Part I.
Existing Methods for Forecasting Mortality

In Part I, we survey some of the best methods currently available for forecasting mortality by scholars—including demographers, public health researchers, economists, sociologists, and others—as well as public policy makers responsible for targeting health spending, planning for intergenerational transfers such as social security–related retirement programs, and other areas. The chapters that follow provide an opportunity for us to identify the work we build on and offer our perspective on—setting the stage, extracting key elements that will provide building blocks for our approach, and highlighting the important intuitions from prior literature that will prove useful for the rest of this book.

2 Methods without Covariates

In this chapter, we discuss existing approaches to forecasting mortality (and other continuous variables) that do not include covariates. The underlying assumption of these methods is that all the information about the future is contained in the past observed values of the log-mortality rate. We ignore exogenous shocks to mortality, such as those from the discovery of new medical technologies, economic crises, education campaigns, public health innovations, or comorbidity patterns, while we include predictable epidemiological cycles due to biological or behavioral responses reflected in past mortality (see Gutterman and Vanderhoof, 1998).

Many results about how these approaches work appear here, including some proposed resolutions to several open disputes in the literature. We also identify several previously unrecognized limitations of some of these approaches. These results also motivate the introduction of our new methods in part II.

The methods discussed in this chapter have all been introduced by or used in the field of classical demography. A strength of this field is the propensity to stay very close to the data, which permits scholars to gain a detailed understanding of the data's strengths, weaknesses, and features. The care and attention they give to data reminds one of the way a botanist might pick up a delicate leaf in her hand, gently turning it over and closely examining and describing every feature.

Demographers thus treat data the way statisticians typically recommend that data analyses begin, although the mathematical tools of the two disciplines often differ. Demographers implicitly treat data as fixed rather than as a realization of a stochastic process. They are often less interested than statisticians in modeling the full data generation process that gave rise to the observations than in correcting errors, filling in missing values, and studying the numbers. Demographers are obviously aware of statistics, and they use some of its technology, but they are somewhat less likely to care about confidence intervals and standard errors or theories of inference. A disadvantage of this approach is that they sometimes do not take advantage of the powerful theories of inference, optimality properties of estimators, and general estimation and modeling techniques developed in the quantitative methods fields existing within other disciplines. More importantly from the perspective of this work, they miss the opportunity to include their deep knowledge

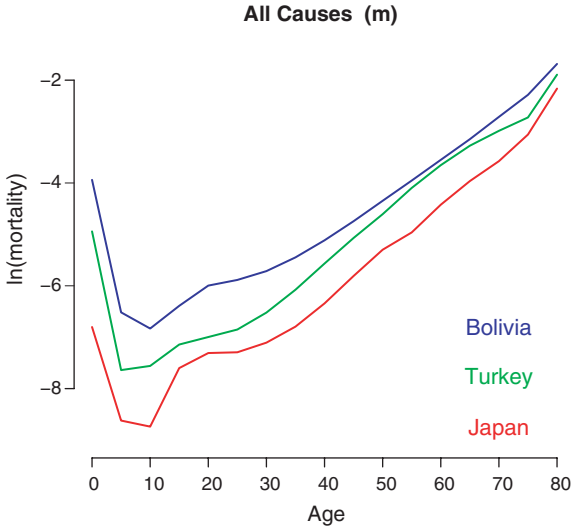

FIGURE 2.1. All-cause mortality age profiles: The age profile of log-mortality from all causes in males in the year 2000, for three countries. This shape is typical of that found in most countries. (The order and color of the countries listed matches the order of the lines at the right side of the graph.)

of demographic patterns in their models, which leaves their quantitative techniques impoverished relative to their qualitative knowledge.

We begin in section 2.1 by briefly describing a few of the common patterns found in mortality data and then, in subsequent sections, introduce models intended to fit these patterns. Section 2.2 offers a unified statistical framework for the remaining approaches described in this chapter. By identifying the limitations and implicit assumptions hard-coded into the quantitative methods in this chapter, we will be well positioned to build improved models in part II.

2.1 Patterns in Mortality Age Profiles

The relationship between mortality and age is "the oldest topic in demography" (Preston, Heuveline, and Guillot, 2001), dating to the political work by Graunt (1662). Demographers have shown not only that different age groups evidence markedly different mortality rates but that mortality varies as a function of age in systematic and predictable ways. Indeed, the evidence for systematic patterns, especially in all-cause mortality, is striking. In developed countries with good data, large populations, and no calamitous events, the all-cause log-mortality rate tends to decline from birth until about age five and then increases almost linearly until death. Some examples of countries with this well-known pattern (resembling the Nike® "swoosh") can be seen in figure 2.1, with age groups on the horizontal axis and

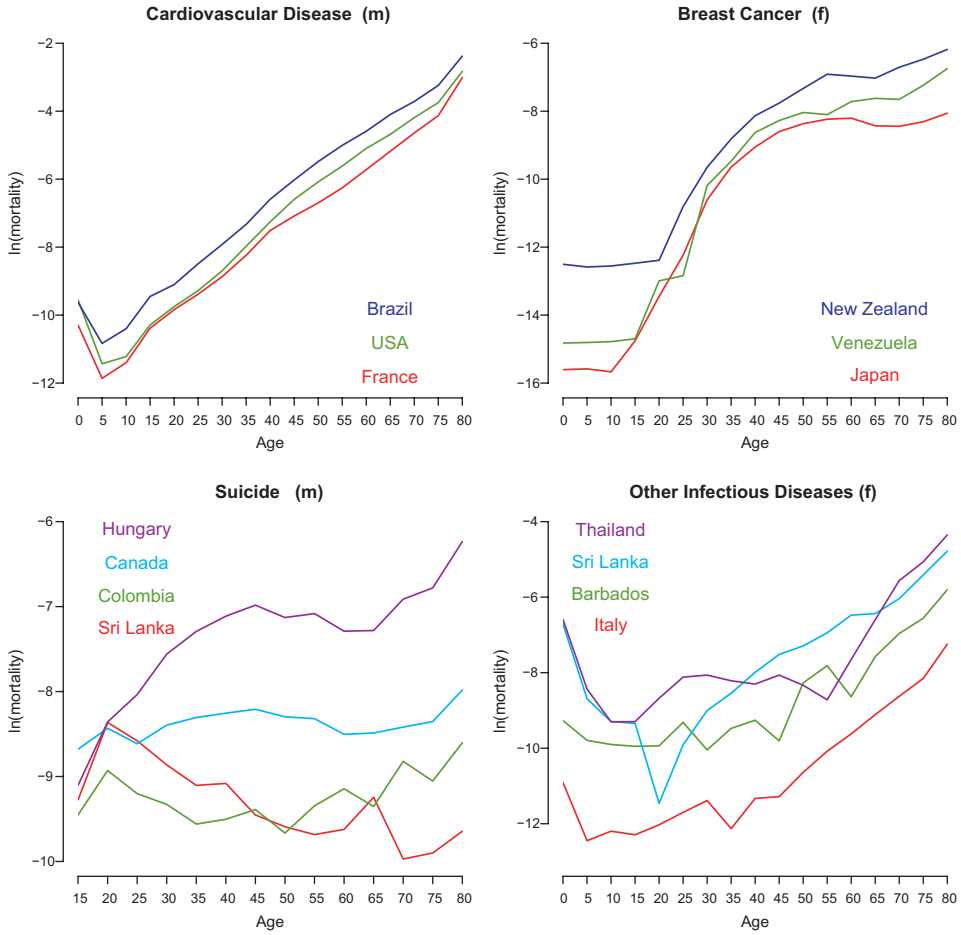

FIGURE 2.2. Cause-specific mortality age profiles: The top two graphs (for cardiovascular disease in males and breast cancer in females) show similar age patterns of mortality among the countries displayed; the same pattern is common among other countries as well. The bottom two graphs (for suicide in males and other infectious diseases in females) portray considerable cross-country variation, which is also typical of countries not displayed here. All the graphs refer to year 2000. (The order of the countries listed matches the order of the lines at the right side of the graph.)

the log-mortality rate (the log of the ratio of the number of deaths to population) on the vertical axis.

Mortality for specific causes also often follows clear, systematic patterns, but these patterns sometimes differ across causes (as we saw in figures 1.3 and 1.4), or countries or years. For example, the age profile of the log-mortality rate for cardiovascular disease among males, given in the top left graph in figure 2.2 for Brazil, the United States, and France, is closer to linear after age five than all-cause mortality. This is typical of other countries as well. In contrast, the pattern of female mortality due to breast cancer (on the

top right of figure 2.2) is also highly systematic, but the pattern differs markedly from the all-cause or cardiovascular disease swoosh patterns.

Figure 2.2 also portrays suicides among males and infectious diseases (other than AIDS, tuberculosis, and malaria) among females, both of which have log-mortality rate age profiles that differ markedly over the countries shown (and also over most other countries that do not appear here). A key pattern represented in all four graphs in figure 2.2 and for all-cause mortality in figure 2.1 and many others in figures 1.3 and 1.4 is that *the log-mortality rate is relatively smooth over age groups. Adjacent age groups have log-mortality rates that are closer than age groups farther apart*. We make use of this observation more directly in part II.

2.2 A Unified Statistical Framework

We can gain much insight into what seem to be (and what are proposed as) apparently ad hoc deterministic forecasting methods by translating them into formal statistical models. This translation helps in understanding what the calculations are about by revealing and focusing attention on the assumptions of the model, rather than the mere methods of computing estimates. And, as important, only by delineating the underlying statistical model is it possible to ascertain the formal statistical properties of any estimator. Without such a translation, establishing the formal statistical properties of an estimator is a much more arduous task, and one that is not usually attempted. Improving ad hoc methods is also a good deal harder because the assumptions one might normally relax are not laid bare.

We raise this issue because some forecasting methods in classical demography have this apparently ad hoc character. They are deterministic calculation rules that have no (stated) statistical models associated with them, no procedure for evaluating their statistical properties, and no accurate method of computing uncertainty estimates, such as standard errors or confidence intervals. In many fields, it is easier to ignore such ad hoc methods and to start building models from scratch. This would be a mistake with models in demography. Here, researchers know their data deeply and have developed calculation techniques (along the line of physics rather than social science methods) that work well in practice. Why they work in theory is not the subject of much research, and how they connect to formal statistical models has only sometimes been studied. But no statistical researcher can afford to ignore such an informative set of tools.

We now outline a unified framework that encompasses different, often competing, models proposed by several researchers in the field and used by many more. Although the details of the techniques can be very different, we show that the underlying methodological approach is the same. The basic idea is to reduce the dimensionality of the data to a smaller number of parameters by directly modeling some of the systematic patterns demographers have uncovered. We begin with a brief overview of some of these patterns and then discuss a statistical formalization.

We begin by defining m as a matrix of log-mortality rates (each element being the log of deaths per capita in a year for one age group), possibly cause- and sex-specific, for a single country, with A rows corresponding to age groups and T columns corresponding to

time periods, usually measured in years. For example,

$$m = \begin{array}{c} \\ 0 \\ 5 \\ 10 \\ 15 \\ 20 \\ 25 \\ 30 \\ 35 \\ \vdots \\ 80 \end{array} \begin{array}{c} \begin{array}{ccccc} 1990 & 1991 & 1992 & 1993 & 1994 \end{array} \\ \begin{pmatrix} m_{0,0} & m_{0,1} & m_{0,2} & m_{0,3} & m_{0,4} \\ m_{5,0} & m_{5,1} & m_{5,2} & m_{5,3} & m_{5,4} \\ m_{10,0} & m_{10,1} & m_{10,2} & m_{10,3} & m_{10,4} \\ m_{15,0} & m_{15,1} & m_{15,2} & m_{15,3} & m_{15,4} \\ m_{20,0} & m_{20,1} & m_{20,2} & m_{20,3} & m_{20,4} \\ m_{25,0} & m_{25,1} & m_{25,2} & m_{25,3} & m_{25,4} \\ m_{30,0} & m_{30,1} & m_{30,2} & m_{30,3} & m_{30,4} \\ m_{35,0} & m_{35,1} & m_{35,2} & m_{35,3} & m_{35,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{80,0} & m_{80,1} & m_{80,2} & m_{80,3} & m_{80,4} \end{pmatrix} \end{array}, \tag{2.1}$$

where each element is m_{at}, the log of the all-cause mortality rate in age group a ($a = 1, \ldots, A$) and time t ($t = 1, \ldots, T$) for one country. The starting point of many methods often consists of summarizing this mortality matrix, which has $A \times T$ elements, with a more parsimonious model of the form

$$m_{at} = f(a, \beta_a, \gamma_t) + \epsilon_{at}, \tag{2.2}$$

where β_a and γ_t are age and time multidimensional parameters, respectively, and ϵ_{at} is a zero mean disturbance, usually assumed to be independent and identically distributed (i.i.d.) and normally distributed. The function f is usually assumed known, but its specific form varies greatly from one class of models to another. Once a particular f has been chosen, then most methods proceed in three stages:

1. Estimate the vectors of parameters β_a and γ_t in model 2.2, using nonlinear least squares.
2. Treat the estimates of $\gamma_1, \ldots, \gamma_T$ as data in a multivariate time series (remember that γ_t can be a vector), or as a collection of independent univariate time series, depending on the particular model. Use standard autoregressive methods to forecast these time series.
3. To obtain a forecast for m_{at}, plug the forecasts for γ_t and the estimated values of β_a into the systematic component of model 2.2, $f(a, \beta_a, \gamma_t)$.

In the following, we review in more detail approaches that can each be seen as special cases of this framework.

2.3 Population Extrapolation Approaches

The simplest approach to forecasting is based on pure extrapolation. The idea is to define the function f in equation 2.2 based on one year (or an average of recent years) of mortality data (Lilienfeld and Perl, 1994; Armstrong, 2001). The data are classified by age and

perhaps other variables like sex, race, or country. These same mortality rates are assumed either to hold constant over time or to drop by some fixed proportion. The only changes over time in the number of deaths would then be a function of population changes, which are often just taken from projections computed by the U.S. Census Bureau or other national or international organizations.

Sometimes mortality rates are averaged prior to assuming constancy over time via age-sex-period or age-sex-cohort regression models with no exogenous variables (other than indicators for the cross-sectional strata). For example, Jee et al. (1998) estimate the lung cancer mortality rates in South Korea by a model with a constant and indicators for sex, age, cohort, and an interaction of sex and age. They then assume that predicted mortality rates from this model will remain constant over time.

In applying these methods, the rate of decline in mortality is often adjusted for expert opinions in various areas. For example, Goss et al. (1998) review the methods behind official government projections in the United States, Mexico, and Canada, and all three countries use similar extrapolative methods, with rates of mortality decline in various cross sections a function of expert judgment. The United Kingdom's Government's Actuary Department (2001) discusses the approaches taken in a variety of countries, many of which use some combination of extrapolative approaches along with some of the others we discuss here. Under this category of methods falls a huge range of reasonable but ad hoc approaches to forecasting or projecting mortality, most of which we do not list.

2.4 Parametric Approaches

Demographers have tried to reduce log-mortality age profiles, such as those portrayed in section 2.1, to simple parametric forms for centuries. The first attempt was by Gompertz (1825), who observed that the log of all-cause mortality is approximately linear after age 20, and so he used this form:

$$f(a, \beta) = \beta_0 + \beta_1 a,$$

which, of course, provides a simple special case of equation 2.2.

Since 1825, literally dozens of proposals for f have appeared in the literature (Keyfitz, 1968, 1982; Tabeau, 2001). A relatively elaborate current example of this approach is offered by McNown and Rogers (1989; 1992), who have led the way in recent years in marshaling parametric forms for the log-mortality age profile for forecasting (see also Rogers, 1986; Rogers and Raymer, 1999; and McNown, 1992). McNown and Rogers use the functional form due to Heligman and Pollard (1980):

$$f(a, \gamma_t) = \gamma_{1t}^{(a+\gamma_{2t})^{\gamma_{3t}}} + \gamma_{4t} \exp[-\gamma_{5t}(\ln a - \ln \gamma_{6t})^2] + \frac{\gamma_{7t}\gamma_{8t}^a}{(1 + \gamma_{7t}\gamma_{8t}^a)}. \tag{2.3}$$

This particular choice of f does not have parameters β_a depending on age but has eight time-dependent parameters $\gamma_{1t}, \ldots, \gamma_{8t}$. The particular parametric form is unimportant here, however, because many others have been proposed and used for forecasting and other

purposes. The key point is that the A age groups are being summarized by this simpler form with somewhat fewer adjustable parameters. Once the parameters $\gamma_{1t}, \ldots, \gamma_{8t}$ have been estimated, they are forecast separately, using standard independent univariate time-series methods. The forecasted parameters are then plugged into the right side of the same equation to produce forecasts for mortality. Of course, this procedure does not necessarily guarantee much smoothness over the forecasted age profile unless the parameters are highly constrained. Although adding constraints is common practice (e.g., McNown and Rogers, 1989, sometimes constrain all but three parameters to fixed points), it can be difficult to interpret the specific parameters and the appropriate constraints without examining the entire age profile, particularly as they evolve out of sample.

A more serious problem with parameterized mortality forecasts is that the parameters are forecast separately. Put differently, the forecast for γ_{1t} is not "aware" of the forecast for γ_{2t}. Although each alone might fit the data well, the combination of all the parameters may imply an empirically unreasonable age profile.

An equally serious problem, which seems to have gone unnoticed in the literature, stems from the fact that for each time t the parameters of equation 2.3 are estimated as the solution of a complex nonconvex optimization problem. Unless we take great care to make sure that for each time t the global minimum is achieved (assuming that is unique!), there is always the risk that the estimated parameters jump from one local minimum to another as we move from one year to the next, rather than tracking the global optimum, therefore leading to meaningless time-series forecasts.

Still another issue is that many of the parametric forms used in the literature involve combinations and compositions of infinitely smooth functions, such as the exponential, which can be represented by infinite power series converging over a certain (possibly infinite) range. Unfortunately, this is typically not an optimal choice in smoothing problems like this. Many of these functions behave like polynomials (of infinite degree) and share with polynomials the undesirable property of being highly "inflexible": constraining the behavior of a polynomial at one age alters its behavior over the entire range of ages, so that any notion of "locality" in the approximation is lost. Many of these resulting parametric forms are thus highly nonrobust to data errors or idiosyncratic patterns in log-mortality.[1]

The idea of reducing the complexity of the data prior to forecasting is exceptionally powerful, and McNown and Rogers have taken this very far and produced many articles and forecasts using it. In our experience, the specific methods they have proposed can work well and poorly, depending on the structure of the data. However, no one has been able to ascertain when such forms are appropriate and when they miss important features of the data. As they have made clear in their work, this ambiguity is to be expected. Their methods work only to the extent that the data reduction does not lose critical features and the parameter forecasting turns out to produce a consistent age profile of mortality. The methods have some implementation difficulties, in that fitting 8 or more parameters to 20 data points can be tricky. With the gracious help of McKnown and Rogers, we were

[1] Parsimonious function representations that are smooth and preserve local properties are available and go under the generic name of splines. Splines are classes of piecewise polynomial functions, and algorithms for estimating them are readily available, robust, and easy to use. It would be interesting to see whether they could be used to improve (and simplify) the current methods based on parametric curve fitting.

able to achieve convergence following their procedures only by restricting a number of the coefficients to very narrow ranges, although we were unable to replicate anything close to their specific numerical results or implied age profiles. Despite technical difficulties, the general idea of identifying structure, if not the details of any specific approach, will surely endure.

2.5 A Nonparametric Approach: Principal Components

2.5.1 Introduction

The key feature of the approaches described thus far is an explicit systematic component that specifies the mathematical form of an expected age profile at any point in time. An alternative approach consists of using nonparametric descriptions of the log-mortality age profile, in which one estimates details of the functional form f (from equation 2.2) rather than specifying them a priori. More precisely, the idea is to represent the full set of age profiles by a linear combination of k "basic" shapes of age profiles ($k \leq A$), where the coefficients on the shapes *and* the shapes themselves are both estimated from the data. If all the age profiles look alike, then only a few shapes should be necessary to describe well each age profile at any point in time, providing a parsimonious representation of the data. This idea is formalized by the method of principal component analysis (PCA).

PCA made its first appearance in demography with Ledermann and Breas (1959), who used factor analysis to analyze life table data from different countries. It was then used by Bozik and Bell (1987) and Sivamurthy (1987) for the projection of age-specific fertility rates. The method of Bozik and Bell was then extended by Bell and Monsell (1991) to forecast age-specific mortality rates, but it was not until Lee and Carter's (1992) somewhat simpler formulation that PCA methods became widely used (see also Lee, 1993, 2000a, 2000b; and Lee and Tuljapurkar, 1994, 1998a, 1998b). We discuss the Lee-Carter model in section 2.6.

Because PCA may not be familiar to all the readers, we outline here the intuition behind it. From a mathematical standpoint, the method is easiest to understand as an application of the singular value decomposition (SVD), the technical details of which appear in appendix B.2.4 (page 233).

Our goal is build a parsimonious representation of the data, consisting of a collection of T log-mortality age profiles $m_t \in \mathbb{R}^A$. A simple representation of the data is the empirical mean age profile (i.e., the average over the existing age profiles). That would imply that we model log-mortality as follows:

$$m_t = \bar{m} + \epsilon_t,$$

where the average $A \times 1$ age profile is

$$\bar{m} = \frac{1}{T} \sum_{t=1}^{T} m_t. \tag{2.4}$$

This model is parsimonious, because it summarizes the entire log-mortality data matrix, composed of $A \times T$ entries, with a vector of only A numbers, $\bar{m}$. However, it is also

obviously too restrictive: while it captures some of the variation across age groups, it ignores all variation over time.

Thus, we next consider a marginal improvement over this model by allowing the average age profile to shift rigidly up and down as a function of time. Formally, this model is expressed as

$$m_t = \bar{m} + \gamma_t v + \epsilon_t, \quad v = (1, 1, \ldots, 1) \in \mathbb{R}^A,$$

where $\gamma_1, \ldots, \gamma_T$ represent an additional set of T unknown-time fixed effects that are easily estimated by least squares. To forecast mortality from this model, we would estimate the parameters, use a univariate forecasting model applied to the estimated values of γ_t, and plug the future values of γ_t in the foregoing specification. This model has a total of $A + T$ parameters and has the implication that the rate of change in mortality is the same across all the age groups and that the age profile has the same shape for all time periods.

This model is closer in spirit to what we set out to create, as the basic shapes used here to represent log-mortality are the average age profile $\bar{m}$ for all years, and the constant age profile v shifting over the years as a function of γ_t. However, while we derived the average age profile from the data, we chose the constant age profile v by assumption, which can be thought of as a particular age profile parametrization. In order to obtain a model more in line with a nonparametric approach, we replace the fixed (constant) age profile v with an unknown age profile β. The vector β is known as the *first principal component*, and we will compute it from the data. The model then becomes

$$m_t = \bar{m} + \gamma_t \beta + \epsilon_t \quad \beta \in \mathbb{R}^A, \tag{2.5}$$

where we estimate the vectors $\bar{m}$ and β (as well as $\gamma_1, \ldots, \gamma_T$) from the data. We will refer to the product $\gamma_t \beta$ as the portion of log-mortality explained by the first principal component. Under the assumption that the disturbances ϵ_t are standard normal, the maximum likelihood estimators of γ_t and β are easily computed in terms of the SVD of the log-mortality matrix m, as we explain later.

Figure 2.3 illustrates in more detail the model in equation 2.5. In the top graph in this figure, all-cause male log-mortality age profiles in Italy, are plotted for each year from 1981 to 2000. The remaining graphs in this figure decompose these observed data. The second graph plots the mean age profile, and the third plots the portion of log-mortality explained by the first principal component, where for clarity we are showing only the graph corresponding to year 2000. Because the age profiles in the first graph have fairly narrow variance in these data, the mean accounts for most of the pattern. The combination of the mean and the term containing the first principal component account for a large fraction of the observed variation. We can see this by examining the residuals plotted in the last graph, which have zero mean, very small variance, and no obvious remaining patterns.

Obviously, we cannot know a priori whether model 2.5 will represent any data set accurately—that is, whether linear combinations of the two basic shapes are enough to describe the features of the age profiles for all the years we are interested in. Fortunately, it is straightforward to generalize this model to an arbitrary number of basic shapes. Such a model can be written as

$$m_t = \bar{m} + \gamma_{1t} \beta_1 + \gamma_{2t} \beta_2 + \ldots \gamma_{kt} \beta_k + \epsilon_t. \tag{2.6}$$

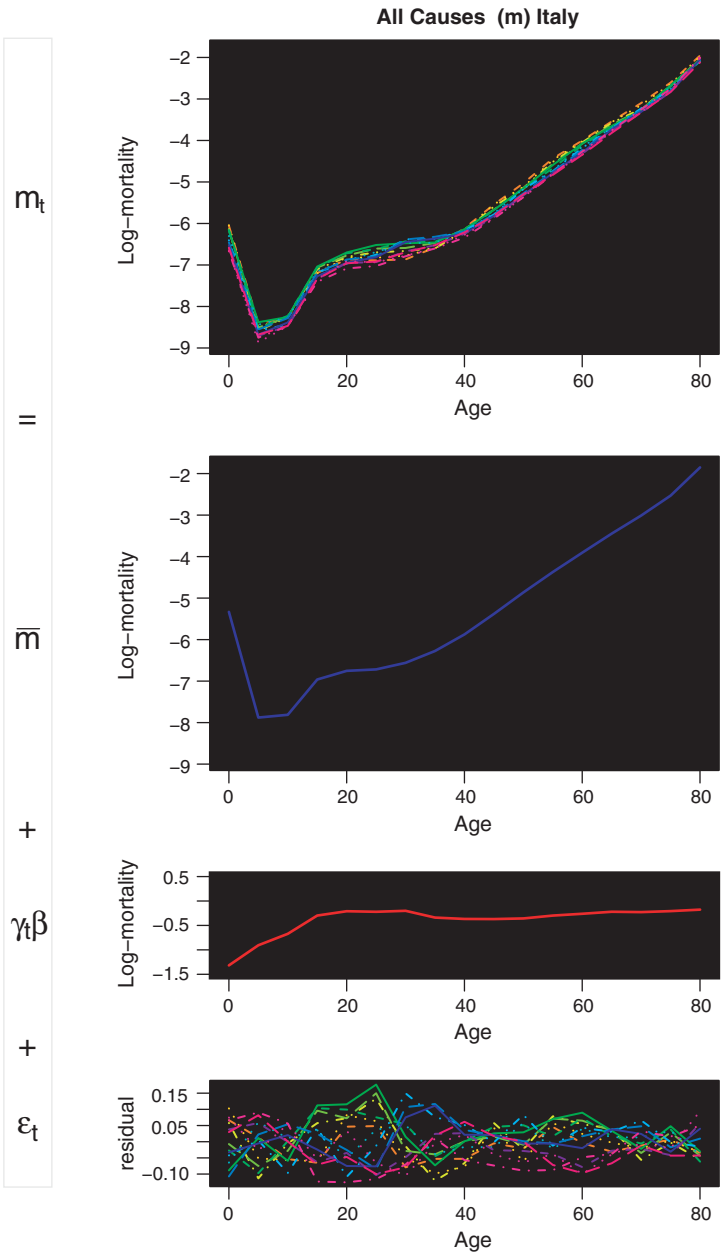

FIGURE 2.3. Decomposing age profiles with principal components: This figure parallels equation 2.5, which decomposes the 20 male all-cause age profiles of log-mortality in Italy (*top graph*) into the mean age profile (*second graph*), the first principal component (for clarity, just for the year 2000; *third graph*), and the residuals (*fourth graph*).

Because the vectors $\beta_1, \ldots, \beta_k$ are unknown and so must be estimated, we do not restrict the solution in any substantive way by assuming they are mutually orthogonal, and so we do so. We refer in the following to equation 2.6 as a specification with k principal components. As in the case of only one component, if the disturbance vector ϵ_t is assumed to be normally distributed, the maximum likelihood estimators of $(\gamma_{1t}, \ldots, \gamma_{kt})$ and $(\beta_1, \ldots, \beta_k)$ correspond to specific rows and columns of the SVD of the log-mortality data matrix m. The estimated values of $(\beta_1, \ldots, \beta_k)$, are known as the *first k principal components* of the data matrix m.

A key point is that the principal components are an intrinsic property of the data and do not depend on the particular specification they belong to. In other words, the maximum likelihood estimators of β_1 in a specification with one component and in a specification with five are identical, so that β_1 is always the first principal component, so long as the data remain the same. This pattern implies that there is a natural order of importance among the principal components: the first is more important than the second, which is more important than third, and so on.

To ascertain what "more important" means, we can, for example, estimate a specification with two components and ask, If we must drop one of the two principal components, which one should we drop? We clearly should drop the second, because the first principal component is optimal (in the sense of maximum likelihood) for the specifications with only one component. Therefore, we should think of the principal components as a nested set of models of the data. We start from a model with one principal component only, which explains a certain percentage of the variance in the data. Then we can refine this model by adding a second principal component and explain an additional percentage of the variance. If this model is not accurate enough, a third principal component can be added, and so on. At each step the added principal component explains, optimally, a percentage of the variance in the data that could not be explained at the previous step. For this reason, the principal components tend to look like age profiles of increasing complexity, because each is a refinement over the previous one.

As an example, we plot in figure 2.4 the 1st, the 2nd, and the 17th principal components for cardiovascular disease in females in the United Kingdom (with 17 age groups). Notice how the 2nd and 17th principal components are more "complex" than the 1st.

Obviously, with A age groups, we can explain 100% of the variance using A principal components. However, the usefulness of principal component analysis lies in the fact that in many data sets, relatively few principal components provide a good approximation to the original data. Suppose for example that four principal components provide enough flexibility to model the age profiles for a particular combination of country, cause, and gender and that we have 17 age groups ($A = 17$). That means that instead of having to forecast 17 time series, one for each group, we have to forecast only 4 time series, those corresponding to $\gamma_1, \ldots, \gamma_4$. However, the method can even be used with $k = A$, in which case the dimensionality of the problem has not been reduced, and so we still have to forecast A time series, but it has been shown that the time series of γ are still often much better behaved than the time series of the raw log-mortality data (Bell and Monsell, 1991; Bozik and Bell, 1987).

As an example, we report in figure 2.5 the maximum likelihood estimates of each of the time series $\gamma_{1t}, \ldots, \gamma_{4t}$, for the category "other malignant neoplasms" (which include all types of cancer other than lung, stomach, liver, and mouth and esophagus) in Japanese males.

In these data, the third and fourth time series hover around zero. Because the principal components are mutually orthogonal, we can interpret γ_{nt} as a measure of how much the

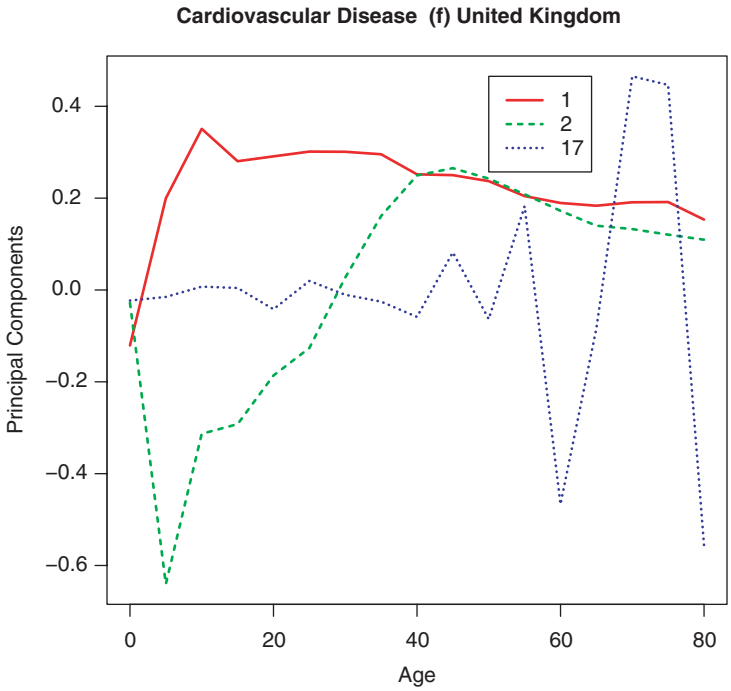

FIGURE 2.4. Principal components of log-mortality: 1st, 2nd, and 17th (there are 17 age groups). This figure refers to log-mortality from cardiovascular disease in females, United Kingdom.

n-th principal component explains of the deviation of the log-mortality age profile from the average age profile. This implies that the third and fourth principal components do not play a crucial role in explaining the shapes of the age profiles in this case.

2.5.2 Estimation

We find it simplifying to think of PCA as a simple application of SVD (See appendix B.2.4, page 233). The SVD theorem asserts that any $A \times T$ matrix Q can be written uniquely as $Q = BLU'$, where B is an $A \times A$ orthonormal matrix, L is an $A \times A$ diagonal matrix with positive or zero entries known as *singular values* ordered from high to low, and U is a $T \times A$ matrix whose columns are mutually orthonormal.

By defining the matrix $G \equiv LU'$, we write the SVD of Q as

$$Q_{at} = B_{a1}G_{1t} + B_{a2}G_{2t} + \cdots + B_{aA}G_{At},$$

where each term of the sum on the right side of this equation defines an $A \times T$ matrix of rank 1. If we assume that $T > A$ and that Q has full rank, the SVD of Q is a unique way of decomposing a matrix of rank A as the sum of A matrices of rank 1. SVD also says that if we want to approximate, in the least-squares sense, the matrix Q with $k < A$ matrices of rank 1, the best of way of doing it is to take the first k terms of the preceding SVD decomposition. Denoting by b_n the n-th column of B and by Q_t the t-th column of Q, the

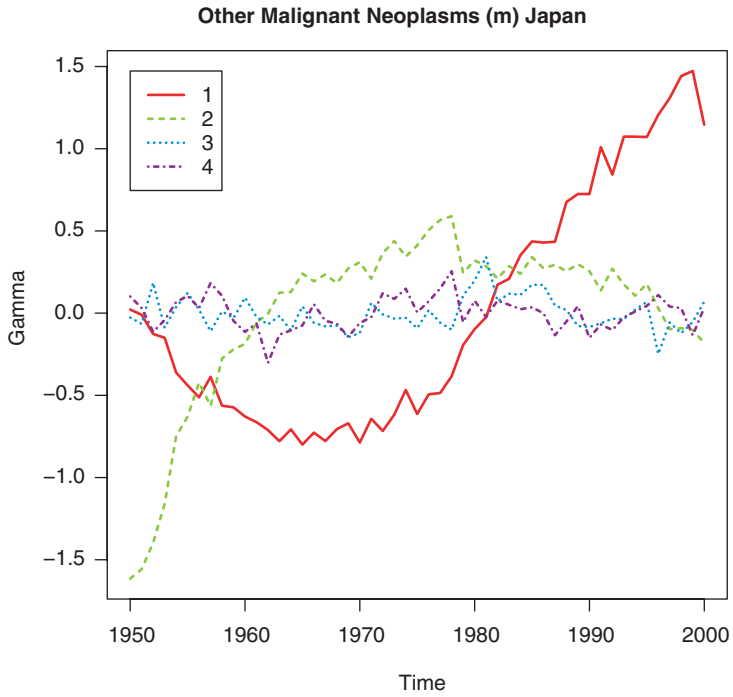

FIGURE 2.5. Maximum likelihood estimates of time series of γ_1, γ_2, γ_3, and γ_4 for other malignant neoplasms in Japanese males.

k-term approximation of the preceding expression can be written in vector form as

$$Q_t \approx b_1 G_{1t} + b_2 G_{2t} + \cdots + b_k G_{kt}.$$

Now we apply this approximation to the matrix of *centered age profiles*, defined as

$$\tilde{m}_t \equiv m_t - \bar{m}, \tag{2.7}$$

by relabeling the variables:

$$Q_t \rightsquigarrow \tilde{m}_t, \quad b_i \rightsquigarrow \beta_i, \quad G_{it} \rightsquigarrow \gamma_{it}.$$

Thus, we conclude that there exist k A-dimensional vectors (age profiles) β_i and k time series $\gamma_{1t}, \ldots, \gamma_{kt}$ such that the following approximation is optimal in the least-squares sense:

$$m_t \approx \bar{m} + \beta_1 \gamma_{1t} + \beta_2 \gamma_{2t} + \cdots + \beta_k \gamma_{kt}, \tag{2.8}$$

which we recognize as the maximum likelihood estimate of the specification in equation 2.6. Hence, in order to compute the first k principal components, we need to compute only the SVD of the matrix of centered age profiles $\tilde{m} = BLU'$ and take the first k columns of B, which are also known as *the first k left singular vectors*. The set of time series

$\gamma_{1t}, \ldots, \gamma_{kt}$ can be read as the first k rows of $G = LU'$. Alternatively, because $G = B'\tilde{m}$, we can also obtain the time series γ_{kt} as

$$\gamma_{kt} = \beta_k' \tilde{m}_t. \tag{2.9}$$

This last expression makes clear that γ_{kt} is simply the projection of the centered age profile $\tilde{m}_t$ on the k-th principal component.

Several algorithms are available for computing the SVD of a matrix, and most statistical packages have an intrinsic SVD routine. If an SVD routine is not available, an alternative is a routine for the computation of the eigenvectors of a symmetric matrix, because this is a simpler, more restrictive problem. In fact, the columns of the matrix B in the SVD of a matrix Q coincide with the eigenvectors of the matrix QQ'. As a consequence, the first k principal components can be computed as the eigenvectors of QQ' corresponding to the largest k eigenvalues. In many approaches to PCA, this is the given definition of principal components. In the rest of the book, we switch freely between these two definitions, because they are equivalent and differ only in the particular procedure used to compute the result.

2.6 The Lee-Carter Approach

The principal-components-based model developed by Lee and Carter (1992) is now used by the U.S. Census Bureau as a benchmark for its population forecasts, and its use has been recommended by two recent U.S. Social Security Technical Advisory Panels. Although Lee and Carter intended for it to be used for all-cause mortality in the United States and a few other similarly developed countries, the model is now used widely by scholars forecasting all-cause and cause-specific mortality around the world (Tuljapurkar, Li and Boe, 2000; Preston, 1991; Wilmoth, 1996; Haberland and Bergmann, 1995; Lee, Carter and Tuljapurkar, 1995; Lee and Rofman, 1994; Tuljapurkar and Boe, 1998; NIPSSR, 2002; Booth, Maindonald and Smith, 2002).

We begin with the model in section 2.6.1 and then discuss estimation in section 2.6.2 and forecasting in section 2.6.3. Section 2.6.4 discusses the properties of this approach. A more extensive treatment of this model appears in Girosi and King (2006).

2.6.1 The Model

The first step of the Lee-Carter method consists of modeling the mortality matrix in equation 2.1 as

$$m_{at} = \alpha_a + \beta_a \gamma_t + \epsilon_{at}, \tag{2.10}$$

where α_a, β_a, and γ_t are parameters to be estimated and ϵ_{at} is a set of disturbances, which is obviously a special case of PCA with $k = 1$ principal components (see section 2.5). This expression is also a special case of the unified statistical model given in equation 2.2 (page 25): it differs structurally from parametric models of the type in equation 2.3, given that the dependence on age groups is nonparametric and represented by the parameters β_a.

The parametrization in equation 2.10 is not unique, because it is invariant with respect to the transformations:

$$\begin{aligned} &\beta_a \rightsquigarrow c\beta_a && \gamma_t \rightsquigarrow \frac{1}{c}\gamma_t && \forall c \in \mathbb{R}, c \neq 0 \\ &\alpha_a \rightsquigarrow \alpha_a - \beta_a c && \gamma_t \rightsquigarrow \gamma_t + c && \forall c \in \mathbb{R}. \end{aligned}$$

This invariance is not a conceptual obstacle; it merely means that the likelihood associated with the preceding model has an infinite number of equivalent maxima. It is then sufficient to pick a consistent rule to identify the parameters, which can be done by imposing two constraints. We follow Lee and Carter in adopting the constraint $\sum_t \gamma_t = 0$. Unlike Lee and Carter, however, we set $\sum_a \beta_a^2 = 1$ (they set $\sum_a \beta_a = 1$). This last choice is done only to simplify some calculations later on and has no bearing on empirical applications.

The constraint $\sum_t \gamma_t = 0$ immediately implies that the parameter α_a is simply the empirical average of log mortality in age group a: $\alpha_a = \bar{m}_a$. Because the Lee-Carter model is consistent with an assumption that the disturbances ϵ_{at} in the preceding model are normally distributed, we rewrite equation 2.10 as

$$\begin{aligned} m_{at} &\sim \mathcal{N}\left(\mu_{at}, \sigma^2\right) \\ \mu_{at} &= \bar{m}_a + \beta_a \gamma_t, \end{aligned} \tag{2.11}$$

which is equivalent to a multiplicative fixed-effects model for the centered age profile:

$$\begin{aligned} \tilde{m}_{at} &\sim \mathcal{N}\left(\bar{\mu}_{at}, \sigma^2\right) \\ \bar{\mu}_{at} &= \beta_a \gamma_t. \end{aligned} \tag{2.12}$$

In this expression, we use only $A + T$ parameters ($\beta_a \gamma_t$, for all a and t, represented on the bottom and right margins of the following matrix) to approximate the $A \times T$ elements of the matrix:

$$\tilde{m} = \begin{array}{c} \\ 5 \\ 10 \\ 15 \\ 20 \\ 25 \\ 30 \\ 35 \\ \vdots \\ 80 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} 1990 & 1991 & 1992 & 1993 & 1994 \end{array} \\ \begin{pmatrix} \tilde{m}_{5,0} & \tilde{m}_{5,1} & \tilde{m}_{5,2} & \tilde{m}_{5,3} & \tilde{m}_{5,4} \\ \tilde{m}_{10,0} & \tilde{m}_{10,1} & \tilde{m}_{10,2} & \tilde{m}_{10,3} & \tilde{m}_{10,4} \\ \tilde{m}_{15,0} & \tilde{m}_{15,1} & \tilde{m}_{15,2} & \tilde{m}_{15,3} & \tilde{m}_{15,4} \\ \tilde{m}_{20,0} & \tilde{m}_{20,1} & \tilde{m}_{20,2} & \tilde{m}_{20,3} & \tilde{m}_{20,4} \\ \tilde{m}_{25,0} & \tilde{m}_{25,1} & \tilde{m}_{25,2} & \tilde{m}_{25,3} & \tilde{m}_{25,4} \\ \tilde{m}_{30,0} & \tilde{m}_{30,1} & \tilde{m}_{30,2} & \tilde{m}_{30,3} & \tilde{m}_{30,4} \\ \tilde{m}_{35,0} & \tilde{m}_{35,1} & \tilde{m}_{35,2} & \tilde{m}_{35,3} & \tilde{m}_{35,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \tilde{m}_{80,0} & \tilde{m}_{80,1} & \tilde{m}_{80,2} & \tilde{m}_{80,3} & \tilde{m}_{80,4} \end{pmatrix} \\ \begin{array}{ccccc} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \end{array} \end{array} \cdot \begin{array}{c} \\ \beta_5 \\ \beta_{10} \\ \beta_{15} \\ \beta_{20} \\ \beta_{25} \\ \beta_{30} \\ \beta_{35} \\ \vdots \\ \beta_{80} \\ \\ \end{array} \tag{2.13}$$

For example, Lee-Carter approximates $\tilde{m}_{5,0}$ in the top left cell by the product of the parameters at the end of the first row and column $\beta_5 \gamma_0$.

Seen in this framework, the Lee-Carter model can also be thought of as a special case of log-linear models for contingency tables (Bishop, Fienberg, and Holland, 1975; King, 1989b: ch. 6), where many cell values are approximated with estimates of parameters representing the marginals. Indeed, this model closely resembles the most basic version of contingency table models, where one assumes *independence* of rows (age groups) and columns (time periods), and the expected cell value is merely the product of the two parameter values from the respective marginals: $E(\tilde{m}_{at}) = \beta_a \gamma_t$. In a contingency table model, this assumption would be appropriate if the variable represented as rows in the table were independent of the variable represented as columns. The same assumption for the log-mortality rate is the absence of age $\times$ time interactions—that β_a is fixed over time for all a and γ_t is fixed over age groups for all t.

2.6.2 Estimation

The parameters β_a and γ_t in model 2.12 can be estimated via maximum likelihood applied to equation 2.11. We do not need to go through this derivation because we have already shown this result in the context of PCA. In fact, the Lee-Carter specification 2.10 is a particular case (with $k = 1$) of the principal components expansion 2.6, whose estimation has been discussed in section 2.5.2 in the context of SVD. As a consequence, the algorithm for the estimation of the parameters in the Lee-Carter model is as follows:

1. Compute the SVD decomposition of the matrix of the centered age profiles: $\tilde{m} = BLU'$.
2. The estimate for β is the first column of B
3. The estimate for γ_t is $\beta' \tilde{m}_t$.

If for any reason the singular values are not sorted in descending order, then the estimate for β is the column of B, which corresponds to the largest singular value. If β does not have length one, then it should be replaced by $\beta / \|\beta\|$. Alternatively, if the SVD decomposition of $\tilde{m}$ is not available, one can compute as the normalized eigenvector of the matrix $C \equiv \tilde{m}\tilde{m}'$ corresponding to the largest eigenvalue.

In practice, Lee and Carter suggest that, after β and γ have been estimated, the parameter γ_t be reestimated using a different criterion. This reestimation step, often called "second stage estimation," does not always have a unique solution for the criterion outlined in Lee and Carter (1992). In addition, different criteria have been proposed more recently (Lee and Miller, 2001; Wilmoth, 1993), and some researchers skip this reestimation stage altogether. These procedures, and difficulties with them, are described in Girosi and King (2007).

2.6.3 Forecasting

To forecast, Lee and Carter assume that β_a remains constant over time and forecast future values of γ_t with a standard univariate time-series model. After testing several autoregressive integrated moving average (ARIMA) specifications, they find that a random walk with drift is the most appropriate model for their data. They make clear that other ARIMA models might be preferable for different data sets, but in practice the random walk

with drift model for γ_t is used almost exclusively in applications. This model is as follows:

$$\begin{aligned}\hat{\gamma}_t &= \hat{\gamma}_{t-1} + \theta + \xi_t \\ \xi_t &\sim \mathcal{N}(0, \sigma^2_{\text{rw}}),\end{aligned} \tag{2.14}$$

where θ is known as *the drift parameter*. The maximum likelihood estimates of the model preceding are as follows:

$$\begin{aligned}\hat{\theta} &= \frac{\hat{\gamma}_T - \hat{\gamma}_1}{T-1} \\ \hat{\sigma}^2_{\text{rw}} &= \frac{1}{T-1}\sum_{t=1}^{T-1}(\hat{\gamma}_{t+1} - \hat{\gamma}_t - \hat{\theta})^2\end{aligned} \tag{2.15}$$

with

$$\text{Var}[\hat{\theta}] = \frac{\sigma^2_{\text{rw}}}{T-1}. \tag{2.16}$$

Once these estimates have been computed, we obtain a forecast for $\hat{\gamma}_t$ in both stochastic and deterministic form. For example, to forecast two periods ahead, we substitute for $\hat{\gamma}_{t-1}$ in equation 2.14:

$$\begin{aligned}\hat{\gamma}_t &= \hat{\gamma}_{t-1} + \theta + \xi_t \\ &= (\hat{\gamma}_{t-2} + \theta + \xi_{t-1}) + \theta + \xi_t \\ &= \hat{\gamma}_{t-2} + 2\theta + (\xi_{t-1} + \xi_t).\end{aligned} \tag{2.17}$$

Conditioning on the estimate of (i.e., ignoring the uncertainty in) θ enables one to substitute in $\hat{\theta}$:

$$\hat{\gamma}_t = \hat{\gamma}_{t-2} + 2\hat{\theta} + (\xi_{t-1} + \xi_t).$$

Hence, to forecast $\hat{\gamma}_t$ at time $T + (\Delta t)$ with data available up to period T, we iterate equation 2.14 Δt time periods forward, plug into it the estimate for θ, and obtain

$$\hat{\gamma}_{T+(\Delta t)} = \hat{\gamma}_T + (\Delta t)\hat{\theta} + \sum_{l=1}^{(\Delta t)} \xi_{T+l-1}.$$

Because the random variables ξ_t are assumed to be independent with the same variance σ^2_{rw}, the last term in the preceding equation is normally distributed with variance $(\Delta t)\sigma^2_{\text{rw}}$, and therefore it has the same distribution as the variable $\sqrt{(\Delta t)}\xi_t$. This allows us to rewrite the preceding equation more simply as

$$\hat{\gamma}_{T+(\Delta t)} = \hat{\gamma}_T + (\Delta t)\hat{\theta} + \sqrt{(\Delta t)}\xi_t. \tag{2.18}$$

Because the variance of ξ_t can be estimated using equation 2.16, we can use equation 2.18 to draw samples for the forecast at time $T + (\Delta t)$, conditionally on the

realization of $\hat{\gamma}_1, \ldots, \hat{\gamma}_T$. Doing so for increasing values of (Δt) yields a *conditional* stochastic forecast for the time series $\hat{\gamma}_t$. We emphasize the word "conditional," because $\hat{\gamma}_1, \ldots, \hat{\gamma}_T$ are random variables themselves, as well as $\hat{\theta}$, and if we included the variation due to these variables as well, we would obtain forecasts with much higher variance. The conditional variance of the forecast in equation 2.18 is

$$\text{Var}[\hat{\gamma}_{T+(\Delta t)}|\hat{\gamma}_1, \ldots, \hat{\gamma}_T] = (\Delta t)\sigma^2_{\text{rw}}. \tag{2.19}$$

Therefore, the conditional standard errors for the forecast increase with the square root of the distance to the forecast "horizon" (Δt).

For most practical purposes scholars use the point estimates of the stochastic forecasts, which follow a straight line as a function of (Δt), with slope $\hat{\theta}$:

$$\text{E}[\hat{\gamma}_{T+(\Delta t)}|\hat{\gamma}_1, \ldots, \hat{\gamma}_T] = \hat{\gamma}_T + (\Delta t)\hat{\theta}. \tag{2.20}$$

We now plug these expressions into the empirical and vectorized version of equation 2.11 to make (point estimate) forecasts for log-mortality as

$$\mu_{T+(\Delta t)} = \bar{m} + \hat{\beta}\hat{\gamma}_{T+(\Delta t)} = \bar{m} + \hat{\beta}(\hat{\gamma}_T + (\Delta t)\hat{\theta}). \tag{2.21}$$

For example, the Lee-Carter model computes the forecast for year 2030, given data observed from 1950 to 2000, as

$$\begin{aligned} \hat{\mu}_{2030} &= \bar{m} + \hat{\beta} \times [\hat{\gamma}_{2000} + 30\hat{\theta}] \\ &= \bar{m} + \hat{\beta} \times \left[\hat{\gamma}_{2000} + 30\frac{(\hat{\gamma}_{2000} - \hat{\gamma}_{1950})}{50}\right]. \end{aligned} \tag{2.22}$$

2.6.4 Properties

Although the Lee-Carter model can be estimated with any time-series process applied to forecast γ_t, the random walk with drift specification in equation 2.14 accounts for nearly all real applications. With that extra feature of the model, Girosi and King (2006) prove that:

1. When considered together, the two-stage Lee-Carter approach is a special case of a simple random walk with drift model. (The random walk with drift model can have any arbitrary error structure, whereas the Lee-Carter model requires the error to have a particular restricted structure. The two models are otherwise identical.)
2. An unbiased estimator of the drift parameter in the random walk with drift model, and hence also of the analogous parameter in the Lee-Carter model, requires one stage and no principal component or singular value decomposition. It is simply $(m_T - m_1)/(T - 1)$.
3. If data are generated from the Lee-Carter model, then the Lee-Carter estimator and the random walk with drift estimator are both unbiased.
4. If the data are generated by the more general random walk with drift model, then the two-stage Lee-Carter estimator is biased, but the simple random walk with drift estimator is unbiased.

These results thus pose the question of why one would prefer to use the Lee-Carter estimator rather than the simpler and more broadly unbiased random walk with drift estimator (restricting ourselves for the moment to the choice of these two models). Perhaps the restrictions in the Lee-Carter error structure could be justified as plausible for some applications, but they obviously will not apply all the time and would be easy to reject in most cases using conventional statistical tests. For reasons we highlight later, empirical differences between the two estimators are often not major, but unless we are in the hypothetical situation where the Lee-Carter assumptions are known to apply, the random walk with drift model would appear to be preferred whenever the two differ.

A simple implication of these results is that Lee-Carter forecasts may work in the short run when mortality moves slowly over time, as is common in all-cause mortality in the United States, for which the method was designed. However, long-run forecasts in other data can shoot off in different directions for different age groups, have a variance across age groups for any year that always eventually increases no matter what the data indicate, will not normally maintain the rank order of the age groups' log-mortality rates or any given age profile pattern, and will always produce a mortality age profile that becomes less smooth over time, after a point. Anything is possible in real data out of sample, but these properties are functions of the model and not necessarily the data or most demographers' priors.

The fact that age profiles evolve in implausible ways under the Lee-Carter model has been noted in the context of particular data sets (Alho, 1992). Our result means that this point is general and does not depend on the particular time-series extrapolation process used or idiosyncrasies of the data set chosen. The result would also seem to resolve a major point of contention in the published debates between Lee and Carter (1992) and McNown (1992) about whether the Lee-Carter model sufficiently constrains the out-of-sample age profile of mortality: almost no matter what one's prior is for a reasonable age profile, Lee-Carter forecasts made sufficiently far into the future will eventually violate it.

To illustrate, figures 2.6 and 2.7 offer examples of six data sets, one in each row. We chose these data to highlight the assumptions of Lee-Carter and the random walk with drift models, not because Lee and Carter or anyone else would apply their method to data like these. The left graph in each row is a time-series plot of the log-mortality rate for each age group (color-coded by age group and labeled at the end of each line), and the right graphs include the age profiles (color-coded by year). For each, the data are plotted up to year 2000 and the Lee-Carter forecasts are plotted for subsequent years.

An easy way to understand Lee-Carter is as forecasts from the more general and simpler random walk with drift model. In this situation, the forecast for each age group is merely a straight line drawn through the first and last observed data point and continued into the future. The independence assumption can be seen by the forecast from one age group being "unaware" of the forecasts from the other age groups. A consequence of these two properties is that, except in the knife-edged case where all the lines happen to be exactly parallel, the *time-series plots of age groups will always fan out after a point*, or in other words *the age profiles of log-mortality will always eventually become less smooth over time*. The fanning out of the Lee-Carter forecasts can be seen clearly in all-cause male mortality in New Zealand and Hungary (the left graph in the first two rows of figure 2.6) and female mortality from digestive disease (the left graph in the second row of figure 2.7). Age group forecasts that fan out have age profiles that become progressively less smooth over time, as can be seen in the increasingly exaggerated age profile graphs in each of these examples. These patterns account for the vast majority of the cross sections in our data set.

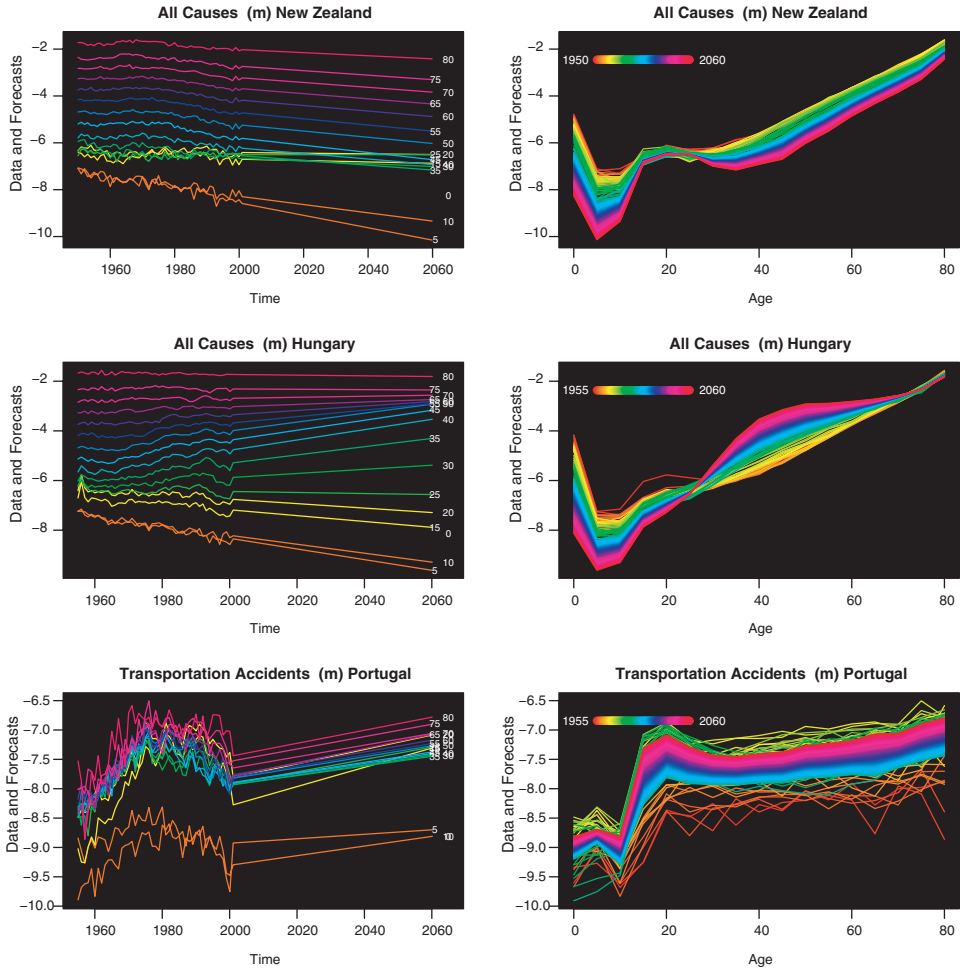

FIGURE 2.6. Data and Lee-Carter forecasts by age and time, part I.

In other data sets, the forecast lines converge for a period, but after converging, the lines cross and from then on fan out too—as in male suicide in the United States (figure 2.7, row 1, left graph). For data like these, the age profile pattern (in the right graph) inverts, with the forecasted pattern the opposite of that indicated in the observed data. In most circumstances, this inversion would be judged to be highly implausible.

The knife-edged case of exactly parallel time-series forecasts is very rare, but we found one that was close: male transportation accidents in Portugal (figure 2.6, row 3, left graph). The forecast lines do not fan out (much) in this data set, and so the age profile stays relatively constant. Coincidentally, however, this example also dramatically illustrates the consequences of a forecasting method that ignores all but the first and last data points. In this case, it misses the sharp downward pattern in the data in the last 20 years of the series. Using such a method in this data set would also ignore a great deal of accumulated knowledge about predictable mortality transitions in countries developing economically. Of course, no demographer (and especially not Lee and Carter) would forecast with a

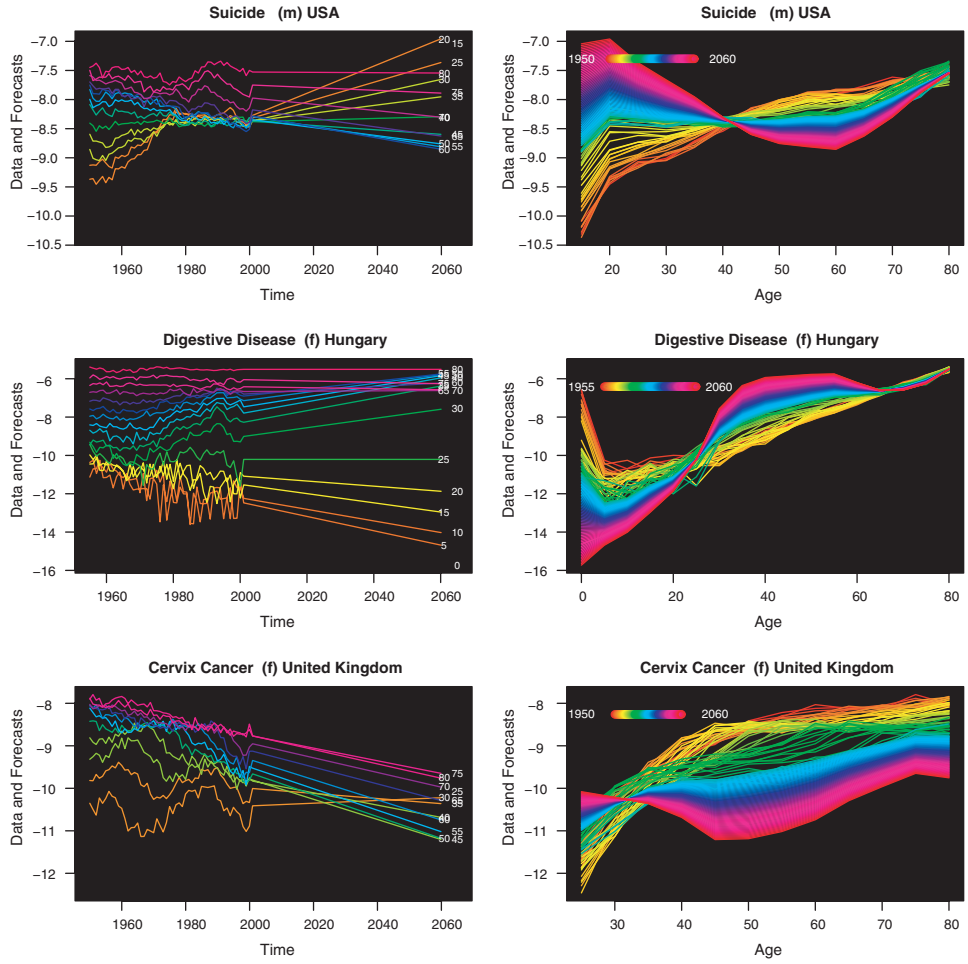

FIGURE 2.7. Data and Lee-Carter forecasts by age and time, part II.

linear model like this in such data, in part because mortality from transportation accidents typically follow a fairly well known up and then down pattern over time. Mortality from transportation accidents, which is predominantly road traffic fatalities, is almost always low when the means of transportation are fairly primitive. Then, as the use of cars increases, mortality steadily rises. Finally, as roads are built, road safety is improved, traffic control measures are implemented, and safer vehicles are imported and then required, deaths from transportation accidents decline. The same inverted U-shape can be seen in many countries, and should obviously be taken into account in any forecasting method. Indeed, from the patterns in many data sets, we should also expect in forecasting that the degree of noise around this inverted U-shape is fairly large, relative to many other causes of death. This noise too should be taken into account in forecasting. Obviously, in the case of transportation accidents we should know from the start not to use a method that makes linearity assumptions. One must also pay close attention to noise or measurement error, because even a single errant data point at the start or end of a data series can send the forecasts off in the wrong direction.

Except in the knife-edged case, the independence of the separate age group forecasts frequently produces implausible changes in the out-of-sample age profiles. We have already seen the dramatic example of suicides in U.S. males. For another example, consider forecasts of all-cause male mortality in New Zealand (figure 2.6, first row). In these forecasts, the lines crossing in the left graph produce implausible patterns in the age profiles, which can be seen in the right graph, with 20-year-olds dying at a higher rate than 40-year-olds. Mortality from cervix cancer in the United Kingdom (figure 2.7, last row) is another example with implausible out-of-sample age profiles. Cervix cancer is a disease known biologically to increase with age, with the rate usually slowing after menopause. Although this familiar pattern can be seen in the raw data in the right graph, the forecasts have lost any biological plausibility.

2.7 Summary

This chapter includes our attempt to extract and illuminate the key insights of statistical methods of forecasting mortality that do not use covariates. We build our model in part II by incorporating, formalizing, and generalizing these insights.

For almost two centuries, the driving goal in the literature has been to simplify data patterns by reducing the dimensionality in the data. Discovering a method that accomplishes this task has the potential to uncover the deep structure that is likely to persist and discard the idiosyncratic noise that can invalidate forecasts and complicate models. The idea of dimension-reduction comes from knowledge, based on considerable experience of scholars poring over data from all over the world, that log-mortality rates tend to be smooth over age groups and follow recognizable but differing patterns over causes of death, sex, and country. The literature has used a variety of techniques to try to model these patterns, but as we show in this chapter, none is sufficient to the task. No known low-dimensional parametric function fits in all situations, or even a predictable subset of applications. Similarly, no uniformly applicable nonparametric dimension-reduction technique that has been tried (such as principal component analysis) is up to the job for the highly diverse applications of interest. Yet, the goal of reducing the data to its core elements and producing forecasts that include known smoothness features of the data remain critical. The methods we introduce in part II are designed to accomplish these tasks, and to enable researchers to include additional information along similar lines, such as information based on similarities across countries and time.