Identification and forecasting in the Lee-Carter model

B. Nielsen
Nuffield College, Oxford OX1 1NF, U.K.
bent.nielsen@nuffield.ox.ac.uk
and J.P. Nielsen
Cass Business School, City University London, 106 Bunhill Row, London EC1Y 8TZ, U.K.
Jens.Nielsen.1@city.ac.uk
8 December 2010

Summary: We consider the identification problem for the model of Lee and Carter (1992). The parameters of this model are known only to be identified up to certain transformations. Forecasts from the model may therefore depend on the arbitrarily chosen identification scheme. A condition for invariant forecasts is proposed. A number of standard forecast models are analyzed.
Keywords: Age-period-cohort model; Cointegration; Forecasting; Identification; Lee-Carter model; Multi-sample problem.

1 Introduction

The model of Lee and Carter (1992) is used widely in demography, actuarial mathematics, and social sciences. It describes the logarithm of mortality through two interlinked time scales

$$\mu_{x,t} = \alpha_x + \beta_x \kappa_t, \tag{1}$$

where $x = 1, \ldots, X$ is the age and $t = 1, \ldots, T$ is the calendar time. Lee and Carter (1992) recognised from the outset that this paramerization is not identified and suggested a particular identification scheme. An important part of the model is that it is easy to extrapolate the time-varying parameter κ_t in order to make forecasts of future mortality. Other identification schemes could, however, be used. It is therefore important to chose an extrapolation method, so that the forecast of future mortality does not depend on the particular identification scheme.

This type of issue is common for over-parametrized models where it is of interest to interpret and forecast the parameters involved. The problem has recently been analysed for the age-period-cohort model by Kuang, Nielsen

and Nielsen (2008a,b, 2010). In that work it is characterized which plots of the estimated age-period-cohort parameters are meaningful and which are not meaningful and it is characterized which forecast methods are invariant to arbitrary identification schemes. Here the same analysis is made for the Lee-Carter model. It is found that when the calendar parameter is identified by an arbitrary identification scheme such as that proposed by Lee and Carter then forecasts must be location-scale preserving. It is interesting to note that the issues that arise for the age-period-cohort model forecasts of arbitrarily identified parameters have to be linear-trend preserving. The implications for applied work are therefore somewhat different for the Lee-Carter model and the age-period-cohort model.

In practice Lee-Carter models are often applied to different data sets such as for women and men or for different countries. It is then of interest to compare the estimated calendar effects from different data sets. Due to the identification issue this has to be done with some care.

2 Identification

Lee and Carter (1992) describe the over-parametrization issue as follows. Suppose that a set of parameters

$$\theta = (\alpha_1, \ldots, \alpha_X, \beta_1, \ldots, \beta_X, \kappa_1, \ldots, \kappa_T) \in \mathbb{R}^{2X+T}$$

is given. Then for any scalar c and any scalar $d \neq 0$ it holds that

$$\mu_{x,t} = \alpha_x + \beta_x \kappa_t = (\alpha_x - \beta_x c) + (\frac{\beta_x}{d}) \{d(\kappa_t + c)\}. \tag{2}$$

This shows that the parametrization θ is equivalent to a parametrization $\theta^\dagger$ where $\alpha_x^\dagger = \alpha_x - \beta_x c$, $\beta_x^\dagger = \beta_x / d$, $\kappa_t^\dagger = d(\kappa_t + c)$. There are two approaches to address this lack of identification. The first approach is to construct a new parametrization which is invariant to the identification problem. Such a parametrization is suggested below. This parametrization can be estimated, interpreted and extrapolated freely without any worries about the identification issue. The second approach, which is used in the literature, is to choose some arbitrarily chosen identification scheme which also solves the estimation problem. However, interpretation and extrapolation have to be done with great care to ensure that this is invariant to the arbitrarily chosen identification scheme.

The standard identification choice is that of Lee and Carter (1992). They suggested a parameter θ°, say, by the identifying constraints $\sum_{x=1}^X \beta_x^\circ = 1$ and $\sum_{t=1}^T \kappa_t^\circ = 0$, which implies that α_x° has interpretation as the average over time of $\mu_{x,t}$, that is $\alpha_x^\circ = T^{-1}\sum_{t=1}^T \mu_{x,t}$.

A noteworthy alternative identification scheme was also mentioned by Lee and Carter (1992, §6). This is to let $\alpha_x^* = \mu_{x,T}$, which comes about by choosing $\beta_1^* = 1$ and $\kappa_T^* = 0$. This parametrisation can be derived from the standard Lee-Carter identification using the relations (2). With the standard identification it holds $\mu_{1,T} = \alpha_1^\circ + \beta_1^\circ\kappa_T^\circ$ while the second identification gives $\mu_{1,T} = \alpha_1^* + \beta_1^*\kappa_T^* = \alpha_1^*$ where $\alpha_1^* = \alpha_1^\circ - \beta_1^\circ c$, which shows that $c = -\kappa_T^\circ$. Moreover, the second identification gives $\beta_1^* = 1$ where $\beta_1^* = \beta_1^\circ/d$ showing that $d = \beta_1^\circ$. The opposite transformation back to the standard Lee-Carter identification is then given by $c^\dagger, d^\dagger$ solving similar equations. First, by the Lee-Carter identification $1 = \sum_{x=1}^X \beta_x^\circ$ where $\beta_x^\circ = \beta_x^*/d^\dagger$ showing $d^\dagger = \sum_{x=1}^X \beta_x^*$. Secondly, by the Lee-Carter identification $\alpha_1^\circ = T^{-1}\sum_{t=1}^T \mu_{1,t} = \alpha_1^* - \beta_1^* c^\dagger$ while $T^{-1}\sum_{t=1}^T \mu_{1,t} = \alpha_1^* + \beta_1^* T^{-1}\sum_{t=1}^T \kappa_t^*$ with $\beta_1^\circ = 1$ showing that $c^\dagger = -T^{-1}\sum_{t=1}^T \kappa_t^*$.

In the following it is checked that the standard Lee-Carter identification identifies the parameters uniquely. The proof is inspired by Kuang, Nielsen and Nielsen (2008a) and is given in the appendix.

Theorem 1 *Let $\mu = (\mu_{x,T}, x = 1,\ldots,X, t = 1,\ldots,T)$, where $\mu_{x,t}$ satisfies (1) for some θ. Then the standard Lee-Carter parametrization θ° where $\sum_{x=1}^X \beta_x^\circ = 1$ and $\sum_{t=1}^T \kappa_t^\circ = 0$, satisfies*

(i) θ° is a function of θ.

(ii) μ is a function of θ through θ°.

(iii) The parametrization of μ by θ° is exactly identified. That is, if $\theta^\dagger \neq \theta^\ddagger$ are two parameters satisfying the standard Lee-Carter identification then $\mu(\theta^\dagger) \neq \mu(\theta^\ddagger)$.

(iv) Equivalent results could be formulated for parametrizations $\theta^\dagger$ implied by (2) for arbitrary choices of c and $d \neq 0$.

The transformation (2) shows that a graph of estimated calendar parameters $\widehat{\kappa}_t$ can be interpreted up to scale and location. In other words the y-axis of the graph does not have any particular meaning. This is not a problem

when looking at a single graph from a single data set. As will be discussed below, problems can arise when comparing estimates from different samples or when extrapolating.

The age-period model, $\mu_{x,t} = \alpha_x + \beta_t$, and the age-period-cohort model, $\mu_{x,t} = \alpha_x + \beta_t + \gamma_{x-t} + \delta$, have related but different interpretational issues as discussed in Kuang, Nielsen and Nielsen (2008a). In the age-period model parameters are determined up to their location since $\mu_{x,t} = (\alpha_x + c) + (\beta_t - c)$. This means that the estimated scale is comparable accross datasets, whereas the levels are not. For the age-period-cohort model parameters are determined up to linear trends since $\mu_{x,t} = (\alpha_x + a - dx) + (\beta_t + b + dt) + \{\gamma_{x-t} + c + d(x-t)\} + (\delta - a - b - c)$. In that case parameters are only interpretable up to an arbitrary linear trend, which in practice makes them uninterpretable even when looking at a single data set.

The arbitrariness of the above mentioned parametrizations can be avoided through a parametrization which is invariant to the transformations outlined in (2). A parametrization that is a maximal invariant under those transformations can be constructed by choosing a suitable $(2X+T-2)$-dimensional subset of the parameters $\mu_{t,x}$. To do this note that the model expression (1) implies that

$$\frac{\beta_x}{\beta_1} = \frac{\mu_{x,T} - \mu_{x,1}}{\mu_{1,T} - \mu_{1,1}}, \quad \beta_1(\kappa_t - \kappa_1) = \mu_{1,t} - \mu_{1,1}, \quad \alpha_x + \beta_x \kappa_1 = \mu_{x,1}. \tag{3}$$

It then follows that the original parameter $\mu_{x,t}$ satisfies the relation

$$\mu_{x,t} = \mu_{x,1} + \frac{\mu_{x,T} - \mu_{x,1}}{\mu_{1,T} - \mu_{1,1}}(\mu_{1,t} - \mu_{1,1}). \tag{4}$$

This shows that $\mu_{x,t}$ is a function of

$$\mu_{1,1}, \ldots, \mu_{X,1}, \mu_{1,T} - \mu_{1,1}, \ldots, \mu_{X,T} - \mu_{X,T}, \mu_{1,2} - \mu_{1,1}, \ldots, \mu_{1,T-1} - \mu_{1,1},$$

or equivalently of the parameter vector

$$\xi = (\mu_{1,1}, \ldots, \mu_{X,1}, \mu_{1,T}, \ldots, \mu_{X,T}, \mu_{1,2}, \ldots, \mu_{1,T-1})' \in \mathbb{R}^{2X+T-2}. \tag{5}$$

As the invariance issue only relates to the parameters θ and not to the parameter μ from which ξ is defined then ξ is invariant. The parameter ξ is also a maximal invariant since for $\xi^\dagger \neq \xi^\ddagger$ then $\mu(\xi^\dagger) \neq \mu(\xi^\ddagger)$. Therefore ξ is the basis for a unique parametrization of μ/ These results are summarized as follows.

Theorem 2 *The parameter $\xi \in \mathbb{R}^{2X+T-2}$ is a maximal invariant of θ under the transformations (2) and satisfies*

1. *for any θ then μ is a function of θ through ξ due to (4),*
2. *for any $\xi \neq \xi^\dagger$ then $\mu(\xi) \neq \mu(\xi^\dagger)$.*

Since ξ is a maximal invariant any plot based on ξ is meaningful. A plot of the calendar effect with a meaningful scale can therefore be done by plotting $\mu_{1,t}$ or, equivalently, $\mu_{x,t}$. The plot of $\mu_{x,t}$ will show the development in mortality over time for individuals of age x. Alternatively, a plot of $\mu_{x,t} - \mu_{x,1} = \beta_x(\kappa_t - \kappa_1)$ will avoid the parameter $\alpha_{x,t}$. Such plots based on different data sets are comparable accross datasets.

Parameters following a particular identification scheme can be recovered from the invariant parametrization. With the identification $\beta_1^* = 1$ and $\kappa_T^* = 0$ the individual parameters are recovered by

$$\alpha_x^* = \mu_{x,T}, \quad \beta_x^* = \frac{\mu_{x,T} - \mu_{x,1}}{\mu_{1,T} - \mu_{1,1}}, \quad \kappa_t^* = \mu_{1,t} - \mu_{1,1}.$$

Likewise, with the Lee-Carter identification $\sum_{x=1}^X \beta_x^\circ = 1$ and $\sum_{t=1}^T \kappa_t^\circ = 0$ then $\alpha_x^\circ = \mu_{x,t} - \beta_x^\circ \kappa_t^\circ$ where

$$\beta_x^\circ = \frac{\mu_{x,2} - \mu_{x,1}}{\sum_{i=1}^X (\mu_{i,2} - \mu_{1,i})}, \quad \kappa_t^\circ = \frac{(\mu_{x,t} - \mu_{x,1}) - \frac{1}{T}\sum_{j=1}^T (\mu_{x,j} - \mu_{x,1})}{\beta_x^\circ}.$$

A consequence of the latter derivation is that Lee-Carter β_x°-parameters will be differ by a scale factor when applied to a full data set on the one hand and a sub-set of the data only including ages up to age X_0 say. This in turn implies that the κ_t°-parameters will differ by a scale factor in the full data and in the sub-set. This effect will also be found when comparing two unrelated data sets.

3 Forecasting

Suppose now that an estimate $\widehat{\theta}$ is available for a particular identification scheme for the original parameter θ. The aim is to forecast $\mu_{x,t}$ for some $t = T + h$ beyond the last observed period T. This is done by extrapolating the time-varying parameter estimates $\widehat{\kappa}_t$ into $\tilde{\kappa}_{T+h}$ and then constructing

the forecast as $\tilde{\mu}_{x,T+h}(\widehat{\theta}) = \widehat{\alpha}_x + \widehat{\beta}_x\tilde{\kappa}_{T+h}$. The forecast is age-specific which Lee and Carter (1992) saw as an advantage although it comes about in a somewhat uncontrolled fashion which has prompted extentions and variations of the model in various directions, see for instance Girosi and King (2008). However, the question of interest here is whether the forecast depends on the chosen parametrization. Ideally the forecast should be invariant to the parametrization but this invariance will depend on the choice of forecasting method. Two approaches that give invariant forecasts are discussed in the following.

The first approach is to base forecasts on the maximal invariant ξ given in (5). Suppose an estimate $\widehat{\xi}$ is available which implies a series $\widehat{\mu}_{1,t}$ for $t = 1, \ldots, T$. Since this series is invariant to the transformations in (2) then an h-step ahead forecast $\tilde{\mu}_{1,T+h}$ based on $\widehat{\mu}_{1,t}$ for $t = 1, \ldots, T$ will also be invariant regardless of the forecasting method applied.

The second approach is to use an identification such as the Lee-Carter identification and choose forecasts which are invariant. To characterize the invariant forecast methods transform θ into $\theta^\dagger$ by (2) and apply the forecast chosen rule, $\tilde{\kappa}_{T+h}(\widehat{\kappa}^\dagger)$, that is

$$\tilde{\mu}_{x,T+h}(\widehat{\theta^\dagger}) \;=\; \widehat{\alpha}^\dagger_x + \widehat{\beta}^\dagger_x\tilde{\kappa}_{T+h}(\widehat{\kappa}^\dagger) = (\widehat{\alpha}_x - \widehat{\beta}_x c) + (\frac{\widehat{\beta}_x}{d})\tilde{\kappa}_{T+h}\{d(\widehat{\kappa} + c)\}.$$

This forecast will be invariant to the chosen identification scheme when this expression does not depend on the arbitrary constants c, d. A condition for invariance then follows.

Theorem 3 *The forecast $\tilde{\mu}x, T + h$ is invariant to the identification scheme if and only if for arbitrary c and $d \neq 0$*

$$\tilde{\kappa}_{T+h}\{d(\widehat{\kappa} + c)\} = d\tilde{\kappa}_{T+h}(\widehat{\kappa}) + dc.$$

In other words, the forecast method for the time-varying component κ_t must be location-scale preserving.

It is interesting to compare this result with the corresponding result for the age-period-cohort model in Kuang, Nielsen and Nielsen (2008b). For that model forecasts of for instance the period parameter must be location and trend preserving, but need not preserve the scale.

Fortunately there are many location-scale preserving forecasts. First of all, the forecast method preferred by Lee and Carter (1992) is location-scale

preserving. This is a random walk with an estimated intercept:

$$\tilde{\kappa}_{T+h} = \tilde{\kappa}_{T+h-1} + \nu_c + \varepsilon_h,$$

where ν_c is estimated by the average of growth rates $\widehat{\nu}_c = (T-1)^{-1}\sum_{t=2}^{T}(\widehat{\kappa}_t - \widehat{\kappa}_{t-1})$. For a moment let $\varepsilon_h = 0$ so as to focus on point forecasting. The point forecast is a linear trend: $\tilde{\kappa}_{T+h} = \widehat{\kappa}_T + \widehat{\nu}_c h$. This is location-scale preserving since $\widehat{\kappa}_T$ is location-scale preserving and $\widehat{\nu}_c$ is location invariant, that is $\tilde{\kappa}_{T+h}\{d(\widehat{\kappa}+c)\} = d(\widehat{\kappa}_T + c) + d\widehat{\nu}_c h = d\tilde{\kappa}_{T+h}(\widehat{\kappa}) + dc$. Note that the slope of line, in h, is $d\widehat{\nu}_c$ which is proportional to the arbitrarily chosen scale coefficient d. The coefficient d or equivalently the scaling of β_x and hence of κ_t may be different for different samples and have to be handled with care as discussed in §4.

Theorem 3 also covers distribution forecasts. To see this include random innovations ε_h in the derivations above to generate a random walk around the linear trend. Suppose ε_h is chosen as zero-mean normal with variance $\widehat{\sigma}^2(\widehat{\kappa}) = (T-2)^{-1}\sum_{t=2}^{T}(\widehat{\kappa}_t - \widehat{\kappa}_{t-1} - \widehat{\nu}_c)^2$. Then it holds that the standard error is location invariant and scale preserving in that $\widehat{\sigma}\{d(\widehat{\kappa}+c)\} = d\widehat{\sigma}(\widehat{\kappa})$. When combined with the linear trend which is location-scale preserving the overall forecast is location-scale preserving.

For the above Lee-Carter forecast the intercept may be substantially important in generating a linear trend. When evaluating this forecast in terms of its location-scale preserving properties the intercept is, however, not crucial. The pure random walk model $\tilde{\kappa}_{T+h} = \tilde{\kappa}_{T+h-1} + \varepsilon_h$ would also be location-scale preserving. Both of these methods evolve around random walks and are characterised as I(1)-methods in the econometric literature. If the estimates $\widehat{\kappa}_t$ are trending but not exactly linearly trending it may be appropriate to forecast using a cummulated random random walk which extrapolates the linear trend in the last two estimates $\widehat{\kappa}_{T-1}$, $\widehat{\kappa}_T$. This would be called an I(2)-method.

Methods that are based on stationary time series can also be location-scale preserving. Such methods are called I(0)-methods. For these methods it is important to include an intercept. Some simple examples of location-scale preserving forecasts are the average $\overline{\kappa} = T^{-1}\sum_{t=1}^{T}\widehat{\kappa}_t$ and the estimate for the last observed calendar time $\widehat{\kappa}_T$, as well as a linear trend fitted to the estimates $\widehat{\kappa}$ by least squares.

A simple autoregression without an intercept will, however, not preserve location-scale. In its simplest form let $\tilde{\kappa}_{T+1} = \rho\widehat{\kappa}_T$ for some $\rho \neq 1$. This

could come about if the time-varying series were considered on the original scale and it were desirable to forecast $\exp(\kappa_{T+1})$ by $\{\exp(\widehat{\kappa}_T)\}^\rho$. For this forecast it holds that $\tilde{\kappa}_{T+1}\{d(\widehat{\kappa}+c)\}\rho d(\widehat{\kappa}_T+c) = d\rho\widehat{\kappa}_T + \rho c \neq d\rho\widehat{\kappa}_T + \rho c$. In practice one might estimate ρ by the least squares coefficient $\widehat{\rho} = \sum_{t=2}^T \widehat{\kappa}_t\widehat{\kappa}_{t-1}/\sum_{t=2}^T \widehat{\kappa}_{t-1}^2$ which would in general be different from unity and therefore have the invariance problem. Another forecast that is not location-scale preserving would be $\tilde{\kappa}_{T+1} = \widehat{\kappa}_T^\delta$.

An autoregression with an intercept will, however, be location-scale preserving. That is

$$\tilde{\kappa}_{T+h} = \rho\tilde{\kappa}_{T+h-1} + \nu_c + \varepsilon_h,$$

where ρ and ν_c are estimated by the least squares method. To see this define $\overline{\kappa} = (T-1)^{-1}\sum_{t=2}^T \widehat{\kappa}_t$ and $\overline{\kappa}_{-1} = (T-1)^{-1}\sum_{t=2}^T \widehat{\kappa}_{t-1}$. Then ρ is estimated by $\widehat{\rho} = \sum_{t=2}^T \widehat{\kappa}_t(\widehat{\kappa}_{t-1}-\overline{\kappa}_{-1})/\sum_{t=2}^T(\widehat{\kappa}_{t-1}-\overline{\kappa}_{-1})^2$ which is location-scale invariant while ν_c is estimated by $\overline{\kappa} - \widehat{\rho}\overline{\kappa}_{-1}$ which becomes $d(\overline{\kappa} - \widehat{\rho}\overline{\kappa}_{-1}) + d(1-\widehat{\rho})c$ when transforming by (2). The expression of Theorem 3 is then, for a one-step-ahead point forecast,

$$\begin{aligned}\tilde{\kappa}_{T+1}(d\kappa+c) &= \widehat{\rho}\{d(\widehat{\kappa}_{T-1}+c)\} + d(\overline{\kappa} - \widehat{\rho}\overline{\kappa}_{-1}) + d(1-\widehat{\rho})c \\ &= d(\widehat{\rho}\widehat{\kappa}_{T-1} + \overline{\kappa} - \widehat{\rho}\overline{\kappa}_{-1}) + c\end{aligned}$$

as desired.

It is popular to use autoregressive integrated moving average (ARIMA) models in the forecasting. As long as an intercept is included the theoretical maximum likelihood estimators will be location-scale preserving. In practice these models are estimated using numerical algorithms that depend on for instance the chosen convergence criteria. The standard advice in such situation is to seek to standardize data and orthogonalize regressors. However, with the Lee-Carter identification the scale is deliberately chosen to be rather extreme so that the numerical accuracy may be poor. The bottom line is that the ARIMA estimation algorithms are only location-scale preserving up to an approximation. We compared algorithms in R and OX in this respect and found that the the OX algorithm based on Doornik and Ooms (2003) is more precise but still with problems with respect to numerical accuracy when used uncritically. A further development is to use ARIMA models with structural breaks as in Coelho and Nunes (2011). These will be location-scale preserving up to the accuracy of the estimation algorithm.

The results are summarised in Table 1. Clements and Hendry (1999, §5) discuss the relative merits of I(0), I(1) and I(2)-forecasting methods in a

Table 1: Invariance properties of various forecasting models

Order of integration	Invariant forecasts	Non-invariant forecasts
I(0)	$x_t = \nu_c + \varepsilon_t$	$x_t = \varepsilon_t$
	$x_t = \nu_c + \nu_l t + \varepsilon_t$	
	$x_t = \rho x_{t-1} + \nu_c + \varepsilon_t$	$x_t = \rho x_{t-1} + \varepsilon_t$
I(1)	$\Delta x_t = \nu_c + \varepsilon_t$	
	$\Delta x_t = \varepsilon_t$	

standard time series context. They are concerned with possible structural changes near the end of the sample or in the beginning of the forecast period which can be detrimental to forecasts. The I(0)-methods tend to be preferable if they describe the sample variation in-sample and structural changes are neither observed at the end of sample nor expected out-of-sample, whereas the higher order integrated methods tend to be more robust to structural changes out-of-sample. The same issues arise when forecasting with the age-period-cohort model as shown in Kuang, Nielsen and Nielsen (2008b). Kuang, Nielsen and Nielsen (2010) discuss the relative merit of the different forecasting methods in the context of insurance data using an extended chain ladder model which is a variant of the age-period-cohort model.

4 Multi-sample problem

In applications it is often of interest to compare the development of mortality in muliple populations such as women/men or accross countries. One popular approach is to perform a Lee-Carter analysis separately on each data set and then compare the estimates of the time-varying parameters. It has been found that the relative performance of such graphs accross samples does not reflect what is otherwise known about the substantial context. Li and Lee (2005) attributes this finding to an unpublished manuscript by Lee and Nault from 1993 and suggest that the different samples are modelled jointly so that the calendar effect for women/men are viewed as deviations from an population average.

It is, however, not impossible to compare estimates from separate Lee-Carter analyses. An increasingly popular approach is to fit a vector autoregression to the estimates $\widehat{\kappa}_{i,t}$ from the two or more samples and apply a cointegration analysis. The motivation would be that the series $\widehat{\kappa}_{i,t}$ for

$t = 1, \ldots, T$ share a common random walk trend as in the Lee-Carter forecast, while the individual series from the different samples deviate from that common trend in a stationary fashion. Such an approach was suggested by Renshaw and Haberman in a discussion paper from 2003 and followed up by Lazar and Denuit (2009). Since the scale and level of the series $\widehat{\kappa}_{i,t}$ for $t = 1, \ldots, T$ are not defined cointegration analysis should be done with some care. Using the above analysis the potential problems can be addressed.

Cointegration analysis of vector autoregressions has been suggested by Johansen (1988, 1995), see also Hendry and Nielsen (2007). The first step of a cointegration analysis is to determine the cointegration rank. Collect first the estimates $\widehat{\kappa}_{i,t}$ from samples $i = 1, \ldots, p$ as a vector k_t, say. Various choices can be made with respect to the specification of the vector autoregression. The choice matching the Lee-Carter forecast would be

$$\Delta k_t = \Pi \begin{pmatrix} k_{t-1} \\ t \end{pmatrix} + \nu + \varepsilon_t, \qquad t = 2, \ldots, T,$$

where $\Delta k_t = k_t - k_{t-1}$ and $\Pi \in \mathbb{R}^{p \times (p+1)}$. The cointegration rank $r = \text{rank}(\Pi)$ is determined by first finding the residuals from regressing Δk_t and $(k'_{t-1}, t)'$, on an intercept and then finding the squared canonical correlations of these residuals. The squared canonical correlations are invariant to the undetermined level and scale of the series k_t, see discussion of Nielsen and Rahbek (2000). The asymptotic distribution of the relevant rank test statistic is non-standard and tabulated by Johansen (1995).

When the matrix Π is found to have reduced rank r it can be written as $\Pi = \alpha(\beta', \delta')$ where $\alpha, \beta \in \mathbb{R}^{p \times r}$ and $\delta \in \mathbf{R}^{1 \times r}$. The Granger-Johansen representation shows that $\kappa_t = C \sum_{s=2}^{t} \varepsilon_s + Y_t + \tau_c + \tau_l t$, where C is a function of the dynamic parameters with the property that $\beta' C = 0$, Y_t is a mean-zero stationary process, and τ_c, τ_l depend on parameters and initial observations in such a way that $\beta' \kappa_t + \delta' t$ is stationary around a constant level.

The cointegrated model where $\Pi = \alpha(\beta', \delta')$ has an identification problem of its own in that α, $(\beta', \delta')'$ are not identified, only the linear vector spaces spanned by them are identified. This is a point that is dealt with by Johansen, who suggests just-identified estimates of these parameters based on the canonical correlation method.

When applying the cointegration model to the time-varying estimates k_t a different identification issue arises. Since the scales of the series $k_{i,t}$ are not identified then the linear spans of α, β are not invariant to the scales

of $k_{i,t}$. This lack of invariance is not detrimental in that Johansen's just-identified estimates combine in such a way that predictions from the model are location-scale preserving as required in Theorem 3.

Invariance problems only arise when interpreting or restricting the cointegration parameters. In the context of two samples representing women and men it would be tempting to impose the restriction $\beta = (1, -1)'$. The idea is that the difference between the series for women and men should be stationary. Since the linear spans of α, β depend on the scaling of the series $k_{i,t}$ this will however result in predictions that are no longer location-scale preserving.

There is one restriction that can be imposed without problems. This is the restriction that the linear trend parameter is zero, $\delta = 0$, so that stationary variation for women/men around the common random walk trend is not following a linear trend. In practice this restriction ensures that the different populations are not drifting apart.

The estimated cointegration model for the time-varying parameter k_t can be extrapolated beyond the last observed period T and combined with estimates of the age-dependent parameters to provide a forecast of the future mortality. If the applied time series model is stable within sample and continues to be correct out of sample then simple predictions from the time series model will result in forecasts with good properties. However, if there is a structural change in the level of the time-dependent parameter k_t near the end of the sample or in the begining of the forecast period then forecasts will have poor properties. Hendry (2006) discusses how to forecast from cointegration models in the presence of level shifts. Suppose the estimated model satisfies the linear trend restriction $\delta = 0$ so

$$\Delta k_t = \widehat{\alpha}\widehat{\beta} k_{t-1} + \widehat{\nu} + \widehat{\varepsilon}_t.$$

In face of structural shifts near the last sample period T a robust forecast is achieved by leaving the cointegration approach aside and double differencing the time-varying parameter k_t through the forecast equation $\Delta\tilde{k}_{T+h} = \Delta\tilde{k}_{T+h-1}$. As discussed above this idea approach has been found to be favourable in the context of out-of-sample level shifts by Clements and Hendry (1999, §5). The reason this works is that it gets around the problem that shifts to the deterministic term ν are rather pernicious to forecasting. Hendry (2006) discusses how to forecast robustly in the light of the cointe-

gration model. The idea is to difference the cointegration equation as

$$\Delta\widetilde{k}_{T+h} = \Delta\widetilde{k}_{T+h-1} + \widehat{\alpha}\widehat{\beta}\Delta\widetilde{k}_{T+h-1}.$$

This forecast is found to perform favourably to double differencing in that forecast variances do not build up. Comparing to the double differencing equation the difference is inclusion of the term $\widehat{\alpha}\widehat{\beta}\Delta\widetilde{k}_{T+h-1}$ which has zero mean and is therefore robust to level shifts and which also captures the dynamic term that was found to be of significance in-sample.

A different type of structural breaks are those happening in the middle of the sample corresponding to wars or variations in economic conditions. In contrast to breaks around the last sample period, T, those breaks can be modelled more actively. Johansen, Mosconi and Nielsen (2000) have suggested a cointegration model that allows for such changes in the slope of the linear trend.

Appendix

Proof of Theorem 1

(i) For any θ then construct θ_{LC} by $d = \sum_{x=1}^{X} \beta_x$ and $c = -T^{-1}\sum_{t=1}^{T} \kappa_t$

(ii) Use that one can transform θ_{LC} into the original θ by $d_{LC} = 1/d$ and $c_{LC} = -cd$ and that the parametrisation (2) is invariant to c, d.

(iii) consider $\theta^\dagger \neq \theta^\ddagger$.

If $\alpha_x^\dagger \neq \alpha_x^\ddagger$ for some x then $T^{-1}\sum_{t=1}^{T} \mu_{x,t}^\dagger = \alpha_x^\dagger \neq \alpha_x^\ddagger = T^{-1}\sum_{t=1}^{T} \mu_{x,t}^\ddagger$.

If $\alpha_x^\dagger = \alpha_x^\ddagger$ for all x but $\kappa_t^\dagger \neq \kappa_t^\ddagger$ for some t then, since $\sum_{x=1}^{X} \beta_x = 1$, it holds $\sum_{x=1}^{X} \mu_{x,t}^\dagger = \kappa_t^\dagger - \sum_{x=1}^{X} \alpha_x^\dagger \neq \kappa_t^\ddagger - \sum_{x=1}^{X} \alpha_x^\ddagger = \sum_{x=1}^{X} \mu_{x,t}^\ddagger$

If $\alpha_x^\dagger = \alpha_x^\ddagger$ for all x and $\kappa_t^\dagger = \kappa_t^\ddagger$ for all t but $\beta_x^\dagger \neq \beta_x^\ddagger$ for some x then $\mu_{x,2}^\dagger - \mu_{x,1}^\dagger = \beta_x^\dagger(\kappa_2^\dagger - \kappa_1^\dagger) \neq \beta_x^\ddagger(\kappa_2^\ddagger - \kappa_1^\ddagger) = \mu_{x,2}^\ddagger - \mu_{x,1}^\ddagger$.

(iv) Exploit the relation (2).

References

Clements, M.P. and Hendry, D.F. (1999) *Forecasting non-stationary time series* Cambridge, MA: MIT Press.

Coelho, E. and Nunes, L.C. (2011) Forecasting mortality in the event of a structural change. To appear in *Journal of the Royal Statistical Society* A.

Doornik, J.A. and Ooms, M. (2003) Computational Aspects of Maximum Likelihood Estimation of Autoregressive Fractionally Integrated Moving Average Models. *Computational Statistics and Data Analysis* 42, 333–348.

Girosi, F. and King, G. (2008) *Demographic forecasting*. Princeton: Princeton University Press.

Hendry, D.F. (2006) Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics* 135, 399–426.

Hendry, D.F. and Nielsen, B. (2007) *Econometric Modeling*. Princeton: Princeton University Press.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–254.

Johansen, S. (1995) *Likelihood-based inference in cointegrated vector autoregressive models* Oxford: Oxford University Press.

Johansen, S., Mosconi, R. and Nielsen, B. (2000) Cointegration analysis in the presence of structural breaks in the deterministic trend. *Econometrics Journal* 3, 216–249.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008a) Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* 95, 979–986.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008b) Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* 95, 987–991.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2010) Forecasting in an extended chain-ladder-type model. To appear in *Journal of Risk and Insurance*. See http://www.nuffield.ox.ac.uk/economics/papers/2010/w5/Forecast24jun10.pdf

Lazar, D. and Denuit, M.M. (2009) A multivariate time series approach to projected life tables. *Applied Stochastic Models in Business and Industry* 25, 806–823.

Lee, R.D. and Carter, L.R. (1992) Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87, 659–671.

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

Nielsen, B. and Rahbek, A. (2000) Similarity issues in cointegration models. *Oxford Bulletin of Economics and Statistics* 62, 5–22.