builds on the theory developed in Kuang et al. [25] for age-cohort data arrays.

In the context of an age-period data array $\mathcal{I}_{\text{ap}}$ it is often of interest to forecast h periods ahead. Suppose it is of interest to forecast the mortality at age i in period $J + h$, so that the cohort is $k = I + J + h - i$. This requires an extrapolation of the period effect. If the cohort index is sufficiently large, that is, $k > K$, then the cohort effect needs to be extrapolated too. Thus, there are two forecast index arrays of interest:

$$\begin{aligned} \mathcal{I}_{\text{ap}} &= \{(i,j) : i = 1,\ldots,I;\ j = J+1,\ldots,J+h;\ k \le K\}, \\ \mathcal{K}_{\text{ap}} &= \{(i,j) : i = 1,\ldots,I;\ j = J+1,\ldots,J+h;\ k > K\}. \end{aligned} \tag{74}$$

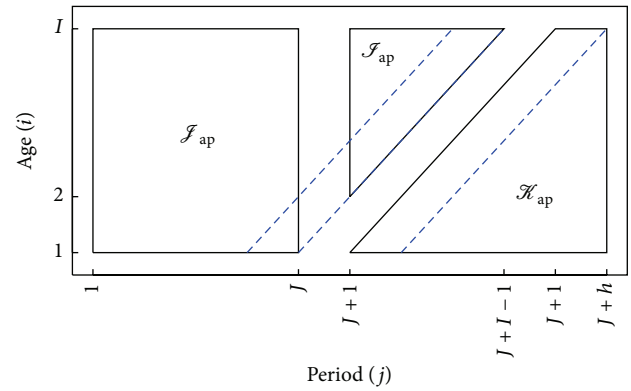

FIGURE 2: $\mathcal{I}_{\text{ap}}$ is the data array. $\mathcal{I}_{\text{ap},1}$ is the forecast array where only period parameters need to be extrapolated. $\mathcal{I}_2$ is the forecast array where both period and cohort parameters need to be extrapolated. Cohorts are indicated by dashed lines.

Figure 2 illustrates these forecast index arrays.

Identification plays a role when extrapolating the estimates obtained on the data array $\mathcal{I}_{\text{ap}}$. The identification issues can be ignored if the investigator simply extrapolates $\Delta^2 \beta_j$ and $\Delta^2 \gamma_k$. In the context of ad hoc identified time effects arbitrary linear trends are introduced in the model. The forecast of the predictor $\mu_{i,J+h}$ is invariant to these if and only if the chosen extrapolation method for β_j, γ_k preserves these linear trends so that they can cancel with the arbitrary linear trend in α_i. The next result gives a precise formulation of this statement. It applies both to point forecasts and distribution forecasts.

Theorem 9. *Consider the predictor μ_{ij} for $i, j \in \mathcal{I}_{\text{ap}}$ as given in (53). Suppose the time effects α_i, β_j, and γ_k are ad hoc identified. Consider the class of h periods-ahead forecasts over $\mathcal{I}_{\text{ap}}$ constructed as $\widetilde{\mu}_{i,J+h} = \widetilde{\alpha}_i + \widetilde{\beta}_{J+h} + \widetilde{\gamma}_{I+J+h-i} + \widehat{\delta}$, where $\widetilde{\beta}_{J+h} + \widetilde{\gamma}_{I+J+h-i}$ is a function of the ad hoc identified estimate $\widehat{\theta}$. Let g be the group (57). Invariance of the forecast $\widetilde{\mu}_{i,J+h}$ with respect to the ad hoc identification is equivalent to either of the following:*

(i) *the extrapolation method for period and cohort effects is linear trend-preserving:*

$$\begin{aligned} &\widetilde{\beta}_{J+h}\left(\widehat{\theta}\right) + \widetilde{\gamma}_{I+J+h-i}\left(\widehat{\theta}\right) \\ &\quad = \left[\widetilde{\beta}_{J+h}\left\{g\left(\widehat{\theta}\right)\right\} - b + d\,(J+h)\right] \\ &\qquad + \left[\widetilde{\gamma}_{I+J+h-i}\left\{g\left(\widehat{\theta}\right)\right\} - c - d\,(I+J+h-i)\right] \\ &\qquad\qquad \forall b, c, d \in \mathbb{R}; \end{aligned} \tag{75}$$

(ii) *functions* f_β, f_γ *exist so that with* $\widehat{\xi}_{\beta,\gamma} = (\Delta^2\widehat{\beta}_3, \ldots, \Delta^2\widehat{\beta}_J, \Delta^2\widehat{\gamma}_3, \ldots, \Delta^2\widehat{\gamma}_K)'$; *then*

$$\widetilde{\beta}_{J+h}\left(\widehat{\theta}\right) + \widetilde{\gamma}_{I+J+h-i}\left(\widehat{\theta}\right) = \left\{\widehat{\beta}_J + h\Delta\widehat{\beta}_J + f_\beta\left(\widehat{\xi}\right)\right\} + \left\{\widehat{\gamma}_K + (h-i+1)\,\Delta\widehat{\gamma}_K + f_\gamma\left(\widehat{\xi}\right)\right\}. \tag{76}$$

To illustrate the use of Theorem 9 consider the extrapolation methods $\widetilde{\beta}_{J+h} = \widehat{\beta}_J$ and $\Delta\widetilde{\beta}_{J+h} = \Delta\widehat{\beta}_J$. The first forecast is a random walk forecast and it is seen to violate (ii). The second forecast is a cumulated random walk and satisfies (ii). The reason is that $\beta_{J+h} = \beta_J + \sum_{\ell=1}^{h}\Delta\beta_{J+\ell}$. Since $\Delta\widetilde{\beta}_{J+\ell} = \Delta\widehat{\beta}_J$, then $\widetilde{\beta}_{J+h} = \widehat{\beta}_J + h\Delta\widehat{\beta}_J$. Further examples of forecasts that are linear trend-preserving as well as some which are not are given Kuang et al. [25, Table 1].

Kuang, Nielsen, and Nielsen [10] apply this to reserving data organised in an age-cohort array $\mathcal{I}_{\text{ac}}$ and discuss the issue of robustification of forecast with respect to structural breaks at the forecast origin. Miranda et al. [24] give an application to asbestos related mortality using an age-period array $\mathcal{I}_{\text{ap}}$.

5.4.4. Bayesian Ad Hoc Identification Using a Dynamic Prior. A Bayesian ad hoc identification using a dynamic prior does not solve the identification problem as discussed in Section 4 and the same care has to be exercised to avoid the problems outlined in Section 3.4. Berzuini and Clayton [6] suggest such an ad hoc identification approach. On page 831 they write "*Identificability problems may be solved by imposing an arbitrary linear constraint on the log-linear trend components of age, period and cohort effects. Happily, such an arbitrary constraint has no effect on the predictions of the model.*" The previous analysis suggests that this is far from innocent.

The Berzuini-Clayton suggestion is to ad hoc identify the model (53) through

$$\begin{aligned}\alpha_i &= \alpha_1 + \alpha_2 i + \sum_{\ell=3}^{i}\sum_{h=3}^{\ell}\Delta^2\alpha_h,\\ \beta_j &= \beta_1 + \beta_2 j + \sum_{\ell=3}^{j}\sum_{h=3}^{\ell}\Delta^2\beta_h,\\ \gamma_k &= \gamma_1 + \gamma_2 k + \sum_{\ell=3}^{k}\sum_{h=3}^{\ell}\Delta^2\gamma_h,\\ \delta &= 0.\end{aligned} \tag{77}$$

A dynamic prior is chosen so that the double differences $\Delta^2\alpha_i$, $\Delta^2\beta_j$, and $\Delta^2\gamma_k$ are independent zero mean normal with variances $\phi = (\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2)$ that have χ^2-type prior. The purpose of this is in part to facilitate extrapolations $\Delta^2\alpha_i$, $\Delta^2\beta_j$, and $\Delta^2\gamma_k$ for $i > I$, $j > J$, and $k > K$, which is done through further draws from normal distributions. The level/trend effects $\theta_{\text{level}} = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)'$ have independent uniform priors on some large intervals.

We will analyse the Berzuini-Clayton model as applied to an age-period data array $\mathcal{I}_{\text{ap}}$. Decompose the canonical parameter ξ from (54) into two parts: the slope and level parameters, say $\xi_\mu = (\mu_{I1}, \mu_{I1} - \mu_{I-1,1}, \mu_{I2} - \mu_{I1})'$, and the collection of double differences, say ξ_Δ. The assumed prior for ξ_Δ is a simple collection of independent normal distributions with variances ϕ. The assumed prior for ξ_μ is a linear combination of not only the independent uniform variables θ_{level}, but also on ξ_Δ, since the age double differences $\Delta^2\alpha_i$ are cumulated backwards in (54), but forwards in (77). Thus, the prior for $\xi = (\xi_\mu', \xi_\Delta')'$ depends on the θ_{level} construction.

We get a hyper-parameter $\lambda_{\text{hyper}} = (\lambda, \phi)$, where λ is some three-dimensional ad hoc identified level/trend effect dependent on $\theta_{\text{level}}, \xi_\Delta$. We will argue that the ad hoc identified level/trend effect λ will wash out in the Berzuini-Clayton model. However, the level/trend parameter ξ_μ is a function of the θ_{level} construction that is tailored to the ad hoc identification. That construction remains in the analysis.

In the presentation of the posterior Berzuini and Clayton are careful only to consider the double differences ξ_Δ and stay clear of the ad hoc identified level/trend effect θ_{level}. Theorem 2 yields the posterior $p(\xi \mid y) = p(y \mid \xi)p(\xi)/p(y)$. Thus, the marginal posterior for the double differences is $p(\xi_\Delta \mid y) = \int p(y \mid \xi_\Delta, \xi_\mu)p(\xi_\Delta, \xi_\mu)d\xi_\mu/p(y)$. This links ξ_Δ to ξ_μ and in turn to the θ_{level} construction.

The extrapolative method is based on double differences so it only depends on λ_{hyper} through ϕ due to Theorem 9 and the subsequent discussion. Thus, the extrapolative method is of the form $p(\widetilde{\mu} \mid \xi, \lambda_{\text{hyper}}, y) = p(\widetilde{\mu} \mid \xi, \phi, y)$. By construction it does not reduce to $p(\widetilde{\mu} \mid \xi, y)$ so that condition (39) for Theorem 3 is not satisfied. The distribution forecast is of the form

$$p\left(\widetilde{\mu} \mid y\right) = \iint p\left(\widetilde{\mu} \mid \xi, \phi, y\right) p\left(\xi \mid y\right) p\left(\phi \mid \xi\right) d\phi\; d\xi, \tag{78}$$

which, apart from depending on the θ_{level} construction, also depends on the conditional prior $p(\phi \mid \xi)$, which is not updated by the likelihood.

In summary, it appears that the Berzuini-Clayton analysis depends on the θ_{level} construction as well as the conditional prior $p(\phi \mid \xi)$. The dependence on the θ_{level} construction could of course be addressed by introducing priors directly on ξ_μ, which in turn would be updated by the likelihood. Since the conditional prior $p(\phi \mid \xi)$ cannot be updated by the likelihood that its sole justification rests on the substantial context.

5.4.5. A Functional Form Hypothesis. It is instructive to consider functional form restrictions on the time effects. Such hypotheses can be analysed using the results outlined in Section 3.4.2. As an example restrict the age effect to be quadratic in a similar way to Yang and Land [5] so that

$$\alpha_i = \sigma_0 + \sigma_1 i + \sigma_2 i^2 \quad \text{for } i = 1, \ldots, I. \tag{79}$$

This restriction on the time effect can be analysed by writing it on the form $R'\theta = \rho$, see (28), and then applying

Theorem A.3. Alternatively, in this particular case, we can show that the restriction actually only affects the ad hoc identified time effect through the canonical parameter, so a simpler analysis can be made.

A quadratic polynomial has constant second order derivative. Therefore the restriction (79) implies

$$\Delta^2\alpha_i = 2\sigma_2 \quad \text{for } i = 3,\dots,I. \tag{80}$$

This expression has one free parameter. Thus, it is useful to consider the third order difference:

$$\Delta^3\alpha_i = \Delta^2\alpha_i - \Delta^2\alpha_{i-1} = 0 \quad \text{for } i = 4,\dots,I. \tag{81}$$

This gives $I-3$ linear restrictions on the canonical parameter. The age time effect α_i then has three remaining parameters, say α_1, α_2, and α_3. These are freely varying since the parameters σ_0, σ_1, and σ_2 are freely varying.

If the constraint is imposed directly on the canonical parameter, the restricted model is a regular exponential family with the advantages outlined in Section 2.4. However, if the analysis is done with the time effect the levels and trend will have to be ad hoc identified while bearing in mind the issues discussed above.

5.4.6. The "Hierarchical Age-Period Cohort Regression Model". In some cases a random effects approach can be used to get an overview of the many parameters of the age-period model. When applied to the time effects this implies an ad hoc identification. An example is the "hierarchical age-period cohort regression model" by Yang and Land [5]. In that paper the age effect is given a quadratic structure, but that does not have to be the case. The model is then given by

$$\begin{aligned} \alpha_i &= \sigma_0 + \sigma_1 i + \sigma_2 i^2, \qquad \beta_j \overset{\mathsf{D}}{=} \mathsf{N}\left(0,\sigma_\beta^2\right), \\ \gamma_k &\overset{\mathsf{D}}{=} \mathsf{N}\left(0,\sigma_\gamma^2\right), \qquad \delta = 0. \end{aligned} \tag{82}$$

Since random effects are only introduced for some of the time effects, the analysis of Section 4.3 has to modified in a similar way to the analysis in Section 5.4.4.

From (80) it is seen that the model restricts $\Delta^2\alpha_i = 2\sigma_2$. Thus, divide the canonical parameter ξ into three elements: the slope and level parameters, say $\xi_\mu = (\mu_{I1}, \mu_{I1} - \mu_{I-1,1}, \mu_{I2} - \mu_{I1})'$, the age-double differences $\xi_\alpha = (\Delta^2\alpha_3,\dots,\Delta^2\alpha_I)$, and the remaining double differences $\xi_{\beta,\gamma}$. Here ξ_α is restricted by the hypothesis $2\sigma_2$ and $\xi_{\beta,\gamma}$ is linear function of the normal random effects, while ξ_μ is a three-dimensional linear function of σ_2 and of the six-dimensional object $\nu = (\sigma_0,\sigma_1,\beta_1,\beta_2,\gamma_1,\gamma_2)'$. This leaves a three-dimensional ad hoc identified level/slope parameter λ which is also a function of ν but not entering the likelihood. Let $\psi = (\sigma_0,\sigma_1,\sigma_2,\sigma_\beta^2,\sigma_\gamma^2)$.

The random effects likelihood are constructed in three steps. First, we have the usual age-period-cohort likelihood $p(y \mid \xi)$. Secondly, the random effects distribution for ξ_μ, $\xi_{\beta,\gamma}$, and λ is multivariate normal, while ξ_α is deterministic function of ψ. Thus, decompose the prior as $p(\xi_\mu,\xi_{\beta,\gamma},\lambda \mid \psi) = p(\xi_\mu,\xi_{\beta,\gamma} \mid \psi)p(\lambda \mid \xi_\mu,\xi_{\beta,\gamma},\psi)$. Thirdly, following Section 4.3 the random effects likelihood will not depend on $p(\lambda \mid \xi_\mu,\xi_{\beta,\gamma},\psi)$ and it is given by

$$\begin{aligned} &p(y \mid \psi) \\ &\quad = \int p\left(y \mid \xi_\alpha,\xi_\mu,\xi_{\beta,\gamma}\right) p\left(\xi_\alpha,\xi_\mu,\xi_{\beta,\gamma} \mid \psi\right) d\left(\xi_\mu,\xi_{\beta,\gamma}\right). \end{aligned} \tag{83}$$

The prior $p(\lambda \mid \xi_\mu,\xi_{\beta,\gamma},\psi)$ is not updated by the data. Plots and inferences based on the posterior $p(\theta \mid \psi, y)$ will then suffer from the ad hoc identification issues outlined in Section 3.4.

5.5. A Two-Sample Age-Period-Cohort Model. When confronted with two samples for women and for men it may be of interest to apply the age-period-cohort model (43) to each of the samples and impose that some of the time effects are the same across samples. The models for samples $r = 1, 2$ are

$$\mu_{ijr} = \alpha_{ir} + \beta_{jr} + \gamma_{kr} + \delta_r \quad \text{for } i,j \in \mathcal{I},\ r = 1,2. \tag{84}$$

The time effect $\theta = (\dots,\alpha_{ir},\beta_{jr},\gamma_{kr},\delta_r,\dots)'$ now varies in $\Theta = \mathbb{R}^q$ where $q = 4(I+J)$.

5.5.1. Analysis of the Unrestricted Two-Sample Model. The unrestricted two-sample model is simply analysed as two copies of the one sample model of Section 5.1. The time effects of each copy are only defined up to linear trends. The group of transformations characterizing the identification problem combines two copies of the one sample group (44). The maximal invariant parameter is $\xi = (\xi_1', \xi_2')' \in \mathbb{R}^p$ where $p = 4(I+J-2)$ and each of ξ_r are of the form (45). The benefits of Section 2 hold when working with that parameter.

5.5.2. Bayesian Ad Hoc Identification Using a Dynamic Model. An application of the unrestricted two-sample model can be found in Cairns et al. [36]. The two samples are the population of England and Wales and the subpopulation of assured lives, so the substantive question is whether there is a selection effect for the assured lives. A Bayesian model with dynamic prior is used. It shares some features with the Berzuini and Clayton [6] model discussed in Section 5.4.4 although the details of the ad hoc identification of the levels and slopes are slightly different. When it comes to forecasting the extrapolative method appears to depend on the ad hoc identified parameter as well as the hyperparameters. This complicates the analysis of the forecast relatively the discussion in Section 5.4.4.

5.5.3. The Hypothesis of Common Period Parameters. The two-sample model allows the possibility for adding cross-sample restrictions on the parameters. As an example we consider the hypothesis of common period parameters.

Working with the canonical parameter the hypothesis is

$$\Delta^2\beta_{i1} = \Delta^2\beta_{i2} \quad \text{for } j = 3,\dots,J. \tag{85}$$

This is a simple linear restriction as that discussed in Section 2.4.3. It is readily seen that the degrees of freedom of

the hypothesis are $p - p_H = J - 2$ so the dimension of the restricted model is $p_H = 4I + 3J - 6$. The canonical parameter under the hypothesis is then

$$\xi_H = \left(\ldots, \mu_{11r}, \mu_{21r} - \mu_{11r}, \mu_{12r} - \mu_{11r}, \Delta^2\alpha_{ir}, \Delta^2\beta_j, \Delta^2\gamma_{kr}, \ldots\right)'. \tag{86}$$

The same result arises when writing the hypothesis in terms of time effects so that

$$\beta_{j1} = \beta_{j2} \quad \text{for } j = 1, \ldots, J. \tag{87}$$

Such hypotheses on the time effect were discussed in Section 3.4.2. It can be analysed using the general result in Theorem A.3. However, we will take the simpler route of arguing that this only restricts the canonical parameter given a hypothesis of the type (85). The argument relies on noting that analysing the restriction for the predictors μ_{ij1} and μ_{ij2} is equivalent to analysing the restriction for the predictors μ_{ij1} and $\mu_{ij2} - \mu_{ij1}$, where the cross-sample differenced predictor is of the form

$$\mu_{ij2} - \mu_{ij1} = (\alpha_{i2} - \alpha_{i1}) + (\beta_{j2} - \beta_{j1}) + (\gamma_{k2} - \gamma_{k1}) + (\delta_2 - \delta_1). \tag{88}$$

Now, the restricted model for the cross-sample differenced predictor $\mu_{ij2} - \mu_{ij1}$ is an age-cohort model:

$$\mu_{ij2} - \mu_{ij1} = (\alpha_{i2} - \alpha_{i1}) + (\gamma_{k2} - \gamma_{k1}) + (\delta_2 - \delta_1). \tag{89}$$

Following the analysis of Section 5.3 the (87) therefore implies the $J - 2$ linear restrictions given by (85). At the same time the predictor for the first sample μ_{ij1} is left unrestricted by (87). In summary, the restrictions (85) and (87) are equivalent.

The restriction has an interesting implication for the interpretation of the involved double differences. For the unrestricted model it was found that only plain double differences, such as $\Delta^2\alpha_{jr}$, are identified. Under the restriction the cross-sample differenced predictor is of age-cohort form (89) so also the cross-double differences $\Delta(\alpha_{i2} - \alpha_{i1})$ and $\Delta(\gamma_{k2} - \gamma_{k1})$ are identified.

5.5.4. Step-Wise Ad Hoc Identification under the Hypothesis. The analysis of Riebler and Held [7] finds that the difference $\alpha_{i2} - \alpha_{i1}$ is identified under the hypothesis (85). This is not consistent with the above analysis showing that the cross-sample differenced predictor is an age-cohort model under the hypothesis, for which levels such as $\alpha_{i2} - \alpha_{i1}$ are identified.

The apparent difference comes about because Riebler and Held follow a step-wise identification approach along the lines of Sections 3.3 and 5.4.1. In a first step the time effects α_{ir}, β_{jr}, and γ_{kr} are constrained to have zero-sums as in (69). In a second step the slopes are ad hoc identified using a Bayesian approach similar to that of Berzuini and Clayton [6]; see Sections 4 and 5.4.4 for a discussion of the consequences.

The identification in the first step implies that $\alpha_{i2} - \alpha_{i1}$ has a zero sum. Under the hypothesis (85) this is exactly what is needed to ad hoc identify the levels in the age-cohort model (89). In other words a different level identification in the first step leads to a different level for the difference $\alpha_{i2} - \alpha_{i1}$.

6. Models with Nonlinear Parametrisations

Some additional issues arise when looking at models with nonlinear parametrisations. A prominent example is the mortality model proposed by Lee and Carter [3] and which is the current benchmark in mortality studies done by government agencies and pension funds. For this model the time effect space Θ has a nondifferentiability which can actually be avoided by working directly with the parameter space M.

We analyze the Lee-Carter model in Section 6.1. In Section 6.2 we turn to a two-sample problem where some additional difficulties can arise when forecasting.

6.1. The Lee-Carter Model. The mortality model proposed by Lee and Carter [3] has predictor of the form

$$\mu_{ij} = \alpha_i + \beta_i\kappa_j \quad \text{for } i, j \in \mathcal{I}_{\text{ap}}. \tag{90}$$

The time effects $\theta = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_I, \kappa_1, \ldots, \kappa_J)$ vary in $\Theta = \mathbb{R}^{2I+J}$.

Lee and Carter pointed towards two identification issues of the model. If α, β, and κ are one solution to (90), then $\alpha - \beta c$, β, $\kappa + c$ is also a solution for any scalar c, just as α, β/d, and κd are a solution for any $d \neq 0$. Consequently, they proposed the ad hoc identification:

$$\sum_i \beta_i = 1, \qquad \sum_j \kappa_j = 0. \tag{91}$$

This is, however, not the full story about the identification issues. To get at this we follow the outline from the linear parametrised models and start by finding the parameter space for the predictor μ.

6.1.1. The Parameter Space. We start by finding the predictor space M. Write the model in matrix form. Let $\underline{\mu}$ denote the $I \times J$-matrix of μ_{ij}. Then

$$\underline{\mu} = \alpha\iota' + \beta\kappa', \tag{92}$$

where α, β, and κ are vectors concatenating α_i, β_i, and κ_j and where $\iota = (1, \ldots, 1)' \in \mathbb{R}^J$. Postmultiply by the projection identity $I_J = \bar{\iota}\iota' + \bar{\iota}_\perp\iota_\perp'$ to get

$$\underline{\mu} = \alpha\iota' + \beta\kappa'\left(\bar{\iota}\iota' + \bar{\iota}_\perp\iota_\perp'\right) = \left(\alpha + \beta\kappa'\bar{\iota}\right)\iota' + \beta\left(\kappa'\bar{\iota}_\perp\right)\iota_\perp', \tag{93}$$

where the orthogonal complement $\iota_\perp$ can be chosen so that $\iota_\perp'\kappa = (\Delta\kappa_2, \ldots, \Delta\kappa_J)'$ but could also be chosen otherwise. Equation (93) shows that the model is composed of two matrices with rank one. Thus, the parameter space is given by

$$M = \left\{\underline{\mu} \in \mathbb{R}^{I\times J} : \underline{\mu} = \gamma\iota' + \delta\iota_\perp' \text{ for } (\gamma, \delta) \in \mathbb{R}^I \times \mathbb{R}^{I\times(J-1)} \text{ so } \operatorname{rank}(\delta) \leq 1\right\}. \tag{94}$$

Note that M does not depend on the normalisation of $\iota_\perp$ since δ is freely varying. The space M is a manifold since the space of matrices δ with an upper bound to the rank is a manifold as opposed to the space where δ has rank of unity. This space can be parametrised parsimoniously by

$$\xi = (\gamma, \delta) \quad \text{where } \gamma = \alpha + \beta\kappa'\bar{\iota}, \ \delta = \beta\kappa'\bar{\iota}_\perp, \tag{95}$$

varying in the manifold

$$\Xi = \left\{(\gamma, \delta) \in \mathbb{R}^I \times \mathbb{R}^{I\times(J-1)} : \operatorname{rank}(\delta) \le 1\right\}. \tag{96}$$

The ξ is the candidate for the maximal invariant describing the equivalence classes of the mapping from the time effect θ to the predictor μ.

The next step is to analyse the time effect space Θ. It is convenient to decompose M into two disjoint sets depending on the rank of δ. These sets are

$$\begin{aligned} M_1 &= \left\{\underline{\mu} \in \mathbb{R}^{I\times J} : \underline{\mu} = \gamma\iota' + \delta\iota'_\perp \text{ for } (\gamma, \delta) \in \mathbb{R}^I \times \mathbb{R}^{I\times(J-1)} \right. \\ &\qquad \left. \text{so } \operatorname{rank}(\delta) = 1\right\}, \\ M_0 &= \left(\underline{\mu} \in \mathbb{R}^{I\times J} : \underline{\mu} = \gamma\iota' \text{ for } \gamma \in \mathbb{R}^I\right). \end{aligned} \tag{97}$$

Correspondingly, the time effect space Θ can be decomposed into two disjoint sets:

$$\begin{aligned} \Theta_1 &= \left(\theta \in \Theta : \exists i, j \text{ so } \beta_i\kappa_j \ne \beta_i\kappa_J\right), \\ \Theta_0 &= \left(\theta \in \Theta : \forall i, j \text{ so } \beta_i\kappa_j = \beta_i\kappa_J\right). \end{aligned} \tag{98}$$

Note that $\delta = 0$ if and only if $\theta \in \Theta_0$. Consider first the time effect space Θ_1, which is implicitly what Lee and Carter had in mind. The mapping $\theta \mapsto \mu$ on Θ_1 to M is invariant to the group of transformations:

$$g_1 : \begin{pmatrix} \alpha_i \\ \beta_i \\ \kappa_j \end{pmatrix} \longmapsto \begin{pmatrix} \alpha_i + \beta_i c \\ \dfrac{\beta_i}{d} \\ (\kappa_j - c)\, d \end{pmatrix}, \tag{99}$$

acting on Θ_1 for all $c \in \mathbb{R}$ and all $d \ne 0$. The parameter $\xi = (\gamma', \delta')'$ is invariant under g_1 acting on Θ_1. Now, consider the space Θ_0 with deficient rank. Then α_i, β_i, and κ_j map into $\alpha_i + \varphi_i$ where $\varphi_i = \beta_i\kappa_j$ is constant in j, so that $\delta = \beta\kappa'\bar{\iota}_\perp = 0$. This mapping is invariant to the group of transformations:

$$g_0 : \begin{pmatrix} \alpha_i \\ \beta_i\kappa_j \end{pmatrix} \longmapsto \begin{pmatrix} \alpha_i + a_i \\ \beta_i\kappa_j - a_i \end{pmatrix}, \tag{100}$$

acting on Θ_0 for all $(a_1, \ldots, a_I)' \in \mathbb{R}^I$.

Theorem 10. *Let $\underline{\mu} \in M$. The parameter $\xi \in \Xi$ of* (95) *satisfies the following:*

(i) *ξ is a function of $\theta \in \Theta$ which is invariant to the groups g_0, g_1 in* (99) *and* (100)*;*

(ii) *$\underline{\mu}$ is a function of ξ;*

(iii) *the parametrisation of $\underline{\mu}$ by ξ is exactly identified in the sense that $\xi^\dagger \ne \xi^\ddagger \Rightarrow \underline{\mu}(\xi^\dagger) \ne \underline{\mu}(\xi^\ddagger)$.*

Theorem 10 shows that ξ varies freely on the space Ξ and it gives a unique parametrisation of μ. As a function of θ it is invariant to g_0, g_1; hence it is a maximal invariant.

It is interesting to compare the properties of the spaces M, Ξ, and Θ. The spaces M and Ξ are spaces of matrices with deficient rank. These are smooth spaces, but they are not vector spaces since the sum of matrices with rank one may have rank larger than one. In contrast Θ is a vector space. The mapping from Θ to M will inevitably be nondifferentiable. This nondifferentiability is avoided by working directly with M. Likewise, in a Bayesian setting it would seem more difficult to introduce a meaningful prior of Θ with its nondifferentiability than on M.

6.1.2. Maximum Likelihood Estimation. The maximum likelihood estimator for ξ can be derived analytically in the normal case.

Consider a situation where the data array is of age-period form so Y_{ij} for $(i, j) \in \mathcal{I}_{\text{ap}}$. Suppose Y_{ij} are independent normal with mean μ_{ij} and variance σ^2. Organise the data in a matrix $\underline{Y}$. Then the log likelihood is of the form

$$\ell\left(\underline{\mu}, \sigma^2; \underline{Y}\right) = -\frac{IJ}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\operatorname{tr}\left\{\left(\underline{Y} - \underline{\mu}\right)\left(\underline{Y} - \underline{\mu}\right)'\right\}. \tag{101}$$

The maximum likelihood estimator is of the following form. Subsequently, this is related to the estimator suggested by Lee and Carter.

Theorem 11. *For a normal age-period array parametrised by* (94) *the maximum likelihood estimators are*

$$\widehat{\gamma} = \underline{Y}\iota\left(\iota'\iota\right)^{-1}, \qquad \widehat{\delta} = \left[\operatorname{svd}_1\left\{\underline{Y}\iota_\perp\left(\iota'_\perp\iota_\perp\right)^{-1}\iota'_\perp\right\}\right]\iota_\perp\left(\iota'_\perp\iota_\perp\right)^{-1}, \tag{102}$$

where $\operatorname{svd}_1(\cdot)$ is the singular value decomposition truncated to one factor.

Thus, γ is estimated by the row-averages of the data matrix, while δ is estimated by the singular value decomposition of the row-wise demeaned data matrix.

6.1.3. Estimation of Ad Hoc Identified Time Effects. The ad hoc identification (91) gives a time effect θ_λ varying in a $2I + J - 2$ dimensional affine subspace of $\Theta = \mathbb{R}^{2I+J}$. The ad hoc identified θ_λ can now be expressed in terms of the maximal invariant parameter ξ using (95). In the case where $\delta \ne 0$ then it has singular value decomposition $\delta = \delta_L\delta_S\delta'_R$ for two vectors $\delta_L \in \mathbb{R}^I$ and $\delta_R \in \mathbb{R}^{J-1}$ so $\delta'_L\delta_L = 1$ and $\delta_R\delta'_R = 1$, while $\delta_S > 0$ is a positive scale. The ad hoc identification of Lee and Carter then gives

$$\alpha_\lambda = \gamma, \qquad \beta_\lambda = \delta_L\left(\iota'\delta_L\right)^{-1}, \qquad \kappa_\lambda = \iota_\perp\delta_R\iota'\delta_L\delta_S. \tag{103}$$

Inserting the maximum likelihood estimators from Theorem 11 yields the estimators proposed by Lee and Carter. However, the disentangling of the singular values and singular vectors of $\widehat{\delta}$ is done by the ad hoc identification $\beta'\iota = 1$ and $\kappa'\iota = 0$. These estimators are therefore specific to the considered data array and data set in parallel with the discussion in Sections 3.2 and 5.4.1.

6.1.4. Consequences of the Possible Rank Deficiency. The parameter space M was split into spaces M_1 and M_0 depending on the rank of δ. The space M_0 is a Lebesgue null set relative to M. Broadly speaking, there are two consequences of the possible rank deficiency. The first consequence is an estimation problem arising in the vicinity of M_0. The second consequence is that the usual normal asymptotic distribution theory does not apply in the vicinity of M_1. Whether this becomes a problem in practice depends on the data. One solution is to ensure that the time effect really is present when using the Lee-Carter model.

Investigate whether the time effects are present amounts to estimating the rank of δ. For a given data set two Lee-Carter models can be estimated. The first model with predictor space M is the unrestricted model in which $\operatorname{rank}(\delta) \leq 1$. The second model has predictor space M_0 so $\delta = 0$. Twice the difference of the likelihood values gives a likelihood ratio test statistic which is asymptotically χ^2. If the smaller model, M_0, is accepted this is used in subsequent analysis. However, if the smaller model M_0 is rejected then it is likely that the predictor is not located in the vicinity of M_0 and it is then safe to work with the predictor space M_1.

The consistency of this step-wise procedure is discussed in a cointegration context by Johansen [37, Section 12]. Even when this procedure points towards working with the parameter space M_1 the rank deficiency may still affect inference under M_1. Analysis of simple canonical correlation models suggests that inference under M_1 will be nearly similar if the distance to M_0 is sufficiently large. A problem is that the distribution for the test statistic will have poor finite sample properties when the parameter value is close to M_0. A simple way to get around this problem is to test for M_0 using a test with lower level than the conventional level. A more complicated way to address this is to employ a finite sample correction when seeking to test for M_0. See Nielsen [38, 39] for further discussion of this issue in the context of simple canonical correlation models.

The rank deficiency issue is typically not encountered in a standard Lee-Carter analysis. The reason is that the analysis is typically applied to data where there is a marked improvement in mortality rates over time. A Lee-Carter analysis could however run into trouble if it were applied to data without a strong calendar effect. The issue becomes more pertinent when extending Lee-Carter model with a cohort component such as

$$\mu_{ij} = \beta_i^{(1)} + \beta_i^{(2)}\kappa_j^{(2)} + \beta_i^{(3)}\gamma_{j-i+I}^{(3)}; \tag{104}$$

see Renshaw and Haberman [40]. If the cohort effect is modest the latter matrix is nearly rank deficient and the likelihood will be nearly flat in certain directions. This is presumably the reason for the estimation problem noted by Cairns et al. [41].

6.1.5. Forecasting. The purpose of Lee-Carter model is usually to forecast future mortality. This issue is considered for the model with parameter space M_1. The standard approach is to extrapolate κ, ad hoc identified through, for instance, $\kappa'\iota = 0$. The h-step ahead extrapolation of κ_{J+h} based on some forecast methods is denoted by $\widetilde{\kappa}_{J+h}(\widehat{\kappa})$. Combined with the estimates $\widehat{\alpha}_i$, $\widehat{\beta}_i$ this gives the overall forecast

$$\widetilde{\mu}_{i,J+h}\left(\widehat{\theta}\right) = \widehat{\alpha}_i + \widehat{\beta}_i\widetilde{\kappa}_{J+h}\left(\widehat{\kappa}\right). \tag{105}$$

The identification question is then for which extrapolation methods this equals

$$\widetilde{\mu}_{i,J+h}\left\{g_1\left(\widehat{\theta}\right)\right\} = \left(\widehat{\alpha}_i + \widehat{\beta}_i c\right) + \left(\frac{\widehat{\beta}_i}{d}\right)\widetilde{\kappa}_{J+h}\left\{\left(\widehat{\kappa} - c\right)d\right\}. \tag{106}$$

The condition for avoiding adverse impact of the ad hoc identification is as follows.

Theorem 12. *Let $\underline{\mu} \in M_2$. The forecast $\widetilde{\mu}_{i,J+h}$ in (105) is invariant to ad hoc identification if and only if the extrapolation method for the period effect is location-scale preserving:*

$$\widetilde{\kappa}_{J+h}\left\{\left(\widehat{\kappa} - c\iota\right)d\right\} = \left\{\widetilde{\kappa}_{J+h}\left(\widehat{\kappa}\right) - c\right\}d \quad \forall c \in \mathbb{R},\ \forall d \neq 0. \tag{107}$$

The default forecast method in the literature is a random walk with a drift, which was the preferred forecast of Lee and Carter [3]. This is given by

$$\widetilde{\kappa}_{J+h} = \widetilde{\kappa}_{J+h-1} + \nu_c + \varepsilon_h, \tag{108}$$

with estimates $\widehat{\nu}_c = (J-1)^{-1}\sum_{j=2}^{J}(\widehat{\kappa}_j - \widehat{\kappa}_{j-1})$ and normal errors ε_h with mean zero and estimated variance $\widehat{\sigma}^2(\widehat{\kappa}) = (J-2)^{-1}\sum_{j=2}^{J}(\widehat{\kappa}_j - \widehat{\kappa}_{j-1} - \widehat{\nu}_c)^2$. This extrapolation method is location-scale preserving as required in Theorem 12. It is even linear trend preserving. Other valid forecasts are a random walk without intercept as given by the equation $\widetilde{\kappa}_{J+h} = \widetilde{\kappa}_{J+h-1} + \varepsilon_h$, or an autoregression given by $\widetilde{\kappa}_{J+h} = \rho\widetilde{\kappa}_{J+h-1} + \nu_c + \varepsilon_h$.

An alternative approach to forecasting would consider the predictor of the model for a particular age ground, say i. That predictor is $\widehat{\underline{\mu}}_i = e_i'(\widehat{\gamma}\iota' + \widehat{\delta}\iota_\perp')$, where e_i is the ith unit vector. From this we can generate forecasts $\widetilde{\mu}_{i,J+h}$ using any time series method. The resulting forecast will in general depend on $\widehat{\kappa}$ as well as $\widehat{\alpha}_i$, $\widehat{\beta}_i$ and it is therefore more general than the forecasts discussed in Theorem 12, which only depends on $\widehat{\kappa}$. The forecast for another age group, say $i^\dagger$, should be the same up to a linear transformation dictated by the Lee-Carter structure. Thus, the h-step ahead forecasts for the entire array are

$$\widetilde{\underline{\mu}}_{J+h} = \frac{\widehat{\delta}\iota_\perp' e_{\check{\imath}}}{e_i'\widehat{\delta}\iota_\perp' e_{\check{\imath}}}\left(\widetilde{\mu}_{i,J+h} - e_i'\widehat{\gamma}\iota'\right) + \widehat{\gamma}\iota', \tag{109}$$

for an index $\check{\imath}$ is chosen so that $e_i'\widehat{\delta}\iota_\perp' e_{\check{\imath}} \neq 0$.

6.1.6. Bayesian Ad Hoc Identification Using a Dynamic Model. A Bayesian model with dynamic specification of the prior has been suggested by Pedroza [42]. Dynamic priors are presented for the time effects $\theta = (\xi, \lambda)$ involving a hyper parameter ϕ. The ad hoc identification (91) is imposed so that analysis is made for an ad hoc identified time effect θ_λ.

Pedroza presents posteriors for θ_λ. When evaluating this posterior one should bear in mind that the conditional prior $p(\lambda \mid \xi)$ is not updated by the data; see Theorem 2. The presented extrapolative method does not depend on λ. Even so, the forecast will depend on conditional prior $p(\phi \mid \xi)$ which is not updated by the data; see Theorem 3.

6.2. The Two-Sample Lee-Carter Model. We now turn to applications of the Lee-Carter model in two-sample problems. Suppose two samples are for women and men. One approach would be to fit separate Lee-Carter models to the two datasets. These Lee-Carter models are of the form

$$\mu_{ijr} = \alpha_{ir} + \beta_{ir}\kappa_{jr} \quad \text{for } i, j \in \mathcal{I}_{\text{ap}},\ r = 1, 2. \tag{110}$$

The objective is now to extrapolate the period effects κ_{jr}. Extrapolating the two models separately using separate random walks is often seen to be volatile, so methods that seek to combine information from both estimated series $\widehat{\kappa}_{jr}$ are sought after. The next result describes the invariance problem in forecasting.

Theorem 13. *Let $\underset{\sim}{\mu}_r \in M_{2r}$ for $r = 1, 2$. The forecast $\widetilde{\mu}_{i,J+h,1}$ for sample $r = 1$ is invariant to ad hoc identification if the extrapolation method $\widetilde{\kappa}_{J+h,1}$ preserves location/scale for sample 1, but is invariant to location and scale for sample 2. That is for all $c_1, c_2 \in \mathbb{R}$ and all $d_1, d_2 \neq 0$; then*

$$\widetilde{\kappa}_{J+h,1}\left\{d_1\left(\widehat{\kappa}_1 - c_1\right), d_2\left(\widehat{\kappa}_2 - c_2\right)\right\} = d_1\left\{\widetilde{\kappa}_{J+h,1}\left(\widehat{\kappa}_1, \widehat{\kappa}_2\right) - c_1\right\}. \tag{111}$$

For one sample the standard forecasting technique appears to be the random walk with a drift as in (108). For the two-sample problem a suggestion could be that women and men should share a common random walk with a drift but deviate from this by a stationary process. In econometrics this idea is referred to as cointegration as proposed by Engle and Granger [43]; see also Johansen [37] for a likelihood based vector autoregressive approach. It is tempting to require that the calendar effects should cointegrate with coefficients of unity, so $\kappa_{j1} - \kappa_{j2}$ should be stationary. However, that apparently intuitive choice violates Theorem 13 because the locations and scales of κ_{jr} are different and arbitrary.

There are two fixes to this problem. The first solution is to work directly with the mortality predictors μ_{ijr} for an arbitrary age group i as outlined for the one-sample case in connection with (109). Since no identification is involved it is permitted to impose that μ_{ij1} and μ_{ij2} cointegrate with coefficients of unity. The forecast for age group i is then carried over to other age groups. The second solution is to work with the estimated series $\widehat{\kappa}_{jr}$ but estimate the cointegrating coefficients from the data. In other words, the cointegrating relation $\widehat{\kappa}_{j1} - \varphi\widehat{\kappa}_{j2} - \psi$ should be zero mean, stationary, with coefficients φ, ψ estimated from the data. This can, for instance, be done by Johansen's approach for a bivariate vector autoregression; see Hendry and Nielsen [44, Section 17].

7. Conclusion

Ad hoc identification is intimately linked to interpretation, inference, numerical analysis, and forecasting. The ad hoc identification will often introduce an arbitrary element in the statistical analysis, whether it is based on frequentist or Bayesian methods. This arbitrary element is entirely avoidable and is in our view best avoided unless there is a clear substantial motivation for ad hoc identification. For decades there has been a debate over how it is best to ad hoc identify mortality models. Our proposal is to bypass this discussion by analysing the surjective mapping between the unidentified time effect parameter and the predictor of the model and then deduce a maximal invariant parametrisation. In our experience there are typically two substantial benefits. First, it simplifies estimation and other statistical computations which is what we have focused on here. Secondly and perhaps more importantly, it helps to focus the substantial question that gives rise to the analysis in the first place.

The issue of dealing with two time scales also occurs in other statistical models, such as the Cox regression model; see Cabrera et al. [45] for a recent application. In future research it would be interesting to consider whether the analysis presented here has any bearing on that problem.

Appendix

A. Proofs

A.1. Some Linear Algebra Results

Lemma A.1. *Consider $A \in \mathbb{R}^{q\times p}$ and $C \in \mathbb{R}^{q\times(q-q_C)}$ so $p, q_C \leq q$. Suppose they have full column rank. Then the following statements are equivalent for some $p_C \leq p$:*

(i) $(A, C) \in \mathbb{R}^{q\times(p+q-q_C)}$ *has rank* $p_C + q - q_C$;

(ii) $A_\perp' C \in \mathbb{R}^{(q-p)\times(q-q_C)}$ *has rank* $p_C + q - q_C - p$;

(iii) $C_\perp' A \in \mathbb{R}^{q_C\times p}$ *has rank* p_C.

Proof of Lemma A.1. (i)⇔(ii) Premultiply the matrix (A, C) with the invertible matrix $(\overline{A}, A_\perp)'$ to get the identity

$$\begin{pmatrix} \overline{A}' \\ A_\perp' \end{pmatrix}(A, C) = \begin{pmatrix} I_p & \overline{A}'C \\ 0 & A_\perp' C \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ 0 & A_\perp' C \end{pmatrix} M \qquad \text{where } M = \begin{pmatrix} I_p & \overline{A}'C \\ 0 & I_{p+q-q_C} \end{pmatrix}. \tag{A.1}$$

Since the first matrix $(\overline{A}, A_\perp)'$ and the last matrix M have full rank then

$$\operatorname{rank}(A, C) = \operatorname{rank}\left(I_p\right) + \operatorname{rank}\left(A_\perp' C\right) = p + \operatorname{rank}\left(A_\perp' C\right). \tag{A.2}$$

(i)⇔(iii) Swap the roles of A, C so $\text{rank}(A,C) = q - q_C + \text{rank}(C_\perp' A)$. □

Lemma A.2. *Consider $A \in \mathbb{R}^{q\times p}$ and $C \in \mathbb{R}^{q\times(q-q_C)}$ so $p \le q_C \le q$. Suppose $(A,C) \in \mathbb{R}^{q\times(p+q-q_C)}$ has full column rank $p+q-q_C$. Then $m = A_\perp' C \in \mathbb{R}^{(q-p)\times(q-q_C)}$ has full column rank and (A,C) has orthogonal complement given by $(A,C)_\perp = A_\perp m_\perp$ where $m_\perp \in \mathbb{R}^{(q-p)\times(q_C-p)}$.*

Proof of Lemma A.2. Since (A,C) has full column rank Lemma A.1 (i, ii) implies that m has full column rank. Since $m_\perp' A_\perp'(A,C) = (m_\perp' A_\perp' A, m_\perp' m) = 0$ we argue that $(A, C, A_\perp m_\perp)$ is invertible. Premultiply by the invertible matrix $(\overline{A}, A_\perp)'$ to get

$$\begin{pmatrix} \overline{A}' \\ A_\perp' \end{pmatrix}(A, C, A_\perp m_\perp) = \begin{pmatrix} I_p & \overline{A}' C & 0 \\ 0 & m & A_\perp' A_\perp m_\perp \end{pmatrix}. \quad (A.3)$$

The matrix $(m, A_\perp' A_\perp m_\perp)$ has full rank. Indeed, its inverse is

$$\left(m, A_\perp' A_\perp m_\perp\right)^{-1} = \left[\left(A_\perp' A_\perp\right)^{-1} m\left\{m'\left(A_\perp' A_\perp\right)^{-1} m\right\}^{-1}, \; m_\perp\left(m_\perp' A_\perp' A_\perp m_\perp\right)^{-1}\right]'. \quad (A.4)$$

Thus, the block triangular matrix (A.3) and, hence, $(A, C, A_\perp m_\perp)$, have full rank. □

A.2. Proofs of Main Theorems

Proof of Theorem 1. Since (A,C) has full column rank then Lemma A.2 shows it has orthogonal complement $A_\perp m_\perp$ so that $(A, C, A_\perp m_\perp)$ is invertible. The orthogonal projection identity shows

$$\begin{aligned} \theta &= \left(\overline{A,C}\right)\begin{pmatrix} A' \\ C' \end{pmatrix}\theta + A_\perp m_\perp\left(\overline{A_\perp m_\perp}\right)'\theta \\ &= \left(\overline{A,C}\right)\begin{pmatrix} \xi \\ \psi \end{pmatrix} + A_\perp m_\perp \zeta, \end{aligned} \quad (A.5)$$

by the constraint $C'\theta = \psi$ and the definitions $A'\theta = \xi$ and $(\overline{A_\perp m_\perp})'\theta = \zeta$. This defines the constrained time effect space Θ_C. Consider now the mapping $\theta \mapsto \mu = XA'\theta$. Premultiply the above expression for θ by $A' = (I_p, 0)(A,C)'$ to get $A'\theta = \xi$ so $\mu = X\xi$. Thus, since ψ is fixed, the equivalence classes in Θ_C are given by $g_C : \theta \mapsto A_\perp m_\perp \zeta$, with $\xi = A'\theta$ as a maximal invariant. □

Proof of Theorem 2. (i) With the likelihood (32) so $p(y \mid \xi, \lambda) = p(y \mid \xi)$ then

$$\begin{aligned} p(y) &= \int\left\{\int p(y \mid \xi, \lambda)\, p(\xi, \lambda)\, d\lambda\right\} d\xi \\ &= \int p(y \mid \xi)\left\{\int p(\xi, \lambda)\, d\lambda\right\} d\xi \\ &= \int p(y \mid \xi)\, p(\xi)\, d\xi. \end{aligned} \quad (A.6)$$

(ii) By Bayes formula and the likelihood (32) then

$$\begin{aligned} p(\xi \mid y) &= \frac{p(y \mid \xi)\, p(\xi)}{p(y)}, \\ p(\lambda \mid \xi, y) &= \frac{p(y \mid \xi, \lambda)\, p(\lambda \mid \xi)}{p(y \mid \xi)} = p(\lambda \mid \xi). \end{aligned} \quad (A.7)$$

(iii) The posterior means are

$$\begin{aligned} \mathsf{E}(\xi \mid y) &= \int \xi p(\xi \mid y)\, d\xi, \\ \mathsf{E}(\lambda \mid \xi, y) &= \int \lambda p(\lambda \mid \xi, y)\, d\lambda = \int \lambda p(\lambda \mid \xi)\, d\lambda, \\ \mathsf{E}(\lambda \mid y) &= \int \lambda p(\lambda \mid y)\, d\lambda = \int \lambda\left\{\int p(\lambda, \xi \mid y)\, d\xi\right\} d\lambda \\ &= \int p(\xi \mid y)\left\{\int \lambda p(\lambda \mid \xi)\, d\lambda\right\} d\xi, \end{aligned} \quad (A.8)$$

noting that $p(\lambda \mid \xi, y) = p(\lambda \mid \xi)$. □

Proof of Theorem 3. Consider the expressions in (37) and (38); that is,

$$p(\tilde{\mu} \mid y) = \iint p(\tilde{\mu} \mid \xi, \lambda, y)\, p(\xi \mid y)\, p(\lambda \mid \xi)\, d\lambda\, d\xi \quad (A.9)$$

and a similar expression involving $p^\dagger$. The question is when they are identical. Assuming $p(\tilde{\mu} \mid \xi, \lambda, y) = p(\tilde{\mu} \mid \xi, y)$ as in (39) the expression reduces to

$$\begin{aligned} p(\tilde{\mu} \mid y) &= \int p(\tilde{\mu} \mid \xi, y)\, p(\xi \mid y)\left\{\int p(\lambda \mid \xi)\, d\lambda\right\} d\xi \\ &= \int p(\tilde{\mu} \mid \xi, y)\, p(\xi \mid y)\, d\xi, \end{aligned} \quad (A.10)$$

since the conditional prior integrates to unity. The same applies for the expression involving $p^\dagger(\lambda \mid \xi)$. □

Proof of Theorem 5. Similar to the proof of Kuang et al. [19, Theorem 1], albeit for a rectangular instead of a triangular data array. □

Proof of Theorem 8. Recall $\theta_{\text{intrisic}} = C_{S\perp} C_{S\perp}' A(A' C_{S\perp} C_{S\perp}' A)^{-1}\xi$ where $C_{S\perp} = C_\perp(S_\perp' C_\perp)^{-1}$ as defined in (72). Premultiply by A' and C' to see that $A'\theta_{\text{intrisic}} = \xi$ and $C'\theta_{\text{intrisic}}$. This does, however, not describe the full variation of θ_{intrisic} since (A,C) is not a square matrix. We must extend the matrix (A,C) with columns so that it is square and invertible.

We find a vector $w \in \mathbb{R}^{q-3}$ so that $(A, C, C_\perp w)$ is invertible. Recall $A \in \mathbb{R}^{q\times p}$ where $q = p + 4$. The matrix $C \in \mathbb{R}^{q\times 3}$ is chosen so that (A,C) has full column rank $q - 1 = p + 3$ as discussed in Section 5.4.1. Apply Lemma A.2, swapping the role of A, C, to see $w_\perp = C_\perp' A$, say, has full column rank. Then (A,C) has orthogonal complement $C_\perp w$.

We show that for any invertible matrix $M \in \mathbb{R}^{q\times q}$ then the 1×2 block matrix $\{(A,C), MC_\perp w\}$ is invertible.

To see this hold, premultiply by $\{(A,C),C_\perp w\}'$ to see that an invertible upper triangular matrix arises.

To analyse the properties of θ_{intrisic} it suffices to analyse $\{(A,C),MC_\perp w\}'\theta_{\text{intrisic}}$, since there is a bijective mapping between the two. Choose $M = \overline{C}_\perp(C_\perp' S_\perp)(S_\perp' C_\perp)\overline{C}_\perp' + CC'$. Then it holds that $C_\perp' M' = (C_\perp' S_\perp)(S_\perp' C_\perp)\overline{C}_\perp'$ so that $C_\perp' M' C_{S\perp} C_{S\perp}' = C_\perp'$ and therefore

$$w' C_\perp' M' \theta_{\text{intrisic}} = w' C_\perp' A\left(A' C_{S\perp} C_{S\perp}' A\right)^{-1}\xi = 0, \tag{A.11}$$

since $w' C_\perp' A = 0$ by construction. Thus, it holds that $w'(C_\perp' S_\perp)(S_\perp' C_\perp)\overline{C}_\perp'\theta_{\text{intrisic}} = 0$ as required. □

Proof of Theorem 9. This is a generalisation of the proof of Kuang et al. [25, Theorems 1, 2]. Let $k_h = J + h + I - i - K = h - i - 1$.

(i) Recall the group g in (57). Then (i) follows by comparing the equations

$$\begin{aligned}\widetilde{\mu}_{i,J+h}\left(\widehat{\theta}\right) &= \widehat{\alpha}_i + \widetilde{\beta}_{J+h}\left(\widehat{\theta}\right) + \widetilde{\gamma}_{K+k_h}\left(\widehat{\theta}\right) + \widehat{\delta},\\ \widetilde{\mu}_{i,J+h}\left\{g\left(\widehat{\theta}\right)\right\} &= \widehat{\alpha}_i + a + di + \widetilde{\beta}_{J+h}\left\{g\left(\widehat{\theta}\right)\right\} + \widetilde{\gamma}_{K+k_h}\left\{g\left(\widehat{\theta}\right)\right\}\\ &\quad + \widehat{\delta} - a - b - c - dI.\end{aligned} \tag{A.12}$$

(ii) As in Section 3.1 there is a bijective mapping from θ to λ, ξ, where ξ is invariant to g, but λ is not. The choice of λ is not important. Any extrapolations of $\widehat{\beta}_j$, $\widehat{\gamma}_k$ can then be written in the form

$$\begin{aligned}\widetilde{\beta}_{J+h}\left(\widehat{\theta}\right) &= \widehat{\beta}_J + h\Delta\widehat{\beta}_J + F_\beta\left(\widehat{\lambda},\widehat{\xi}\right),\\ \widetilde{\gamma}_{K+k_h}\left(\widehat{\theta}\right) &= \widehat{\gamma}_K + k_h\Delta\widehat{\gamma}_K + F_\gamma\left(\widehat{\lambda},\widehat{\xi}\right),\end{aligned} \tag{A.13}$$

for some functions F_β, F_γ. Applying the group g it follows

$$\begin{aligned}\widetilde{\beta}_{J+h}\left\{g\left(\widehat{\theta}\right)\right\} &= \widehat{\beta}_J + b - dJ + h\left(\Delta\widehat{\beta}_J - d\right) + F_\beta\left\{g\left(\widehat{\lambda}\right),\widehat{\xi}\right\},\\ \widetilde{\gamma}_{K+k_h}\left\{g\left(\widehat{\theta}\right)\right\} &= \widehat{\gamma}_K + c + dK + k_h\left(\Delta\widehat{\gamma}_K + d\right) + F_\gamma\left\{g\left(\widehat{\lambda}\right),\widehat{\xi}\right\}.\end{aligned} \tag{A.14}$$

Due to (i) it must hold $F(\lambda,\xi) = F_\beta(\lambda,\xi) + F_\gamma(\lambda,\xi)$ and must equal $F\{g(\lambda),\xi\} = F_\beta\{g(\lambda),\xi\} + F_\gamma\{g(\lambda),\xi\}$. The function F must then be constant in the first argument. This must apply in the forecast region $\mathcal{J}_{\text{ap}}$ where γ is not extrapolated. Therefore, it must also hold that the $F_\beta(\lambda,\xi) = F_\beta\{g(\lambda),\xi\}$, and in turn that $F_\gamma(\lambda,\xi) = F_\gamma\{g(\lambda),\xi\}$. Conversely, if the functions F_β, F_γ are constant in λ then the forecast is invariant to g. □

Proof of Theorem 10. (i) Equation (95) shows that ξ is a function of θ.

(ii) Equation (93) shows that $\underline{\mu} = \gamma\iota' + \delta\iota_\perp'$ is a function of ξ.

(iii) The decomposition $\underline{\mu} = \underline{\mu}\bar{\iota}\iota' + \underline{\mu}\bar{\iota}_\perp'\iota_\perp$ shows that there is a one-one mapping between $\underline{\mu}$ and $(\underline{\mu}\bar{\iota}, \underline{\mu}\bar{\iota}_\perp')$. In turn, (93) shows that $(\underline{\mu}\bar{\iota}, \underline{\mu}\bar{\iota}_\perp') = (\gamma,\delta)$. Thus if $(\gamma^\dagger,\delta^\dagger) \neq (\gamma^\ddagger,\delta^\ddagger)$ then $\underline{\mu}^\dagger \neq \underline{\mu}^\ddagger$. □

Proof of Theorem 11. Rewrite the trace term using the identity $I_J = \bar{\iota}\iota' + \bar{\iota}_\perp\iota_\perp'$ to get

$$\begin{aligned}\mathcal{T} = \operatorname{tr}\left\{\left(\underline{Y} - \underline{\mu}\right)\left(\underline{Y} - \underline{\mu}\right)'\right\} &= \operatorname{tr}\left\{\left(\underline{Y} - \underline{\mu}\right)\bar{\iota}\iota'\left(\underline{Y} - \underline{\mu}\right)'\right\}\\ &\quad + \operatorname{tr}\left\{\left(\underline{Y} - \underline{\mu}\right)\bar{\iota}_\perp\iota_\perp'\left(\underline{Y} - \underline{\mu}\right)'\right\}.\end{aligned} \tag{A.15}$$

By (94) then $\underline{\mu} = \gamma\iota' + \delta\iota_\perp'$ so that $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2$ where

$$\begin{aligned}\mathcal{T}_1 &= \operatorname{tr}\left\{\left(\underline{Y} - \gamma\iota'\right)\bar{\iota}\iota'\left(\underline{Y} - \gamma\iota'\right)'\right\} = \iota'\iota\operatorname{tr}\left\{\left(\underline{Y}\bar{\iota} - \gamma\right)\left(\underline{Y}\bar{\iota} - \gamma\right)'\right\},\\ \mathcal{T}_2 &= \operatorname{tr}\left\{\left(\underline{Y} - \delta\iota_\perp'\right)\bar{\iota}_\perp\iota_\perp'\left(\underline{Y} - \delta\iota_\perp'\right)'\right\}\\ &= \operatorname{tr}\left\{\left(\underline{Y}\bar{\iota}_\perp\iota_\perp' - \delta\iota_\perp'\right)\left(\underline{Y}\bar{\iota}_\perp\iota_\perp' - \delta\iota_\perp'\right)'\right\}.\end{aligned} \tag{A.16}$$

The term $\mathcal{T}_1$ has a minimum of zero if and only if $\underline{Y}\bar{\iota} = \gamma$. In $\mathcal{T}_2$ replace for a moment $\delta\iota_\perp'$ by a matrix ϕ with rank of at most one. The altered $\mathcal{T}_2$ is minimised when ϕ is the singular value decomposition of $\underline{Y}\bar{\iota}_\perp\iota_\perp'$ truncated to rank one due; see Golub and van Loan [46, Theorem 2.5.2]. That singular value decomposition has the property that it is zero when multiplied by ι. Therefore it is also the minimiser of the original problem. □

Proof of Theorem 12. It follows by comparing (105) and (106). □

Proof of Theorem 13. Write $\widetilde{\mu}_{i,J+h,1}(\widehat{\theta}_1,\widehat{\theta}_2) = \widehat{\alpha}_{i1} + \widehat{\beta}_{i1}\widetilde{\kappa}_{J+h,1}(\widehat{\kappa}_1,\widehat{\kappa}_2)$. This is equals to $\widetilde{\mu}_{i,J+h,1}\{g_2(\widehat{\theta}_1), g_2(\widehat{\theta}_2)\} = (\widehat{\alpha}_{i1} + \widehat{\beta}_{i1}c_1) + (\widehat{\beta}_{i1}/d_1)\widetilde{\kappa}_{J+h,1}\{d_1(\widehat{\kappa}_1 - c_1), d_2(\widehat{\kappa}_2 - c_2)\}$ under the given condition. □

A.3. A Further Result on Time Effect Hypotheses. Consider the restriction $R'\theta_R = \rho$ of (28). The following result holds.

Theorem A.3. *Consider the restriction*

$$R'\theta_R = \rho \tag{A.17}$$

of (28) *where* $A \in \mathbb{R}^{q\times p}$ *and* $R \in \mathbb{R}^{q\times(q-q_R)}$ *so* $p, q_R \leq q$. *Suppose A and R have full column rank. Then, for some* $p_R \leq \min(p, q_R)$ *it holds* $p_R = \operatorname{rank}(R_\perp' A) = \operatorname{rank}(A,R) - (q - q_R)$. *Then write* $R_\perp' A = ab'$ *for some matrices* $a \in \mathbb{R}^{q_R\times p_R}$ *and* $b \in \mathbb{R}^{p\times p_R}$ *with full column rank. The hypothesis* (28) *restricts the canonical parameter affinely through*

$$b_\perp'\xi = b_\perp' A'\overline{R}\rho, \tag{A.18}$$

so that the degrees of freedom of the restriction is $p - p_R$. *Introduce the parameters*

$$\varphi_1 = a'\overline{R}_\perp'\theta_R, \qquad \varphi_2 = \overline{b}' A'\xi = \varphi_1 + \overline{b}' A'\overline{R}\rho. \tag{A.19}$$

Then, the predictor μ can be written as the p_R-dimensional affine subspace of the form

$$M_R = \left(\mu \in \mathbb{R}^n : \mu = Xb\varphi_1 + XA'\overline{R}\rho \text{ for } \varphi_1 \in \mathbb{R}^{p_R}\right) \quad (A.20)$$

$$= \left(\mu \in \mathbb{R}^n : \mu = Xb\varphi_2 + X\overline{b}_\perp b_\perp' A'\overline{R}\rho \text{ for } \varphi_2 \in \mathbb{R}^{p_R}\right). \quad (A.21)$$

The mapping $\theta_R \mapsto \mu = XA'\theta_R$ on $\theta_R \in \Theta_R$ is invariant with respect to the group $g_R : \theta_R \mapsto \theta_R + R_\perp a_\perp \zeta_3$ with φ_1 and φ_2 as maximal invariants.

A special case arises if the restriction combines a restriction on A with ad hoc identification. That is, if $R = (Ad, C)$ where $d \in \mathbb{R}^{p\times(p-p_R)}$ and $C \in \mathbb{R}^{q\times\{(q-p)-(q_R-p_R)\}}$ so (A, C) has full column rank. Then $b = d_\perp$ so that the restriction (A.18) *reduces to $d'\xi = (I_{p-p_R}, 0)\rho$.*

Proof of Theorem A.3. Apply Lemma A.1 (i, iii) to see that $\operatorname{rank}(A, R) = p_R + q - q_R$ is equivalent to $\operatorname{rank}(R_\perp' A) = p_R$. We can then write $R_\perp' A = ab'$ for some matrices $a \in \mathbb{R}^{q_R\times p_R}$ and $b \in \mathbb{R}^{p\times p_R}$ with full column rank.

Exploit the projection identity $I_q = R_\perp \overline{R}_\perp' + \overline{R}R'$ to get $\xi = A'\theta_R = A'R_\perp \overline{R}_\perp' \theta_R + A'\overline{R}R'\theta_R$. Insert $A'R_\perp = ba'$ and $R'\theta_R = \rho$ to get $\xi = ba'\overline{R}_\perp'\theta_R + A'\overline{R}\rho$ and therefore

$$\theta \longmapsto \mu = XA'\theta = Xba'\overline{R}_\perp'\theta_R + XA'\overline{R}\rho. \quad (A.22)$$

Inserting the orthogonal projection $I_p = b\overline{b}' + \overline{b}_\perp b_\perp'$ this can also be written as

$$\theta \longmapsto \mu = XA'\theta = Xb\left(a'\overline{R}_\perp'\theta_R + \overline{b}' A'\overline{R}\rho\right) + X\overline{b}_\perp b_\perp' A'\overline{R}\rho. \quad (A.23)$$

Noting that $\varphi_1 = a'\overline{R}_\perp'\theta_R$ is freely varying so is $\varphi_2 = a'\overline{R}_\perp'\theta_R + \overline{b}' A'\overline{R}\rho$. To rewrite φ_2 premultiply the first term by $I_{p_R} = \overline{b}'b$ to get $\varphi_2 = \overline{b}'(ba'\overline{R}_\perp'\theta_R + A'\overline{R}\rho)$. Then insert $\rho = R'\theta_R$ and $ba' = A'R_\perp$ to get $\varphi_2 = \overline{b}' A'(R\overline{R}_\perp' + \overline{R}R')\theta_R = \overline{b}' A'\theta_R = \overline{b}' A'\xi$. This gives the space M_R in (A.21).

To derive the reduced group g_R choose a $\theta_R \in \Theta_R$ so that $R'\theta_R = \rho$. Any $\theta_R^\dagger \in \Theta$ can be written as $\theta_R^\dagger = \theta_R + \overline{R}\zeta_1 + R_\perp \overline{a}\zeta_2 + R_\perp a_\perp \zeta_3$ since $(\overline{R}, R_\perp)$ and $(\overline{a}, a_\perp)$ have full rank. Since $\theta_R \in \Theta_R$ then $\theta_R^\dagger \in \Theta_R$ if and only if $\zeta_1 = 0$. We now consider whether θ_R, $\theta_R^\dagger$ are equivalent with respect to the restricted mapping (A.22). It holds

$$\theta_R^\dagger \longmapsto \mu^\dagger = Xba'\overline{R}_\perp'\left(\theta_R + R_\perp \overline{a}\zeta_2 + R_\perp a_\perp \zeta_3\right) + XA'\overline{R}\rho. \quad (A.24)$$

Noting that $\mu = Xba'\overline{R}_\perp'\theta_R + XA'\overline{R}\rho$ and $a'\overline{R}_\perp' R_\perp \overline{a} = I_{p_R}$, while $a'\overline{R}_\perp' R_\perp a_\perp = 0$, then $\mu^\dagger = \mu + Xb\zeta_2$ which reduces to μ if and only if $\zeta_2 = 0$. Thus, the mapping $\theta_R \mapsto \mu$ on $\theta_R \in \Theta_R$ is invariant with respect to the group $g_R : \theta_R \mapsto \theta_R + R_\perp a_\perp \zeta_3$ with φ_1 and φ_2 as maximal invariants.

Now, suppose $R = (Ad, C)$. Then Lemma A.2 shows that $m = A_\perp' C$ has full column rank, while (A, C) has orthogonal complement $A_\perp m_\perp$. Thus, R has orthogonal complement

$$R_\perp = \left\{\left(\overline{A, C}\right)\begin{pmatrix} d_\perp \\ 0 \end{pmatrix}, (A, C)_\perp\right\} = \left\{\left(\overline{A, C}\right)\begin{pmatrix} I_p \\ 0 \end{pmatrix} d_\perp, A_\perp m_\perp\right\}. \quad (A.25)$$

In particular, it holds $A'R_\perp = (I_p, 0)(A, C)'R_\perp = (d_\perp, 0) = d_\perp(I_{p_R}, 0)$, which is denoted ba' in the general case, so that $b = d_\perp$. Consider the restriction (A.18). Here, $b_\perp'\xi = d'\xi$, while $b_\perp' A' = d'A' = (I_{p-p_R}, 0)(Ad, C)' = (I_{p-p_R}, 0)R'$ so that $b_\perp' A'\overline{R}\rho = (I_{p-p_R}, 0)\rho$. □

Conflict of Interests

The authors do not have any conflict of interests.

Acknowledgment

Comments from Andrew Hunt and María Dolores Martínez Miranda are gratefully acknowledged.

References

[1] A. Ornelas, M. Guillén, and M. Alcañiz, "Implications of unisex assumptions in the analysis of longevity for insurance portfolios," in *Modeling and Simulation in Engineering, Economics, and Management*, vol. 145 of *Lecture Notes in Business Information Processing*, pp. 99–107, 2011.

[2] S. F. Jarner and E. M. Kryger, "Modelling adult mortality in small populations: the SAINT model," *ASTIN Bulletin*, vol. 41, no. 2, pp. 377–418, 2011.

[3] R. D. Lee and L. R. Carter, "Modeling and forecasting U.S. mortality," *Journal of the American Statistical Association*, vol. 87, pp. 659–671, 1992.

[4] Y. Yang, W. J. Fu, and K. C. Land, "A methodological comparison of age-period-cohort models: the intrinsic estimator and conventional generalized linear models," *Sociological Methodology*, vol. 34, pp. 75–110, 2004.

[5] Y. Yang and K. C. Land, "Age-period-cohort analysis of repeated cross-section surveys: fixed or random effects?" *Sociological Methods & Research*, vol. 36, no. 3, pp. 297–326, 2008.

[6] C. Berzuini and D. Clayton, "Bayesian analysis of survival on multiple time scales," *Statistics in Medicine*, vol. 13, no. 8, pp. 823–838, 1994.

[7] A. Riebler and L. Held, "The analysis of heterogeneous time trends in multivariate age-period-cohort models," *Biostatistics*, vol. 11, no. 1, pp. 57–69, 2010.

[8] F. Girosi and G. King, *Demographic Forecasting*, Princeton University Press, Princeton, NJ, USA, 2008.

[9] E. Pitacco, M. Denuit, S. Haberman, and A. Olivieri, *Modelling Longevity Dynamics for Pensions and Annuity Business*, Oxford University Press, Oxford, UK, 2009.

[10] D. Kuang, B. Nielsen, and J. P. Nielsen, "Forecasting in an extended chain-ladder-type model," *Journal of Risk and Insurance*, vol. 78, no. 2, pp. 345–359, 2011.

[11] E. Coelho and L. C. Nunes, "Forecasting mortality in the event of a structural change," *Journal of the Royal Statistical Society A: Statistics in Society*, vol. 174, no. 3, pp. 713–736, 2011.

[12] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman & Hall, London, UK, 1974.

[13] O. Barndorff-Nielsen, *Information and Exponential Families*, John Wiley & Sons, Chichester, UK, 1978.

[14] S. R. Searle, *Matrix Algebra Useful for Statistics*, John Wiley & Sons, New York, NY, USA, 1982.

[15] B. Nielsen, "Power of tests for unit roots in the presence of a linear trend," *Oxford Bulletin of Economics and Statistics*, vol. 70, no. 5, pp. 619–644, 2008.

[16] D. J. Poirier, "Revising beliefs in nonidentified models," *Econometric Theory*, vol. 14, no. 4, pp. 483–509, 1998.

[17] A. F. M. Smith, "Bayes estimates in one way and two way models," *Biometrika*, vol. 60, no. 2, pp. 319–329, 1973.

[18] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, Chichester, UK, 2000.

[19] D. Kuang, B. Nielsen, and J. P. Nielsen, "Identification of the age-period-cohort model and the extended chain-ladder model," *Biometrika*, vol. 95, no. 4, pp. 979–986, 2008.

[20] N. Keiding, "Statistical inference in the Lexis diagram," *Philosophical Transactions of the Royal Society of London A*, vol. 332, no. 1627, pp. 487–509, 1990.

[21] S. E. Fienberg and W. M. Mason, "Identification and estimation of age-period-cohort models in the analysis of discrete archival data," *Sociological Methodology*, vol. 10, pp. 1–67, 1979.

[22] D. Clayton and E. Schifflers, "Models for temporal variation in cancer rates. II: age-period-cohort models," *Statistics in Medicine*, vol. 6, no. 4, pp. 469–481, 1987.

[23] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[24] M. D.M. Miranda, B. Nielsen, and J. P. Nielsen, "Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality," *Journal of the Royal Statistical Society A*, 2014.

[25] D. Kuang, B. Nielsen, and J. P. Nielsen, "Forecasting with the age-period-cohort model and the extended chain-ladder model," *Biometrika*, vol. 95, no. 4, pp. 987–991, 2008.

[26] P. D. England and R. J. Verrall, "Stochastic claims reserving in general insurance," *British Actuarial Journal*, vol. 8, pp. 519–544, 2002.

[27] B. Zehnwirth, "Probabilistic development factor models with applications to loss reserve variability, prediction intervals, and risk based capital," in *Proceedings of the Casualty Actuarial Society Forum*, pp. 447–605, Arlington, Va, USA, 1994.

[28] D. Kuang, B. Nielsen, and J. P. Nielsen, "Chain-ladder as maximum likelihood revisited," *Annals of Actuarial Science*, vol. 4, pp. 105–121, 2009.

[29] B. Carstensen, "Age-period-cohort models for the Lexis diagram," *Statistics in Medicine*, vol. 26, no. 15, pp. 3018–3045, 2007.

[30] D. Clayton and E. Schifflers, "Models for temporal variation in cancer rates. I: age-period and age-cohort models," *Statistics in Medicine*, vol. 6, no. 4, pp. 449–467, 1987.

[31] R. M. O'Brien, "Constrained estimators and age-period-cohort models," *Sociological Methods & Research*, vol. 40, no. 3, pp. 419–452, 2011.

[32] R. M. O'Brien, "Intrinsic estimators as constrained estimators in age-period-cohort accounting models," *Sociological Methods & Research*, vol. 40, no. 3, pp. 467–470, 2011.

[33] W. J. Fu, K. C. Land, and Y. Yang, "On the intrinsic estimator and constrained estimators in age-period-cohort models," *Sociological Methods & Research*, vol. 40, no. 3, pp. 453–466, 2011.

[34] L. L. Kupper, J. M. Janis, A. Karmous, and B. G. Greenberg, "Statistical age-period-cohort analysis: a review and critique," *Journal of Chronic Diseases*, vol. 38, no. 10, pp. 811–830, 1985.

[35] T. R. Holford, "An alternative approach to statistical age-period-cohort analysis," *Journal of Chronic Diseases*, vol. 38, no. 10, pp. 831–836, 1985.

[36] A. J. G. Cairns, D. Blake, K. Dowd, G. D. Coughlan, and M. Khalaf-Allah, "Bayesian stochastic mortality modeling for two populations," *ASTIN Bulletin*, vol. 41, no. 1, pp. 29–59, 2011.

[37] S. Johansen, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford, UK, 1995.

[38] B. Nielsen, "The likelihood-ratio test for rank in bivariate canonical correlation analysis," *Biometrika*, vol. 86, no. 2, pp. 279–288, 1999.

[39] B. Nielsen, "Conditional test for rank in bivariate canonical correlation analysis," *Biometrika*, vol. 88, no. 3, pp. 874–880, 2001.

[40] A. E. Renshaw and S. Haberman, "A cohort-based extension to the Lee-Carter model for mortality reduction factors," *Insurance: Mathematics and Economics*, vol. 38, no. 3, pp. 556–570, 2006.

[41] A. J. G. Cairns, B. David, K. Dowd et al., "A quantitative comparison of stochastic mortality models using data from England and wales and the United States," *North American Actuarial Journal*, vol. 13, no. 1, pp. 1–35, 2009.

[42] C. Pedroza, "A Bayesian forecasting model: predicting U.S. male mortality," *Biostatistics*, vol. 7, no. 4, pp. 530–550, 2006.

[43] R. F. Engle and C. W. Granger, "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, vol. 55, pp. 251–276, 1987.

[44] D. F. Hendry and B. Nielsen, *Econometric Modeling*, Princeton University Press, Princeton, NJ, USA, 2007.

[45] M. A. S. Cabrera, S. M. de Andrade, and R. M. Dip, "Lipids and all-cause mortality among older adults: a 12-year follow-up study," *TheScientificWorldJOURNAL*, vol. 2012, Article ID 930139, 5 pages, 2012.

[46] G. H. Golub and C. F. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Md, USA, 1989.