Demographic Forecasting[1]

Federico Girosi[2] Gary King[3]

March 8, 2006

with contributions from Kevin Quinn and Gregory Wawro

[1]This is an early draft; we would be especially grateful if you would send us any comments or questions you might have. The current version of this manuscript is available at `http://GKing.Harvard.edu` and is meant to be printed in color. Please be aware that the lack of standardization in digital color representation means that colors in this manuscript may differ from what we intend depending on your computer, screen, software settings, and printer.

[2]Policy Researcher, The RAND Corporation; and Institute for Quantitative Social Science, Harvard University (The RAND Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138; `http://www.ai.mit.edu/people/girosi/, girosi@rand.org`).

[3]David Florence Professor of Government, Harvard University (Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; `http://GKing.Harvard.Edu`, `King@Harvard.edu`, (617) 495-2027).

Contents

List of Figures

List of Tables

Preface

We introduce a new framework for forecasting age-sex-country-cause-specific mortality rates that incorporates considerably more information, and thus has the potential to forecast much better, than any existing approach. Mortality forecasts are used in a wide variety of academic fields, and for global and national health policy making, medical and pharmaceutical research, and social security and retirement planning.

As it turns out, the tools we developed in pursuit of this goal also have broader statistical implications, in addition to their use for forecasting mortality or other variables with similar statistical properties. First, our methods make it possible to include different explanatory variables in a time series regression for each cross-section, while still borrowing strength from one regression to improve the estimation of all. Second, we show that many existing Bayesian (hierarchical and spatial) models with explanatory variables use prior densities that incorrectly formalize prior knowledge. Many demographers and public health researchers have fortuitously avoided this problem so prevalent in other fields by using prior knowledge only as an ex post check on empirical results, but this approach excludes considerable information from their models. We show how to incorporate this demographic knowledge into a model in a statistically appropriate way. Finally, we develop a set of tools useful for developing models with Bayesian priors in the presence of partial prior ignorance. This approach also provides many of the attractive features claimed by the empirical Bayes approach, but fully within the standard Bayesian theory of inference.

Software

Accompanying this book is an easy-to-use (free and open source) software package that implements all our suggestions (see Appendix F for a description and `http://GKing.Harvard.edu/` for a copy). 'The software is entitled 'YourCast: Time Series Cross-Sectional Forecasting with Your Assumptions" to emphasize a key point of our book, that choices about the assumptions made by our statistical model are governed entirely by your choices, and the sophistication of those assumptions and the degree to which they match empirical reality are, for the most part, limited only by what you may know or are willing to assume rather than arbitrary choices hidden behind

a complicated mathematical model. Although some of the tools we introduce require technical sophistication to implement, the ideas are conceptually straightforward. As such, the software and methods should be usable even by those who decide not to digest all of our detailed mathematical arguments.

Background

The methods developed in this book rely on fields of statistics and mathematics, at least some of which are likely to be unfamiliar to many interested in mortality forecasting. Yet, given the highly important public policy issues at stake, the advantage to scholars and citizens of any forecasting improvement, even when achieved via unfamiliar mathematical techniques, should, in our view, outweigh higher costs to researchers in learning the methods. We have thus not shied away from introducing new methods but have tried to reduce the associated costs to researchers in a variety of ways. Most importantly, we explain our methodology in a way that should make all of our results accessible to those who are familiar only with linear regression analysis and the basics of Bayesian inference. We also include in Appendix A a detailed glossary of notation. In addition, since different aspects of the necessary mathematical background are likely unfamliar to different audiences, we offer an extensive mathematical refresher in Appendix B that should be almost entirely self-contained.[1]

Although we have attempted to keep the book as readable as possible, we have also included, for more mathematically sophisticated readers, all necessary proofs and evidence so that the work would be relatively self-contained.

Publications

In some of the fields with which the content of this book intersects, almost all new results appear first in articles. Book presses are left to print only texts summarizing prior research. In this project, we resisted the temptation to send preliminary or partial results to scholarly journals since we felt the whole of our book would be greater than the sum of the parts, sliced into smaller articles. Thus, although we have occasionally presented preliminary results in talks over the last five years, this manuscript is the first complete account of our approach.

[1] To be specific, our goal was to make the book readable by anyone with background equivalent to Government 2001 taught by King at Harvard. This class covers likelihood models, simulation, and the basics of Bayesian inference. We tested this hypothesis on several students from that class, and added explanations for concepts with which they felt uncomfortable. The syllabus and detailed lecture notes for this class are available at `http://gking.harvard.edu/class.shtml`. The book is of course an easier read for those who also had a full length class in Bayesian modeling.

Acknowledgements

We are especially grateful for the contributions of Kevin Quinn and Greg Wawro. Although they should not be held responsible for our errors or arguments, we could not have written this book without what we learned from them. Wawro and Quinn were King's post-doctoral fellows before Girosi joined the project at what is now the Institute for Quantitative Social Science (in the 1998–1999 and 1999–2000 academic years, respectively), and each contributed a great deal. Greg Wawro replicated Murray and Lopez's (1996) forecasts, uncovered some surprising errors (traced to a remarkably basic error in the SAS Statistics package) and managed to improve substantially the quality of WHO's mortality database. He also demonstrated the failure of a whole range of classic econometric tests and techniques when applied to our data: Paradoxically, we found that standard tests frequently confirmed the presence of problems such as cointegration, nonstationarity, parameter heterogeneity, nonrandom missingness, dynamic processes, and omitted variable bias, but following the textbook econometric correction for each or a variety of off-the-shelf approaches degraded or only slightly improved the forecasts. We now accept new model features only if they demonstrate their worth by improving out-of-sample forecasting performance, specification tests for the various econometric problems being largely besides the point.

Kevin Quinn had the vision to see the connection between what we were trying to accomplish and the literature on Markov random fields. He worked out how to extend the Markov random field approach, designed for modeling physical space, to processes that varied over conceptual space (such as age). He also designed clever ways to implement these ideas via Gibbs sampling in Gauss code. These contributions were invaluable.

We also thank Chris Murray, who originally challenged us to invent a method of forecasting mortality that out-performed existing approaches — including his own — and his office at the World Health Organization kept us supplied with data and research support until we were successful. Chris, along with David Evans, Majid Ezzati, Alan Lopez, Colin Mathers, and Josh Salomon, taught us a great deal about mortality data and patterns. They were all especially helpful during the sessions we had pouring over thousands of forecasts from successive versions of our model. To say they were good sports every time we showed up in Geneva lugging two linear feet of printouts in a tiny font is no small understatement.

This project is also an outgrowth of a National Institute of Aging grant on the "Global Burden of Disease in Aging Populations" (P01 AG17625-01). The frequent meetings of this distinguished group of scholars have provided invaluable feedback on our work and, through site visits around the world, have greatly informed our analyses. Our thanks to the other project leaders on this grant, David Cutler, Ken Hill, Alan Lopez, Chris Murray, and Richard Peto, to Richard Suzman at NIA, and to the many other participants in our meetings.

Our thanks to Emmanuela Gakidou, James Honaker, Catherine Michaud, Ken

Scheve, Ajay Tandon, and Neils Tomijiman for logistical help with and insightful comments on our work. Thanks also to Sam Abrams, Chris Adolph, Marcus Augustine, Anders Corr, Alexis Diamond, Suzanne Globetti, Mie Inoue, Ethan Katz, Ryan Moore, Claudia Pedroza, Nirmala Ravishankar, Heather Stoll, and Liz Stuart for superb research assistance, and Anita Goldpergel and Elena Villalon for expert programming assistance. Our appreciation goes to Gadi Geiger for his mechanical engineering contributions. For helpful discussions on previous drafts or talks, we thank Barbara Anderson, Heather Booth, Brad Carlin, Majid Ezzati, John Maindonald, Jamie Robins, Don Rubin, Len Smith, Liz Stuart, Leonie Tickle, and Chris Zorn. Our special thanks to Ron Lee and Nan Li for their help with the demographic literature and insights into their models and perspectives. Micah Altman, Shawn Bunn, Matt Cox, and William Wei at the Harvard-MIT Data Center provided a series of first rate computational environments that always matched our quickly evolving needs. Thanks also to Jaronica Fuller for consistently cheering up everyone within miles of our offices, and Marie Cole, Beverly MacMillen, and Kim Schader (King's assistants) for organizing the team we assembled and making difficult administration issues transparent. For research support, in addition to the National Institutes of Aging, we are grateful to the National Science Foundation (SES-0112072, IIS-9874747), the World Health Organization, and, at Harvard, the Weatherhead Initiative, and the Institute for Quantitative Social Science.

Chapter 1

Qualitative Overview

1.1 Introduction

This book introduces a set of methodological techniques designed to build on and contribute to parts of the scholarly disciplines of demography, statistics, political science, macro-epidemiology, public health, actuarial science, regularization theory, spatial analysis, and Bayesian hierarchical modeling. Our approach would also seem applicable in a variety substantive research applications in these and other disciplines. In this chapter, we describe our intended contribution in four ways, the last three being consequences of the first.

The narrowest view of this work is an attempt to address the goal we originally set for ourselves: to create *a new class of statistical methods for forecasting population death rates* that out-performs existing alternatives — by producing forecasts that are usually closer to out-of-sample mortality figures in large scale comparisons, by more reliably fitting well-known patterns in mortality data as they vary over age groups, geographical regions, and time, and generally by incorporating more available quantitative and qualitative information.

Mortality analyses are of widespread interest among academics, policymakers, industry researchers, and citizens worldwide. They are used for retirement fund planning, for directing pharmaceutical research, and planning public health and medical research interventions. They constitute our primary running example and source of data for empirical validation. The World Health Organization (WHO), which has made use of our methods, relies on worldwide mortality forecasts to estimate morbidity (by using the well-known relationship between death rates and the prevalence of certain diseases), to make ongoing public health recommendations to specific countries and regions, and as a source of information for the health ministries in member countries.

The class of methods we introduce are also applicable to a wider range of problems than mortality forecasting. In particular, any research that uses a method like

linear regression for more than one cross-section or time series may benefit from our approach. Although the only data we present in this book are about mortality, the methods would seem to be directly applicable to forecasting variables from classical demography and macro-epidemiology, such as fertility rates and population totals; economics, such as income and trade; political science, such as electoral results; and sociology, such as regional crime rates. Section 1.2 elaborates.

In order to accomplish our original goal of forecasting mortality rates, we found it necessary to develop a series of new tools that turn out to have wider implications. These new tools also suggest alternative ways of understanding the content of this book.

Thus, a second and broader view of this book is as *a better way to specify and run a set of linear regression models, one that improves on existing Bayeisan approaches and automates what would otherwise be highly time consuming decisions*. Our key methodological result is demonstrating how to borrow strength by partially pooling separate cross-sections based on similarities across only the expected values of the dependent variables instead of the traditional practice of pooling only based on similar coefficients. This result also enables us to pool data from regressions that have different explanatory variables in each cross-section, or the same explanatory variables with different meanings. Since in many applications, researchers can directly observe the dependent variable, and never actually see the coefficients, this approach makes it possible to base prior distributions on known information rather than optimistic speculation. We also offer several new ways of building priors that better represent the types of knowledge substantive experts possess and, as importantly, new ways of modeling ignorance and indifference, and understanding when they are appropriate in Bayesian models. These methods incorporate and extend the key attractive features of empirical Bayes models, but without having to resort to a theory of inference outside the standard Bayesian framework.

A third way of understanding our work is an implication of the second: We show that *the most common Bayesian method of partially pooling multiple coefficients in cross-sections thought to be similar is often inappropriate as it frequently misrepresents prior qualitative knowledge*. The idea of pooling is correct and can be powerful, but many Bayesian models with covariates in the literature are flawed in this way. This claim affects several scholarly literatures in statistics and related fields, such as at least some work in hierarchical modeling, spatial smoothing, and applications in the social and natural sciences. We provide the mathematical and statistical tools to correct for this problem in theory and practice.

A final way to view this book is as a small step in *reconciling the open warfare between cross-national comparativists in political science, sociology, public health, and some other fields with the area studies specialists that focus on one country or region separately from the rest of the world*. In political science, for example, the current animosity between quantitative cross-national comparativists and area studies scholars originated in the expanding geographic scope of data collection in the 1960s. As

quantitative scholars sought to include more countries in their regressions, the measures they were able to find for all observations became less comparable, and those that were available (or appropriate) for fewer than the full set were excluded. Area studies scholars appropriately complain about the violence these procedures do in oversimplifying the reality they find from their in-depth (and usually qualitative) analyses of individual countries but, as quantitative comparativists continue to seek systematic comparisons, the conflict continues. By developing models that enable comparativists to include different explanatory variables, or the same variables with different meanings, in the time-series regression in each country, we hope to eliminate a small piece of the basis of this conflict. The result should permit statistical analyses and data collection strategies that are more sensitive to local context and that include more of the expertise of area studies specialists. Indeed, even if the area studies specialist in each country would prefer a unique set of explanatory variables, our methods enable a scholar to estimate all these regressions together, marshaling the efforts of many scholars and enabling them to work together without sacrificing what they they each bring to the table.

We now discuss each of these four perspectives on our work in turn, saving technical discussions for subsequent chapters.

1.2 Forecasting Mortality

Almost all countries — democracies and authoritarian regimes, rich countries and poor countries, nations in the north and those in the south, etc. — make efforts to improve the health of their populations. Indeed, over nine percent of the world's economy (and fifteen percent of the U.S. economy) is devoted to health care spending. As in other areas of public policy, information about the problem can help tremendously in amelioration efforts. For this reason, WHO has regularly forecast mortality and morbidity, with increasing geographic resolution over time, for the entire world. These forecasts are used by WHO, other international institutions, donor countries, and the health ministries and public health bureaucracies within each country to direct the flow of funds in the most effective way possible to the population groups in most need or which can be helped the most. Mortality forecasts are also used to assess the future security of retirement and social security plans, public and private insurance schemes, and other public policies that depend on specific population and mortality counts.

In recognition of the value of this information, but also for other unrelated reasons, enormous quantities of money are spent on vital registration systems in many countries. For example, in the United States, laws in each of the fifty states require death certificates be completed for every person who dies. Federal law then mandates the central compilation and publication of these data, and other vital statistics. Each of the 779,799,281 deaths recorded in our database was (in principle) coded from an official death certificate. Vital registration systems around the world are not of

uniform quality, and examples of systematic bias even from developed countries can be identified. After all, many registration systems were developed for purposes more related to the administrative, taxation, or representative functions of state than for monitoring, forecasting, or improving health care. Yet, most such systems appear to be steadily getting better. International agreements exist on common definitions of causes of death, and on proper procedures for data collection. And in any event the care and resources that goes into collecting these data dwarfs most other variables measured in the social sciences.

1.2.1 The Data

Our mortality data have the following structure. For 191 countries, 24 causes of death[1], 17 age groups, and 2 sexes, we observe an annual time series of the number of people who have died and the population in that subgroup. For our purposes, the death rate is of interest: the number of people who have died divided by the number of people at risk of dying (the population). The time series of death rates for each of these 155,856 cross-sectional units usually ends in 2000, and is between 2 and 50 observations long. Time series cross-sectional data are commonly used in the social sciences. Less common are data like these which have four cross-classified cross-sections for each time period. The methods we develop here usually help more with larger numbers of cross-sections, but they can be advantageous with as few as two.

For each year, sex, age, country, and cause, we also observe several covariates (explanatory variables). When they are all observed, we have gross domestic product (adjusted for purchasing power parity), tobacco consumption (in some cases based on direct information and in others inferred based on present lung cancer rates), human capital, total fertility, fat consumption, a time trend as a rough measure of technology, and the levels of freedom and democracy in each country.

Enormous effort worldwide went into producing and collecting these data, but much still remains missing. Every country in the world has at least two observations, but only 34 countries have at least 35 time series observations; 17 have 20–34 observations; 33 have 2–19; and 107 have only 2 observations. Many of the countries with only two "observations" were imputed by WHO experts.

In Figure 1.1 we provide a graphical representation of the distribution of the number of observations for all-cause mortality. The red points in the graph represent the percentage of countries for which the number of observations is larger than a given

[1]The 24 categories of death are called "clusters" in the international classification of diseases. These include all-causes (which is the sum of all other causes), malaria, AIDS, tuberculosis, other infectious diseases (i.e,. other than malaria, AIDS and tuberculosis), lung cancer, cancer of mouth and esophagus, liver cancer, stomach cancer, breast cancer, cervix cancer, other malignant neoplasms, infectious respiratory disease, chronic respiratory disease, cardiovascular disease, digestive disease, maternal conditions, perinatal conditions, all other diseases, transportation accidents, other unintentional injuries, suicide, homicide, and war.

amount, read on the horizontal axis. In green we show the percentage of the world population living in those countries. For example, the figure shows that 50–60% of the world's population lives in countries whose age-specific mortality time series has more than 10 observations. This implies that any forecasting method which is not applicable to problems with fewer than 10 observations will fail to make forecasts of any kind for 40-50% of the world's population.

Africa, AIDS, and malaria are the areas with the most missing data. Data are usually good for OECD countries but sparse in the rest of the world. If mortality is observed for any age, it is (almost) always observed for all ages. All-cause mortality (the number of people who die regardless of cause) is often observed for more countries than cause-specific mortality, although counts for some diseases are observed when all-cause is not; if any cause is observed, all causes, ages, and sexes are normally observed. We treat the covariates as fully observed, if they exist in a cross-section at all, although parts are exogenously imputed based on statistical techniques in combination with expert judgment.

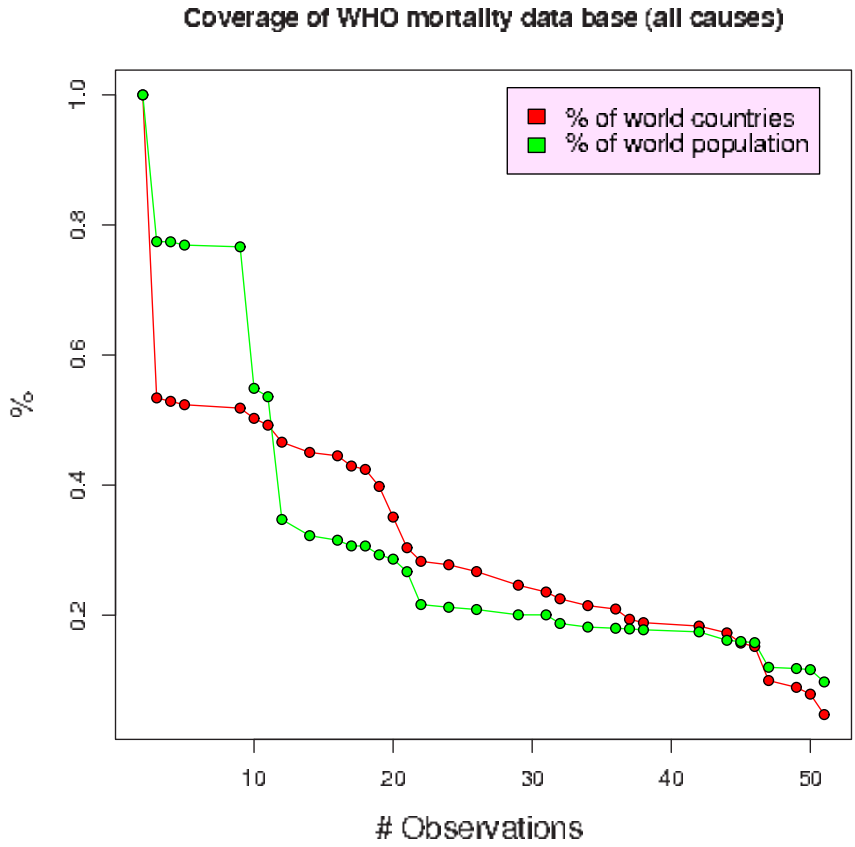

Figure 1.1: Distribution of number of observations for all-cause mortality: the vertical axis represent the percentage of countries (and the percentage of the world population living in those countries) whose time series have a number of observations smaller than a given amount, read on the horizontal axis.

The noise in death rates appears to be mainly a function of the size of the cross-sectional unit, so that small countries, younger age groups, and rarer causes of death are associated with more noise. Measurement error, which is one contributor to noise, is somewhat more prevalent in smaller cross-sections, but many exceptions exist; the real cause would appear to be underfunded data collection facilities in less developed countries. Data outside the OECD also tends to be noisier. Our judgment, based on interviews and some statistical tests, is that fraud does not seem to be a major problem.

1.2.2 The Patterns

Health policymakers and public health, medical, and pharmaceutical researchers are primarily interested in cause-specific, rather than total, mortality. They need cause-specific information to direct appropriate treatments to population subgroups, and so that they might have a chance at understanding the mechanisms that give rise to observed mortality patterns. Researchers also use well-known relationships between cause-specific mortality and some (ultimately fatal) illnesses to estimate the prevalence of these illnesses.

Others, such as economists, sociologists, actuarial scientists, insurance companies, and public and private retirement plans, are primarily interested in total or "all-cause" mortality. The cost to a retirement plan, for example, does not change with changes in the causes of death unless the compositions of cause-specific mortality add up to different totals. However, those interested in forecasting only all-cause mortality are still well-advised to examine closely, and attempt to forecast, cause-specific mortality. For example, the U.S. Social Security Administration takes into account forecasts of the leading causes of death in their all-cause forecasts, albeit in an informal, qualitative way.

We now offer, in Figure 1.2, an overview of the leading causes of death worldwide in the year 2000. In each graph, we have included each cause of death for all age groups that is the leading cause for at least one age group. The graph on the left is for males and shows that worldwide, the leading cause of death for males aged 10 and under is infectious diseases other than AIDS, malaria, and tuberculosis. After age 10, death due to this cause declines and then increases over age groups, but it is not the leading cause of death for any other age group. From 10 to 20, the leading cause of death is transportation accidents; these increase before declining, but AIDS takes over until about age 35 as the leading cause of death among males. For older aged men, cardiovascular disease is the leading killer. The patterns are similar for females, although slightly lower than for males. In addition, transportation accidents never rise to the leading cause and so do not appear.

Figure 1.2 also demonstrates that different causes of death have widely divergent age profiles. Some have mortality rates that start high for infants and rise as they age. Some drop and then rise, and some follow still other patterns. However, through

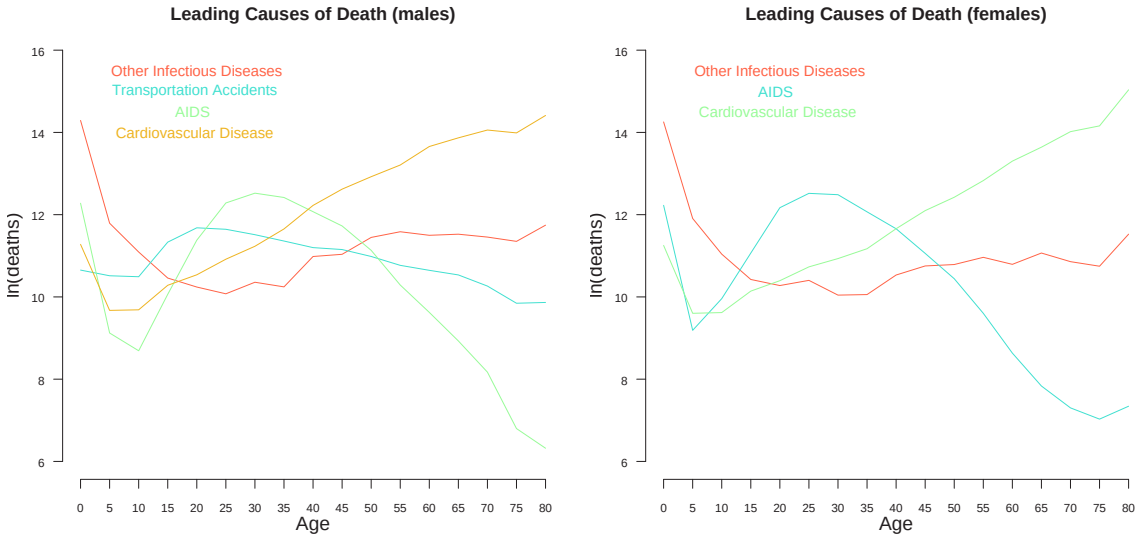

Figure 1.2: Leading Causes of Deaths Worldwide by Sex. Logarithm of the number of deaths worldwide by leading cause for males (on the left) and females (on the right). A cause of death is included in each graph if more members of the respective sex died from it than from any other cause for at least one age group.

all this diversity, a clear pattern emerges: *The age profiles are all relatively smooth.* Five-year-olds and 80-year-olds die at very different rates, but people in any pair of adjacent age groups die of similar rates from any cause.

We now illustrate the same point about the diversity and smoothness of log-mortality age profiles by offering a broader picture of different causes of death. For this figure, we change from the log of deaths to the log of the deaths per capita, otherwise known as the *log-mortality rate*. We average the age profile of the log-mortality rate over all available years and all 191 countries in our database. Figure 1.3 presents these average age profiles for the 23 causes of death in females, one in each plot. Figure 1.4 offers a parallel view for the 20 causes of death in males. (We exclude perinatal conditions in both since the vast majority of the deaths are in the first age group.) Age groups are 5 year increments from 0 to 75 and 80+, and are labeled by the lower bound.

These figures provide a baseline for understanding the units of log-mortality for our subsequent analyses. They also clearly demonstrate that an appropriate method of forecasting mortality must be capable of modeling great diversity in the age profile across diseases (and sex and country) while at the same time guaranteeing that log-mortality remains smooth over the age profile. All 43 age profile graphs are fairly smooth, with some exceptions for younger age groups, but they follow a large diversity of specific patterns.

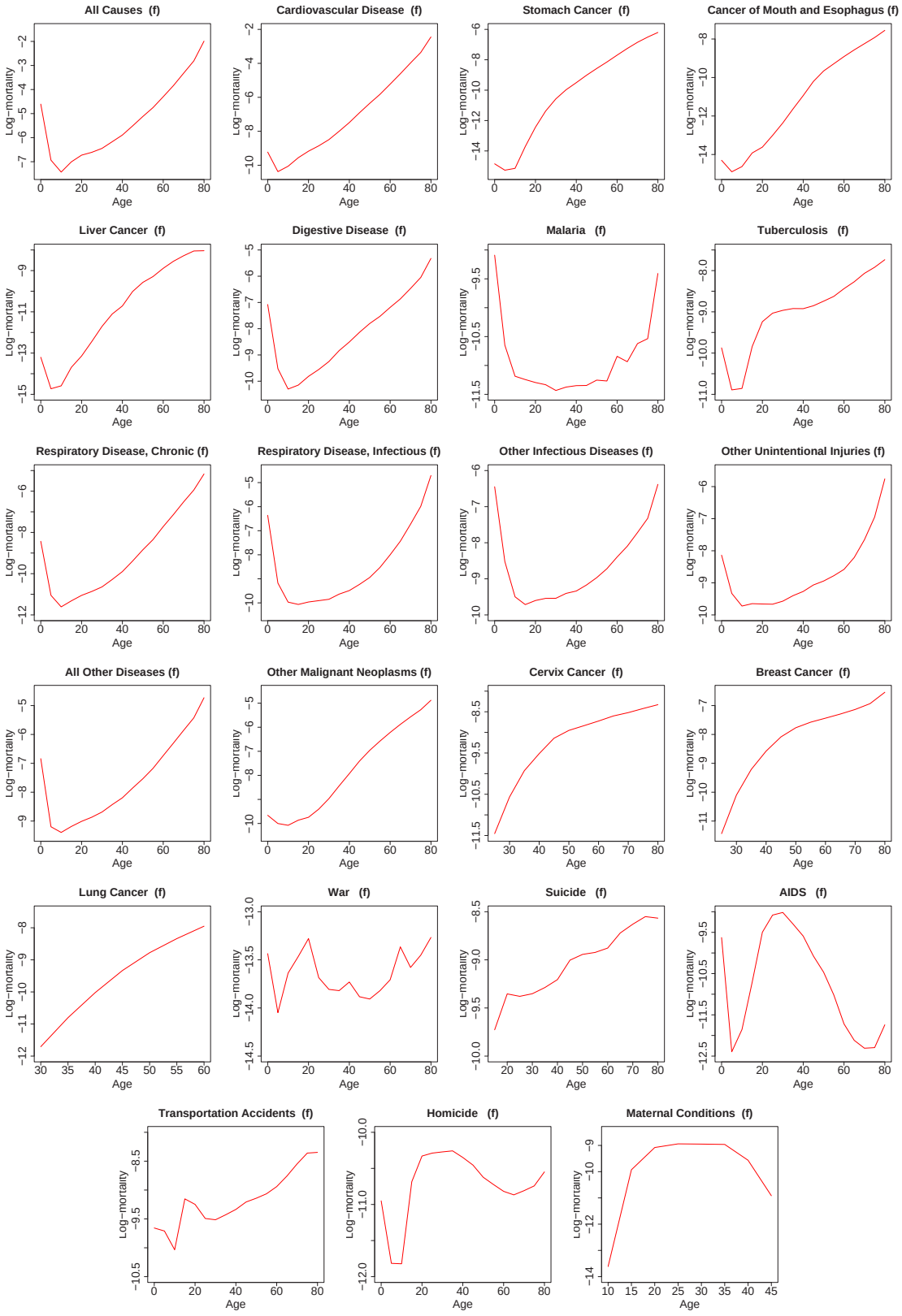

Figure 1.3: World Age Profiles for 23 Causes of Death in Females. The age profiles have been averaged over all 191 countries and over all available years.

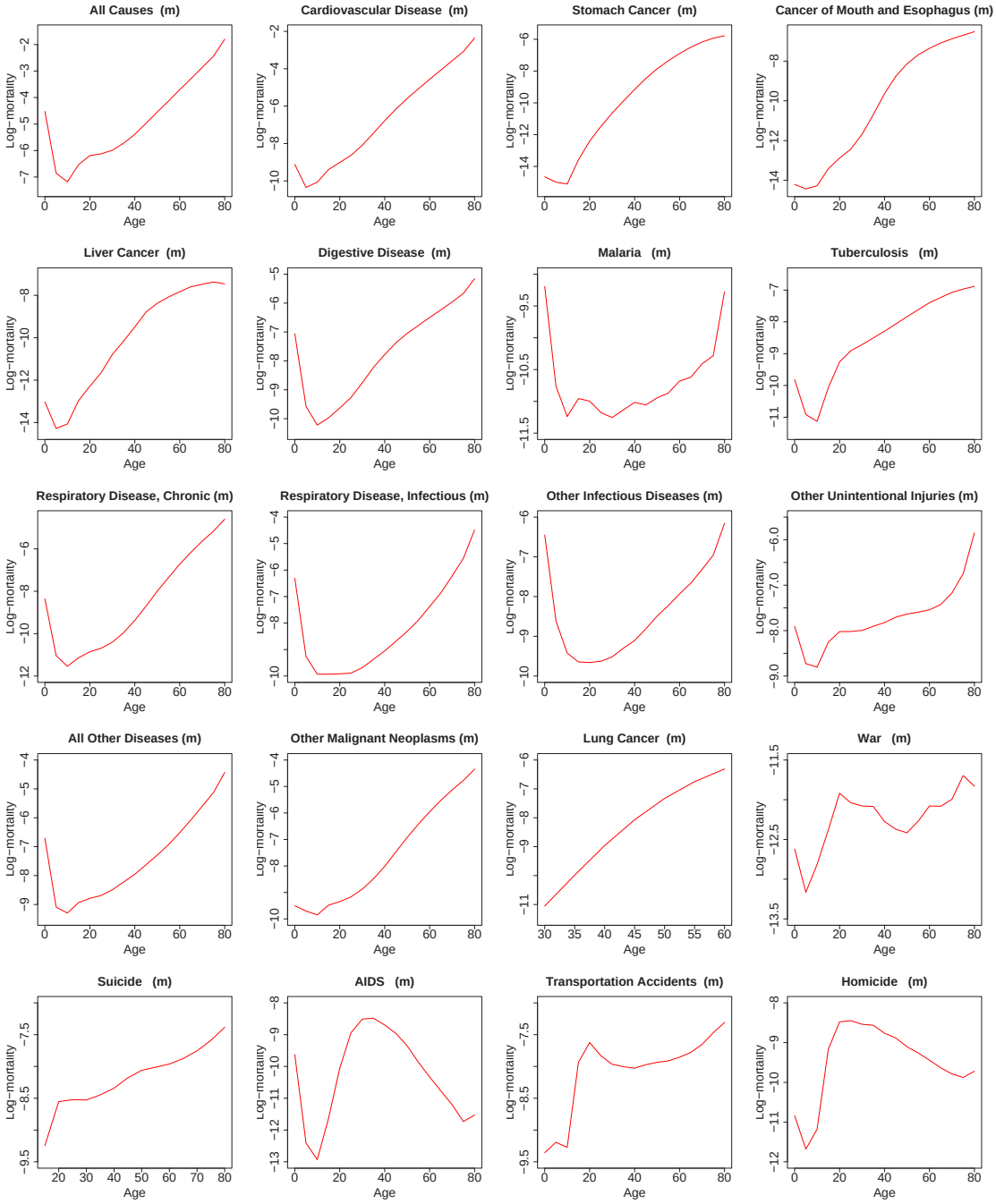

Figure 1.4: World Age Profiles for 20 Causes of Death in Males. The age profiles have been averaged over all 191 countries and over all available years.

1.2.3 Scientific vs. Optimistic Forecasting Goals

A list of all valid methods of learning about the future developed thus far is as follows: waiting. As a methodology, waiting is easy to apply (for some!), requires no special expertise, and is highly certain. Unfortunately, whether researchers make forecasts or not, public policy makers will not wait since getting the future wrong will often produce enormous costs that can be denominated in dollars, deaths, and disease. Policymakers may get it wrong, but they certainly will try. As such, and despite the uncertainties, researchers could hardly have more motivation to make forecasts better and to try to formalize and improve what is often informal qualitative guesswork. Indeed, almost any advance in learning about future mortality rates has the potential to inform and rationalize health care spending, mobilize resources to improve care where it is needed, and ultimately to reduce mortality.

Since modern scientific forecasting is not soothsaying, we need to be clear about what it can and cannot accomplish, even at its best. So here is our definition:

> Forecasting is a (1) systematic distillation and summary of all relevant information about the present that (2) by assumption may have something to do with the future.

Part (1) is the forecasters' first and only real accomplishable scientific goal. Part (2) is pure assumption that is by definition beyond the range of, and not fully defensible on the basis of, the data. Even genuine repeated out-of-sample forecasts that turn out close to the observed truth have no necessary relationship whatsoever to subsequent forecasting success.

Put differently, science is about collecting the facts and the deeper understanding that explains the facts or that the facts imply. Forecasting involves more than science. It also involves unverifiable assumptions, justified by argument, analogy, and reasoned speculation that by definition goes beyond direct evidence. The best forecasts, then, are those that provide the best systematic summaries of all that is known, and also your assumptions (which is why we call our software YourCast). The purpose of this book is to provide methods that make it easy to include more existing information, in as many forms as possible, in our forecasting models so that we can better summarize the present and quantify and understand our assumptions about the future.

Our public policy goal is to provide forecasts for the public health establishment that are based on more information than those now used and are more systematic. The death rate forecasts of interest are those up to 30 years in the future for each time series, with the needed forecast horizon often differing by cause of death. In most applications, time series analysts would not normally be willing to make 30-year-ahead forecasts on the basis of even 50 observations, and we often have fewer. The difference here is that death rates usually move slowly and smoothly over time so that forecasting with low error rates at least a few years out is usually easy. Moreover, when considered as a set, and combining it with qualitative knowledge experts in the field have gathered from analyses of numerous similar datasets, the information

content availabe for making forecasts is considerably greater than when treating each time series separately and without qualitative information.

But the uncertainties should not be overlooked, and valid thirty year forecasts are not likely to be accurate with any known degree of certainty. New infectious diseases can spring up, mutate, and kill millions, as occurred with the HIV pandemic and the 1918 flu, or affect few, as in the SARS outbreak in 2003. Unexpected developments in medical technology can drastically reduce the effects of a disease, as occurred with HIV treatments in the developed countries in the last decade or the eradication of smallpox, and the reduction in polio and shistosomiasis, throughout the world. Warfare and international terrorism have demonstrated potential to wreak havoc without advance notice. Changes can even occur without a cause that we can unambiguously identify, as with the dramatic increase in cardiovascular disease in Eastern Europe in the late 1980s. Physicians are not very good at predicting the time of death of individuals with serious illnesses, and no law of nature guarantees that errors will always cancel out when individuals are aggregated. As Lee and Carter write (1992, p. 675), almost tongue in cheek, "Perhaps we should add the disclaimer that our [confidence] intervals do not reflect the possibilities of nuclear war or global environmental catastrophe, for indeed they do not." In fact, model-based confidence intervals can never be accurately calibrated since the model can never be counted on to reliably include all relevant future events. Forecasts are by definition extrapolations outside the range of available data (King and Zeng, 2005).

These qualifications put all forecasting methods in context. In practice, mortality forecasts are always at best conditional on the state of knowledge we have today. In this book, we go farther than existing methods primarily by incorporating a larger fraction of (known) information than prior approaches, but we obviously cannot include information that is unknown at the time of the forecasts. Thus, to make the same point in yet another way, the forecasts produced by the methods in this book, and all other methods in the literature, can only hope to tell us what will happen in the future if current conditions and patterns hold. If future mortality is driven by factors we have not included, or if the relationship between the factors we include and mortality changes in unexpected ways, our forecasts will be wrong. Moreover, since conditions do change, and will likely change, we do not expect forecasts made today to be accurate. Indeed, much of the stated purpose of policy makers and funding agencies in the health arena is to change these conditions to reduce future mortality rates. To put it even more starkly, the central goal of the global public health, medical, research, governmental, and intergovernmental infrastructures is to make our forecasts wrong! We hope they succeed.

The fact that even we expect that our forecasts will be proven incorrect does not mean that our methods and forecasts should be considered free from scientific evaluation. Far from it. The only issue is finding the right standard. True out-of-sample forecasts are infeasible since it would take several years for data to come in and mortality forecasts are needed immediately and continuously. And even if we waited,

had our forecasts confirmed, and decided to use those forecasts for policymaking, nothing could stop the world from changing at that point to make our subsequent forecasts inaccurate. Since policymakers are effectively trying to do this anyway, actual out-of-sample forecasts are not that useful here.

The optimal scientific standard for us would be to make actual out-of-sample forecasts in an (infeasible and unethical) experiment where some country is held to have the same standards, conditions, types of medical care, etc., in a future period as during the period from which our data come. But although this hypothetical experiment would validate the model, it is obviously of no practical use. To approximate this situation, we set aside some number of years of data (the "test set" or "in-sample period"), fit the model in question to the rest of the data (the "training set" or "out-of-sample period"), and then compare out-of-sample forecasts to the data set aside. This standard makes us somewhat less vulnerable to being proven wrong than we would with real out-of-sample forecasts, which is a perhaps necessary flaw in our approach. Nevertheless, the enormous number of forecasts we need to compute would make any attempt to stack the deck extraordinarily difficult, even if we had set out to do it intentionally. Although we have run hundreds of thousands of such out-of-sample comparisons between our approach and previous methods offered in the literature, and although earlier drafts of this manuscript recorded summaries of many of these runs, we exclude them here. The reason is that our method does better because it includes more infomration and so these tests merely demonstrate the predictive value of available information. This makes comparisons either an unfair fight — since we use more information than relevant competitors — or merely an evaluation of the quality of available information — neither of which are of much use here. In addition, as explained above, forecasts always require choices about assumptions regarding the future, and we wished to evaluate our methods, not our particular assumptions about the future. Thus, with this book we are releasing our full original data set and our software program that makes it easy for readers to choose their own assumptions, make their own forecasts, and if they wish make their own assessments of their forecasting accuracy when using our methods with their assumptions.

1.3 Statistical Modeling

We offer here a qualitative overview of the statistical modeling issues we tackle in this book. We discuss this modeling in the context of our mortality example, although everything in this section would also apply to any outcome variable that a researcher would be willing to specify as a linear function of covariates. The methods introduced are an improvement over linear regression (and seemingly unrelated linear regressions and other related procedures) for any analysis with more than one regression over at least some of the same units. They can be extended to nonlinear models (the theory would be easy, but the computation would need to be worked out), but we have not done so here.

In our application, the over 150,000 time series created substantial difficulties in data management, computational speed, and software architecture. To put the problem in context, if we spent as much as a single minute of computational and human time on each forecast, we would need more than three months (working 24 hours a day, 7 days a week) to complete a single run through all the data using only one of the many possible specifications. Not surprisingly, we found that if a feature of a method was not automated (i.e., if it required human judgment to be applied to each cross-section), it became almost impossible to use. An important goal of the project thus became automating as much as possible. Fortunately, this is entirely consistent with, and indeed almost equivalent to, the goals of good Bayesian modeling, since automation is merely another way of saying that we need to include as much available a priori information in the model as possible. So this has indeed been our goal. Throughout our work, we have unearthed and then incorporated a considerable amount of prior knowledge in our models, which has been the primary reason for our forecasting improvements.

Demographers, in contrast, often use the vast majority of their prior knowledge as a way to evaluate, rather than to improve their methods. While this procedure can be productive, it is inferior from a scientific perspective since the method is effectively expert judgment which is never formalized in a way others can apply and so is less vulnerable to being proven wrong. At its worst, their procedure is to keep adjusting a model specification until it produces forecasts consistent with one's prior expert opinion — in which case the forecast is nothing more than an expert judgment and the statistical model is nothing more than scientific-looking but irrelevant decoration. In our view, the future of demographic modeling and forecasting lies primarily in identifying and conditioning on more and more information in statistical models. Classical demographers have gleaned enormous information from decades of careful data analyses; the key now is to marshal this for making inferences.

To explain our alternative approach, we now begin with a basic building block, a single least squares regression, and then add in other sources of information. Consider a single cross-sectional unit (one age $\times$ sex $\times$ cause $\times$ country, such as the log of the mortality rate in 45 year old South African males who die of cardiovascular disease) with a long time series of observations on the log-mortality rate and a set of covariates (explanatory variables) that code its systematic causes or predictors. If this cross-section has little missing data or measurement error, informative covariates, and a high signal-to-noise ratio[2], then least squares regression might do reasonably well. Unfortunately, in the vast majority of the time series in our data, measurement error, missing data, unexpected changes in the data, and unusual patterns conspire to make a simple least squares time series regression model produce incorrect or even absurd forecasts. Normally, the problem is not bias, but enormous variance. This problem becomes clear when one thinks of the fanciful goal of forecasting thirty years ahead

[2] "Signal-to-noise ratio" is a term from communications research, where the "signal" is the useful information and "noise" represents anything else.

on a single time series with say twenty observations or with more observations but high rates of measurement error, etc.

Thus, the key goal in analyzing these data is to identify and find ways of incorporating additional information in order to increase efficiency. The most common way to do this would be to pool. For example, we could pool all countries within a region, effectively assuming that the coefficients on the covariates for all those countries are identical. This would increase efficiency, but — if in fact the coefficients varied — it could then introduce bias.[3] Strict pooling in this way is also contrary to our prior information, which does not normally indicate that coefficients from neighboring countries would be identical. Instead of having to decide by hand which countries (or cross-sections) to pool, which is infeasible, we could think about automating this decision by relaxing the requirements for pooling. Instead of requiring neighboring countries to have identical coefficients, we could allow them to have similar coefficients. This kind of "partial pooling," or "smoothing," or "borrowing strength" from neighboring cross-sectional units to improve the estimation (and efficiency) of each one, is common in modern Bayesian analysis, and we were therefore able to build on a considerable body of prior work (which we properly cite in subsequent chapters).

Partially pooling countries by assuming similarity of coefficients in neighboring countries is an improvement, and serves to automate some aspects of the analysis. (To better reflect knowledge of public health, we also generalized "neighboring" to refer to similar countries, not always strictly adhering to geographic contiguity.) We extend this logic to combine partial pooling of neighboring countries, and consecutive time periods, simultaneously with partial pooling of adjacent age groups. The fact that 5-year-olds and 80-year-olds die of completely different causes and at very different rates would normally prevent pooling these groups. However, we also know that 5-year-olds and 10-year-olds die at similar rates, as do 10-year-olds and 15-year-olds, and 15-year-olds and 20-year-olds, etc. Thus, we simultaneously pool over neighboring countries, adjacent age groups, and time (and we allow smoothing of interactions, such as trends in neighboring age groups), to result in a form of multidimensional, non-spatial smoothing. This step also provides a more powerful approach to reducing the dimensionality of mortality data than the 175+ year tradition of parametric modeling in classical demography (Gompertz, 1825; Keyfitz, 1982).

Each of these steps generated an improvement in efficiency, fit, and forecasting stability. Yet, it was still very difficult to use with so many forecasts. We had to spend inordinate amounts of time tuning the priors for different data sets, and finding what seemed like the "right" parameters was often impossible. Eventually, we realized that the problem was with the most fundamental assumption we were making — that the coefficients in neighboring units were similar. In fact, the scale

[3] Assuming a regression coefficient is constant when it in fact varies can cause one to underestimate standard errors and confidence intervals. If in addition the actual coefficients assumed to be constant are correlated with one of the measured variables, then the estimated coefficient will be a biased estimate of the average of the true coefficients.

of most of the coefficients in our application is not particularly meaningful and so although our model had been saying they should be similar, and large literatures on Bayesian hierarchical modeling and spatial modeling smooth in this way, experts in public health and demography do not really have prior beliefs about most of these coefficients. For one, most of these coefficients are not causal effects and so are not even of interest to researchers or the subject of any serious analysis. They are at best partial or direct effects — for example, the effect of GDP on mortality, after controlling for some of the consequences of GDP, such as human capital and total fertility. This coefficient is not the subject of study, it is never directly observable, and, since some of the control variables are both causally prior and causally consequent, it has no natural causal interpretation. Moreover, even for variables about which these scholars possess considerable knowledge, such as tobacco consumption, the real biological knowledge is at the individual level, not at the aggregated national level. Finally, even when some knowledge about the coefficients exists, the prior must include a normalization factor that translates the effect of tobacco consumption, say, into the effect of GDP. Unfortunately, no information exists in the data to estimate this normalization and prior knowledge about it almost never exists.

Our software includes a facility for partially pooling coefficients, for the situations for which it may be useful (such as may be the case for case-control studies designed to estimate a specific causal effect and where base probabilities are not estimated). But we also added a new feature that ameliorates many of the remaining problems in our application, and we believe will work in many others. Thus, instead of only partially pooling coefficients, about which we had little real prior knowledge, we developed a way to partially pool expected mortality. Scholars have been studying mortality rates for almost two centuries and know a great deal about the subject. Although we do not observe expected mortality, for which our priors our constructed, every observed mortality rate is a direct and usually fairly good estimate of expected mortality. Priors formulated in this way correspond closely to the nature of the variables we are studying. This procedure also makes it substantially easier to elicit information from subject matter experts. It also directly satisfies the goal of Bayesian analysis by incorporating more prior information appropriately. And it serves our goal of automating the analysis, since far less cross-section-specific tuning is required (and many fewer hyperparameter values need be set) when using this formulation.

Since the mapping from the vector of expected mortality rates, on the scale of our prior, to the coefficients, on the scale of estimation, is a many-to-few transformation, it may seem less than obvious how to accomplish this. We have however developed a relatively straightforward procedure to do this (that we describe in Chapter 4). The resulting prior turns out to require fewer adjustable parameters than partially pooling coefficients and every one of which can be set on the basis of known demographic information. The resulting analysis generates the metric space necessary even to compare regressions with entirely different covariate specifications, including different numbers of covariates, and different meanings of the covariates include in each cross-

section.

We also found that our method of putting priors on the expected outcome variable, rather than the coefficients, turned out to solve a different problem that affects many time series-cross-sectional data collection efforts. The issue in these efforts is that interesting variables are available in some countries or cross-sectional units but not others. The problem is that existing methods require the same variables to be available for all the units. This is easy to see when trying to pool coefficients, since omitting a variable from the time series in one cross-sectional unit will make all the coefficients take on a different meaning and so they become impossible to pool either partially or completely (at least without adding untenable assumptions). The result is that in order to use existing methods, scholars routinely make what would otherwise seem like bizarre data analysis decisions. The uncomfortable choice is normally one among:

1. omitting any variables not observed for all units, which risks attributing differences to biases from omitted variables;

2. excluding cross-sectional units for which some variables are not available, which risks selection bias; or

3. running each least squares analysis separately, equation-by-equation, which risks large inefficiencies.

Researchers have of course been infinitely creative in choosing ad hoc data analysis strategies to limit the effects of these problems, but the lack of a better method clearly hinders the research in numerous fields.

Our method of pooling on expected mortality avoids this choice by allowing researchers to estimate whatever regression in each cross-sectional unit they desire, and to borrow strength statistically from all similar or neighboring units even with different specifications (i.e., different covariates) by smoothing the expected value of the dependent variable instead of each of the coefficients. Borrowing strength statistically in this way greatly increases the power of the analyses compared to simple equation-by-equation regressions. Making choices about priors is also much simpler. And as a result, with these methods, scholars can collect and use whatever data are most appropriate in each country or cross-sectional unit, so long as they have a dependent variable with the same meaning across countries. The data to which our methods are most useful have many cross-sections and a relatively short time series in each.[4]

[4] A different approach to this problem might be to multiply impute entire variables when they are missing in a cross-sectional unit, but this would require further methodological work since methods to do this have heretofore only been developed for independent cross-sections such as survey data (Gelman, King, and Liu, 1999). Moreover, imputation would be inappropriate when scholars prefer to use different covariates in different cross-sections (i.e., even when data happen to be available). This approach would also not help when the variables have different meanings in each cross-section.

Another situation where smoothing the expected outcome variable can be useful is when the same explanatory variables are in the specification for each cross-sectional unit, but they are measured differently in each. For example, measures of gross national product, and other economic variables, are denominated in the currency of each country. The methods we provide to smooth the expected outcome variable enable scholars to use variables in whatever denominations are most meaningful in each country.

What follows in this book contains much technical material but, when viewed from this perspective, the result is simply a better way of running least squares (LS) regressions. The logic is along the lines of seemingly unrelated regressions — such that if you have several regressions to run that are related in some way, estimating them jointly results in more efficient estimation than separate equation-by-equation analyses. Of course, the result is very different since, for example, even identical explanatory variables produce more efficient results and the precise information assumed and modeled is very different. The technology to produce our estimates may seem complicated, but the logic and the end result are quite simple. Indeed, the end result is still a set of regression coefficients, predicted values, and any other quantities of interest that normally come from a linear regression. The only difference is that the new estimates can include considerably more information and will have generally superior statistical properties to the old ones. Formally, they have lower mean square error and can cope far better with measurement error, short time series, noisy data, and model dependence. In our application, this added efficiency also produces much better out-of-sample forecasts, and more accurate regression coefficient estimates.

1.4 Implications for the Bayesian Modeling Literature

Our work has an important implication for the Bayesian modeling literature, and applies to many hierarchical or multilevel models with clusters of exchangeable units and spatial models imposing smoothness across neighboring areas, so long as the model includes multiple covariates. The implication is that *many of the prior densities commonly put on coefficients to represent qualitative knowledge about clustering or smoothness are misspecified*. We summarize our argument here and elaborate it in subsequent chapters.

As described in the previous section, some coefficients on explanatory variables are causal effects about which we might have some real prior knowledge. However, most variables included in regression-type analyses are controls (i.e., "confounders," "antecedent covariates," or "pre-treatment variables"), and the coefficients on these controls are usually viewed as ancillary. The problem is that the claim that researchers have prior knowledge about coefficients that have never been observed directly and most consider a nusance is dubious. These coefficients may have even been estimated,

but they have never really been the subject of study, and so there exists little sound scientific basis for claiming that we possess any substantial degree of prior knowledge about them.

But, one may ask, don't we often have strong knowledge that neighboring cross-sections are similar? Of course, but the issue is what "similar" means in the usual qualitative judgment about prior knowledge, and how we should go about formalizing such similarities. If we are predicting mortality, we might imagine that New Hampshire and neighboring Vermont have similar cause-specific mortality rates, but that does not necessarily imply that the coefficients on the variables that predict mortality in these two American states are similar. In fact, if the covariates differ between the two states, then the *only* way mortality can be similar is if the coefficients are different. As such, imposing a prior that the coefficients are similar in this situation is the opposite of our qualitative knowledge.

Put differently, what many researchers in many applications appear to mean when they say that "neighboring states are similar" is that the dependent variable (or the expected value of the dependent variable) takes on similar values across these states — not that the coefficients are necessarily similar. Moreover, similarity in one does not necessarily imply similarity in the other. In this area, like so many in mathematics and statistics, formalizing qualitative intuitions and knowledge with some precision often leads to counterintuitive conclusions.

We develop tools that enable researchers to put a prior directly on the expected value of the dependent variable, allowing these to be smooth across neighboring areas or all shrunk together in the case of an exchangeable group within a cluster. The result is a different prior and thus model in most applications, even when we translate this prior into its implications for a prior on the coefficients. In our mathematical analyses and empirical experiments, these priors outperform priors put directly on the coefficients: they fit the data better; they forecast better; and they better reflect our qualitative knowledge. They also require fewer adjustable parameters and a natural space within which to make all relevant comparisons, even among coefficients and no matter the scale of the covariates.

1.5 Incorporating Area Studies in Cross-National Comparative Research

In a variety of disciplines, and often independent of the disciplines, area studies scholars have explored individual countries and regions as separate areas of study. These scholars have contributed a large fraction of the basic descriptive information we have about numerous countries, but the advantages of the incredible depth and detailed analyses they perform are counterbalanced by the absence of comparison with other areas. Those focusing on different countries work in parallel but without much interaction and without systematic comparison. In the middle of the last century,

coincident with the rise of behavioralism and quantification in the social sciences, some scholars began to analyze some of the same questions as area studies scholars by systematic quantitative country comparisons.

Although these trends also affected anthropology, sociology, public health, and other areas, we tell the story from the perspective of political science where the conflict is particularly pronounced. Political science is also among the most diverse of scholarly disciplines, and it includes scholars from all the other disciplines affected.

Political scientists began to be comparative on a global scale in the 1960s, vastly expanding the scope of their data collection efforts. Over the rest of the century, they traveled to every corner of the planet to observe, measure, and compare governments, political systems, economies, conflicts, and cultures. Venturing out by oneself to a foreign (which meant primarily non-American) land to collect data became a right of passage for graduate students in the subfield of political science called comparative politics. Other scholars built large cross-national data sets that spanned ever increasing sets of countries. Whereas "comparative study [was once] comparative in name only" (Macridis, 1955), the comparative politics subfield and political science more generally emerged in this period as a more modern, international discipline.

As this period also marked the height of the behavioral movement in the discipline, many researchers became enthusiastic quantifiers and sought to measure concepts across as many nations as possible. Political science made important strides during this period but, in its efforts to expand comparisons across diverse cultures, researchers also created a variety of strained quantifications and mis-measured concepts, which often led to implausible conclusions, or at least to conclusions without a demonstrable connection to empirical reality.

The reaction from traditional area studies scholars and others was fierce. The data gathered was widely recognized as sometimes reliable but rarely valid: "The question with our standard variables on literacy, urbanization, occupation, industrialization, and the like, is whether they really measure common underlying phenomena. It is pretty obvious that, across the world, they do not; and this quite aside from the reliability of the data gathering agencies" (Sartori, 1970: 1039). Sartori talked about "conceptual stretching" and the "traveling problem" (for a more recent statement, see Collier and Mahon, 1993). Research practices were characterized as "indiscriminate fishing expeditions for data" (LaPalombara, 1968). Generally, relations between the two camps resembled some of the wars we have studied more than a staid scholarly debate: "no branch of political science has been in more extreme ferment than comparative politics during the last fifteen years" (LaPalombara, 1968: 52).

In the last 3–4 decades, comparative politics researchers have improved their data collection techniques. Their procedures are more professional, more replicable, and better documented, and the results are often even permanently archived (in the Inter-University Consortium for Political and Social Research, which was formed during this period by political scientists). Political scientists have also developed better theories to help guide data collection efforts, and as a result of all this work the

concepts underlying our quantifications, and the measures themselves, have improved. Data sets have moved from small cross-sectional snapshots to large time series-cross-sectional collections. And methods for analyzing data like these have also become more sensitive and adapted to the problems at hand (Beck and Katz, 1995, 1996; Beck, Katz, and Tucker, 1998; Stimson, 1985; Western, 1998; Zorn, 2001).

As a result of these improvements, the respect for quantitative work among those who know the details of individual countries has improved (as has the use of standards of scientific inference in qualitative research, King, Keohane and Verba 1994, but we still have a long way to go). Indeed, the field has not resolved the key *comparative* problem of quantitative comparative politics. The problem remains a problem in part because it may be inherently unresolvable. Political scientists want broad cross-national and comparable knowledge, and simultaneously need detailed context-specific information. They want unified concepts that apply to places that are so different that the concepts very well may not be comparable. Is bartering for a goat in Ghanzi (a town in land-locked Botswana) and buying a GPS-equipped luxury yacht in Fort Lauderdale really comparable after translating Pula into U.S. dollars, adjusting for purchasing power, and dividing by the implied cost of the goat? And that's money. What about ideas without natural units and without international organizations devoted to making them comparable — concepts like support for the government, partisan identification, social capital, post-industrialism, political freedom, human security, and many of the other rich concepts that political scientists study?

Quantitative scholars are of course painfully aware of these problems even when not explicitly working to solve them. The pain surfaces most obviously during data collection where scholars are torn between the desire to have one large comparable data set which they can stack up to run regressions on — thus needing the same variables measured over every country and time period — and the desire to respond to the area studies critique by collecting contextually and culturally sensitive measures. The issue is that contextually sensitive measures almost by definition involve collecting different variables, or the same variables with different meanings, in each country. The methods that have existed through this entire period, however, required the identical variables with the same meanings in all countries.

Area studies scholars thus seem to have at least two remaining complaints about the quantitative cross-national comparative literature: Some oppose quantification in principle, and others do not like the particular types of cross-national quantifications and comparisons that have been conducted. In our view, a key methodological constraint — requiring scholars to have the same variables with the same meanings across countries — has helped to conflate these two distinct complaints in scholarly debates held over the years. The methods offered here relax this underlying methodological constraint and enable scholars to use different explanatory variables, or those with different meanings, in the time series regressions in different countries. In that way a time series-cross-sectional data analysis can still be done but with more meaningful

measures. We hope this development will eliminate some of the basis of the second area studies complaint, making the resulting analyses more appropriate to each area studied. As importantly, methods along these lines may encourage scholars to collect data in different ways. These methods will not alleviate metaphysical objections to quantification, but if our methods enable or encourage scholars to collect more contextually sensitive, country-specific data, perhaps some of the reasons for the first complaint will also be reduced.

Thus, we can express a formalization and simplification of the preferences of area studies scholars (i.e., those who would allow some quantification) by saying that they would want to run a separate time series regression within each country, using whatever explanatory variables are appropriate to use within that country and without the constraints imposed by having to build comparable measures across countries. Of course, they want more than this, but they surely want at least this. Probably the leading area study, the field of American politics, has employed this strategy increasingly, and with increasing success, over the last several decades. The problem is that in most other countries, long time series are not available and so the results of any such analyses would be highly inefficient, making this strategy mostly infeasible. Most quantitative cross-national comparativists would probably also like to adopt this strategy, if it were feasible, since it would enable them to incorporate the detailed information and insights of the area studies scholars. Our methods enable quantitative area studies scholars to collect whatever explanatory variables they feel is appropriate and nevertheless to enable experts from different regions to work together — by borrowing strength across the different regions to improve estimation for each one individually — without any constraint other than a dependent variable with a common meaning. This hardly solves all problems of cross-national comparison, but it should make it possible to work together more productively.

1.6 Concluding Remark

This book can be viewed as a set of methods for forecasting mortality, a new approach to statistical modeling, a critique and method for improving an aspect of the Bayesian modeling literature, or a step toward resolving some of the disputes between area studies experts and cross-national comparativists. The methods introduced should also have substantive applications well beyond these areas, some of which we discuss in the chapters to come.

Part I

Existing Methods for Forecasting Mortality

In Part I, we survey 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 many other areas. The chapters that follow provide an opportunity for us to identify the work we build on and present our perspective on this work — 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.

Chapter 2

Methods Using No 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. Exogenous shocks to mortality, such as from the discovery of new medical technologies, economic crises, education campaigns, public health innovations, or from comorbidity patterns are ignored, while predictable epidemiological cycles due to biological or behavioral responses reflected in past mortality are included (see Gutterman and Vanderhoof, 1998).

Many results about how these approaches work appear here, including proposed resolutions to several open disputes in the literature. We also identify several previously unrecognized but serious limitations 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 be 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, with scholars forming 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 data analyses begin, although the mathematical tools of the two disciplines often differ. Demographers often 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 results. 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. And more importantly from the perspective of this work, they miss the opportunity to include their deep knowlege 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 their quantitative methods in this chapter, we will be well positions 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 are aware 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 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 1.4 and 1.3), 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 U.S., 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, TB, 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 which 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 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.

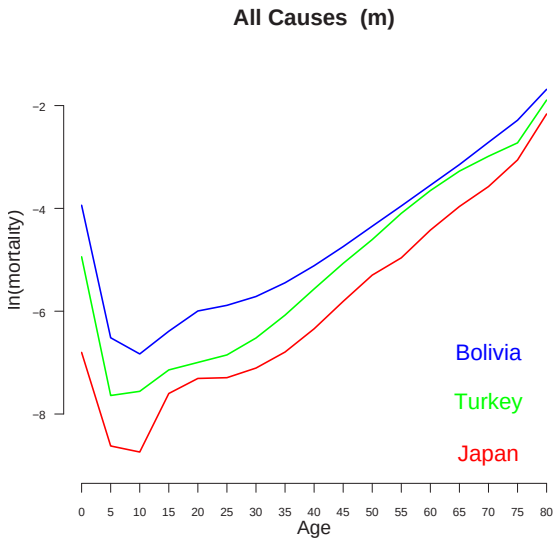

Figure 2.1: All-Cause Mortality Age Profiles: The age profile of log-mortality from all causes in males in the year 2000, for three countries. This shape is typical of that found in most countries. (The order of the countries listed matches the order of the lines at the right side of the graph.)

2.2 A Unified Statistical Framework

Somewhere there is or should be a theorem which proves that for every deterministic method of calculation there exists *some* formal statistical model for which the deterministic method is a reasonable estimator. Whether or not such a theorem would be accurate, we find that we 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 since 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 start building models from scratch. This would be a mistake

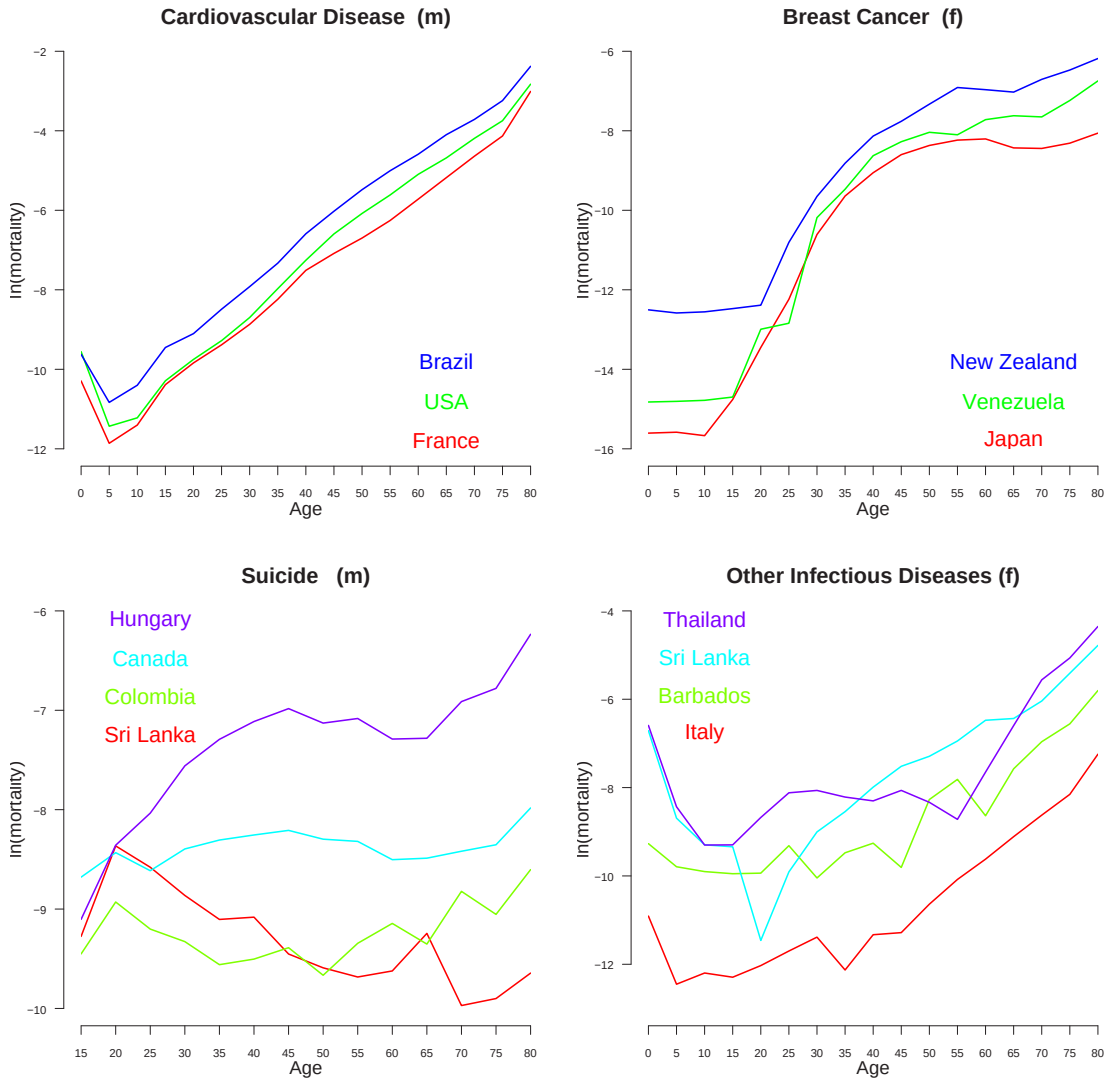

Figure 2.2: Cause Specific Mortality Age Profiles: The top two graphs (for cardiovascular disease in males and breast cancer in females) show similar age patterns of mortality among the countries displayed; the same pattern is common among other countries as well. The bottom two graphs (for suicide in 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. (In addition to the color codes, the order of the countries listed matches the order of the lines at the right side of the graph.)

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. While 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), possibly cause and sex specific, for a single country, with A rows corresponding to age groups and T columns corresponding to years. For example,

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

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 the mortality matrix above, 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 i.i.d. and normally distributed. The function f is assumed known, but its specific form varies greatly from one class of

models to another, and it may or may not reflect some a priori knowledge. 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 of parameters), 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 competing 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. Then these same mortality rates are assumed to hold constant over time, or they are assumed 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 U.S., Mexico, and Canada, and all three use similar extrapolative methods, with rates of mortality decline in various cross-sections a function of expert judgement. See Government's Actuary Department (2001) for discussions of 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 was 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 marshalling parametric forms for the log-mortality age profile for forecasting (see also Rogers, 1986, 1989; Rogers and Raymond, 1999, and McNown 1992). McNown and Rogers use the functional form due to Heligman and Pollard (1980):

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

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, since a variety of 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.

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 non-convex optimization problem. Unless great care is taken 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 property of being quite "inflexible": Constraining the behavior of a polynomial at 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 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 work well sometimes and poorly sometimes, 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 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 eight or more parameters to twenty data points can be tricky. With the gracious help of McKnown and Rogers, we were 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

[1] Parsimonious function representations which are smooth and preserve local properties are available and go under the generic name of splines. Splines are classes of piece-wise 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.

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, 2000, 2000a; and Lee and Tuljapurkar, 1994, 1998, 1998a). We discuss the Lee-Carter model in Section 2.6.

Since 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 248).

More intuitively, 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, since 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 $\gamma_1, \ldots, \gamma_T$ include an additional set of T unknowns 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 specification above. 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 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 hand," which can be thought of as a particular age profile parametrization. In order to obtain a model more in line with a non-parametric 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 Singular Values Decomposition (SVD) of the log-mortality matrix m, as we explain later.

Figure 2.3 illustrates in more detail the 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 1951 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 only showing the graph corresponding to year 2000. Since the age profiles in the first graph have fairly narrow variance, 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 very small variance and a zero mean.

Obviously it is not known a priori whether model 2.5 will represent the data 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:

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

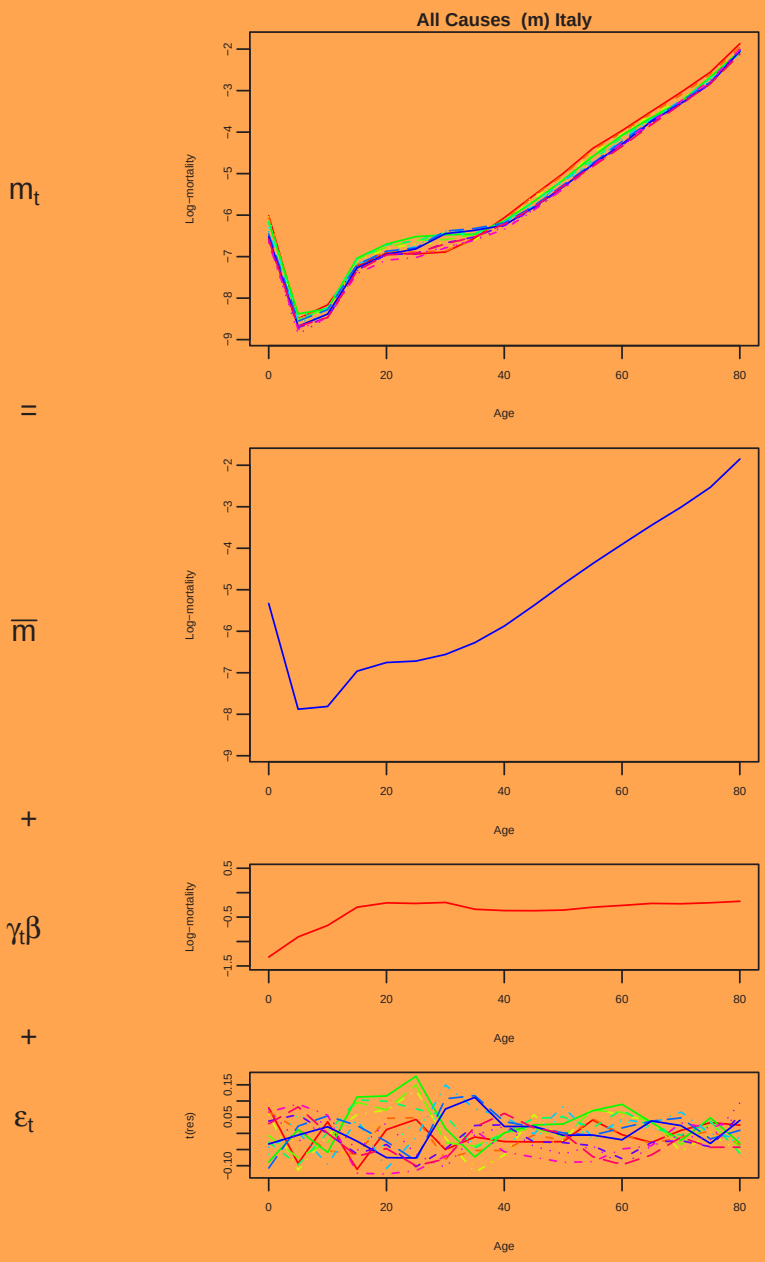

Figure 2.3: Decomposing Age Profiles with Principal Components: This figure parallels Equation 2.5, which decomposes the twenty male all-cause age profiles of log mortality in Italy (top graph) into the mean age profile (second graph), the first principal component (third graph), and the residuals (last graph).

Since 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 γ_{it} and β_i, $i = 1, \ldots, k$, correspond to specific rows and columns of the SVD of the log-mortality data matrix m. The estimated values of β_i, $i = 1, \ldots, k$, 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 of the particular specification they belong to. In other words, the maximum likelihood estimator 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 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 explain what "more important" means, suppose we 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, since 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 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, since each is a refinement over the previous one.

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

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 real data sets, relatively few principal components can 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/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 only have to forecast four 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

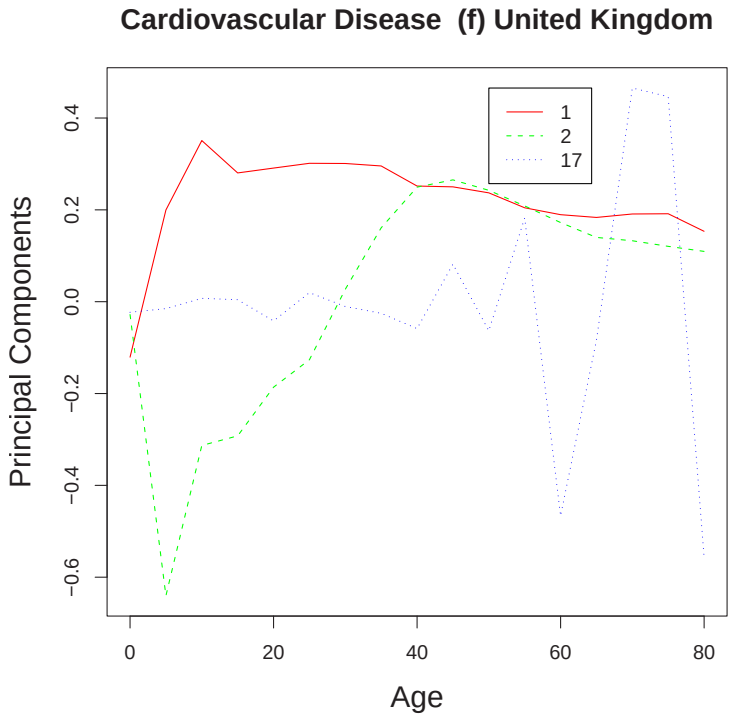

Figure 2.4: Principal Components of Log-Mortality: 1st, 2nd and 17th (there are 17 age groups). This figure refers to log-mortality from cardiovascular disease in females, United Kingdom.

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 is still often much better behaved than the time series of of the raw log-mortality data (Bell, 1988; 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, mouth and esophagus) in Japanese males.

In these data, the third and fourth time series hover around zero. Since the principal components are mutually orthogonal, we can interpret γ_{nt} as a measure of how much the 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 forth principal components do not play a crucial role in explaining the shapes of the age profiles in this case.

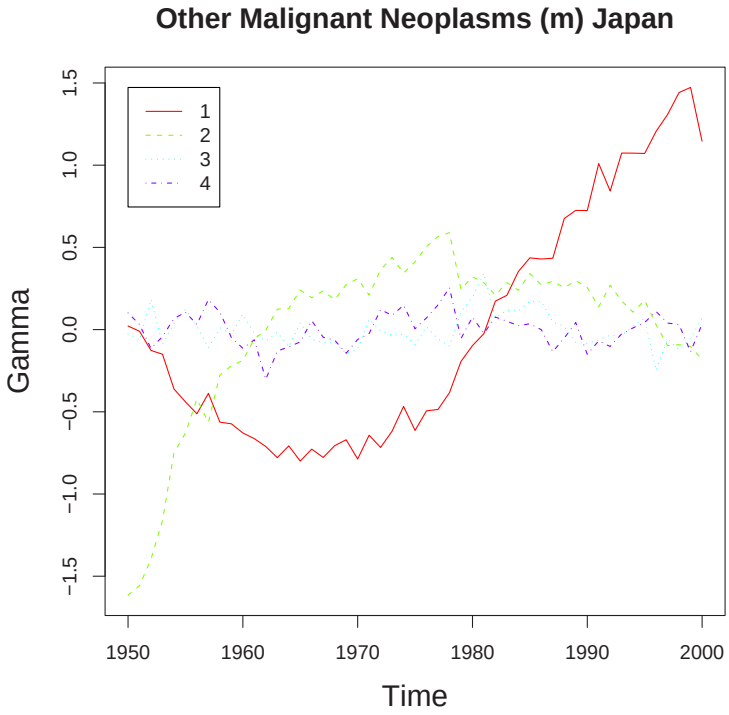

Figure 2.5: First 4 time series γ_{it} for other malignant neoplasms in Japanese males.

2.5.2 Estimation

We find it simplifying to think of PCA as a simple application of SVD (See Appendix B.2.4, Page 248). 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. Assuming 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 SVD decomposition above. Denoting by b_n the n-th column of B and by Q_t the t-th column of Q, the k-term approximation of the expression above can 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 exists 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 only need to compute 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, since $G = B'\tilde{m}$, 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, since 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 will switch freely between these two definitions, since they are equivalent and only differ in the particular procedure used to compute the result.

2.6 The Lee-Carter Approach

The principal components-based model due to Lee and Carter (1992*a*) is now used by the U.S. Census Bureau as a benchmark for their population forecasts, and its use has been recommended by the two most recent U.S. Social Security Technical Advisory Panels. Although Lee and Carter only intended for it to be used for all-cause mortality in the U.S. and a few other similarly developed countries, it is now one of the dominant methods in the academic literature and is 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 (2005).

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 29): It differs structurally from parametric models of the type in Equation 2.3 given that the dependence on age groups is non-parametric, and represented by the parameters β_a.

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

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

This is not a conceptual obstacle; it merely means that the likelihood associated with the model above 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 age profile in age group a: $\alpha_a = \bar{m}_a$. Since the Lee-Carter model implicitly assumes that the disturbances ϵ_{at} in the model above 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 matrix below) to approximate the $A \times T$ elements of the matrix:

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

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, 1989: 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×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: $\tilde{m} = BLU'$. (We assume that the singular values, the elements on the diagonal of L, are sorted in descending order and that the columns of B have length one.)

2. The estimate for β is the first column of B;

3. The estimate for γ_t is $\beta'\bar{m}_t$.

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 re-estimated 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*b*). In addition, different criteria have been

proposed more recently (Lee and Miller, 2001; Wilmoth, 1993), and some researchers skip this re-estimation stage altogether. These procedures, and problems with them, are described in Girosi and King (2005).

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 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 above 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

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

$$\tag{2.17}$$

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

$$\begin{aligned} \hat{\gamma}_t &= \hat{\gamma}_{t-1} + \theta + \xi_t \\ &= (\hat{\gamma}_{t-2} + \theta + \xi_{t-1}) + \theta + \xi_t \\ &= \hat{\gamma}_{t-2} + 2\theta + (\xi_{t-1} + \xi_t) \end{aligned} \tag{2.18}$$

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:

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

Since the random variables ξ_t are assumed to be independent with the same variance σ^2_{rw} the last term in the equation above 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 equation above more simply as:

$$\hat{\gamma}_{T+(\Delta t)} = \hat{\gamma}_T + (\Delta t)\hat{\theta} + \sqrt{(\Delta t)}\xi_t. \tag{2.19}$$

Since the variance of ξ_t can be estimated using Equation 2.16, we can use Equation 2.19 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", since $\hat{\gamma}_1, \ldots, \hat{\gamma}_T$ 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 2.19 is:

$$\text{Var}[\hat{\gamma}_{T+(\Delta t)} \mid \hat{\gamma}_1, \ldots, \hat{\gamma}_T] = (\Delta t)\sigma^2_{\text{rw}} \tag{2.20}$$

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 $\hat{\theta}$:

$$\text{E}[\hat{\gamma}_{T+(\Delta t)} \mid \hat{\gamma}_1, \ldots, \hat{\gamma}_T] = \hat{\gamma}_T + (\Delta t)\hat{\theta} \tag{2.21}$$

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

$$\begin{aligned} \mu_{T+(\Delta t)} &= \bar{m} + \hat{\beta}\hat{\gamma}_{T+(\Delta t)} \\ &= \bar{m} + \hat{\beta}(\hat{\gamma}_T + (\Delta t)\hat{\theta}). \end{aligned} \tag{2.22}$$

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

$$\begin{aligned} \hat{\mu}_{2030} &= \bar{m} + \hat{\beta} \times [\hat{\gamma}_{2000} + 30\hat{\theta}] \\ &= \bar{m} + \hat{\beta} \times \left[\hat{\gamma}_{2000} + 30\frac{(\hat{\gamma}_{2000} - \hat{\gamma}_{1950})}{50}\right]. \end{aligned} \tag{2.23}$$

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 (2005) 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 highly 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 below, empirical differences between the two estimators are often not major, but unless we are in the usual 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 U.S., 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 demographer's 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 quite general and does not depend on the particular time series extrapolation process used or any idiosyncracies 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*b*) 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 will eventually violate it.

To illustrate, Figures 2.6 and 2.7 offer examples of six datasets, one in each row. 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.

In other data sets, the forecast lines converge for a period, but after converging they cross and then from then on they too fan out — as in male suicide in the U.S. (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 twenty years of the series. Using such a method in this

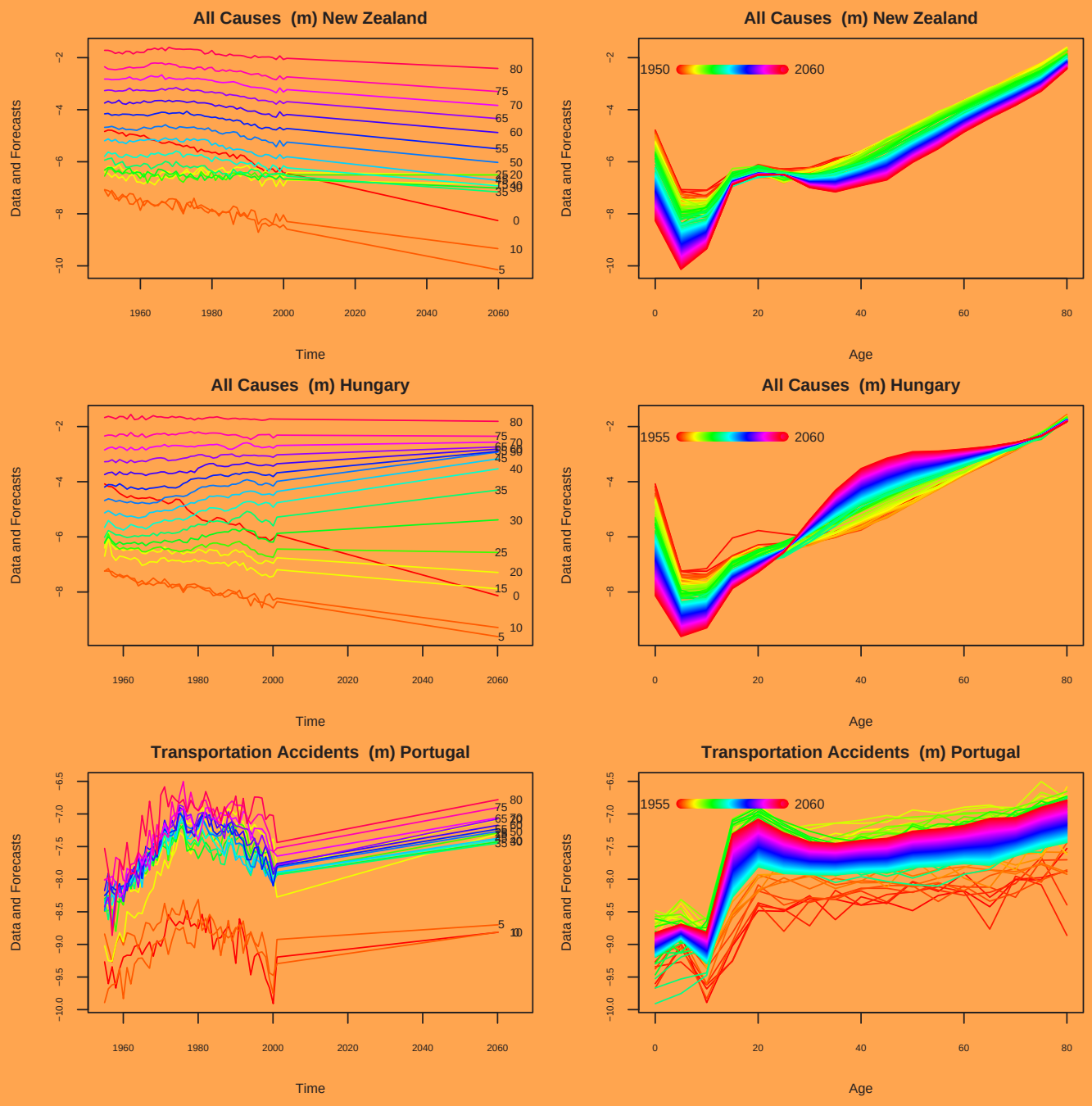

Figure 2.6: Data and Lee-Carter Forecasts by Age and Time, Part I

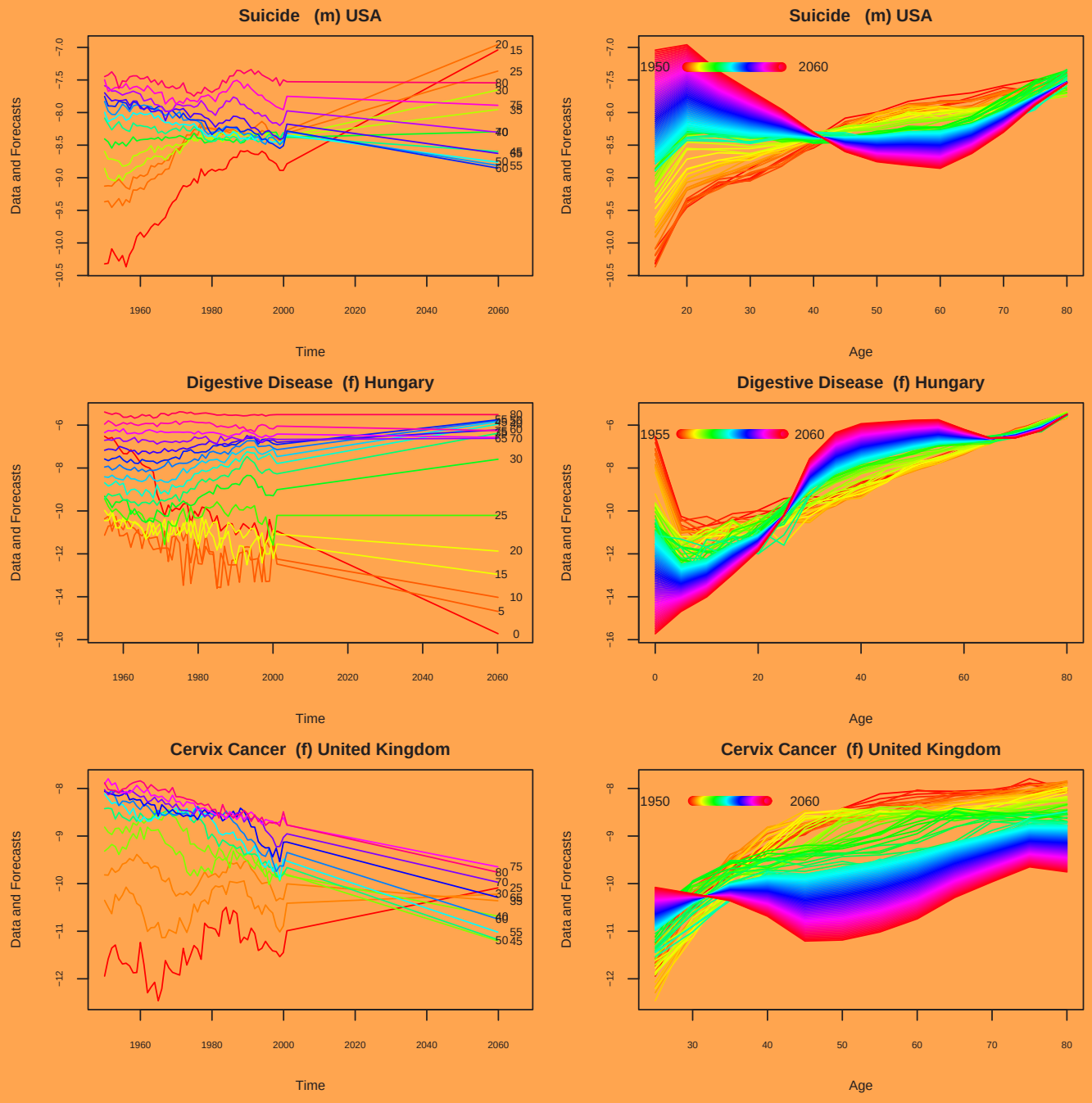

Figure 2.7: Data and Lee-Carter Forecasts by Age and Time, Part II

data set would also ignore a great deal of accumulated knowledge about predictable mortality transitions in countries developing economically. Of course, no demographer would forecast with a linear model like this in such dat, in part since 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 increase, 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 too should be taken into account in forecasting. Obviously, transportation accidents is a case where 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, since 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 produce 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 exmaple 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 Concluding Remarks

In this Chapter, we attempted to extract and illuminate the key insights of statistical methods of forecasting mortality that do not use covariates. Although we also highlighted problems with each method, the insights derived from each method and the collective wisdom of the literature in pursuing these methods is critically important. Indeed, we build our model in Part II precisely by incorporating and formalizing these insights.

For almost two centuries, the driving goal in the literature has been to reduce 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. The idea of dimension-reduction

comes from knowledge, based on considerable experience of scholars pouring 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 are 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 simliar lines, such as based on similarities across countries and time.

Chapter 3

Methods Using Covariates

We now introduce methods that use information from exogenously measured covariates to improve mortality forecasts. Information coded in covariates such as tobacco consumption, GDP, education rates, etc. are ignored in the methods of Chapter 2. To the extent that we understand mortality and can measure its causes, therefore, the methods described herein have the potential to improve forecasts substantially.

However, methods using covariates that have appeared in the literature often do not directly model the age-mortality relationship and exclude the considerable information we have about it. As such, these methods should not be considered generalizations of those in Chapter 2 and will not necessarily produce better forecasts. They will only improve forecasts when the additional information from the covariates outweighs the loss of information about the age profile of mortality. The methods we describe in Part II combine the insights from the approaches in both chapters.

3.1 Equation-by-Equation Maximum Likelihood

The idea of equation-by-equation maximum likelihood (ML) is to analyze each time series (i.e., for one age, sex, country, and cause) with a separate regression. Few scholars would forecast mortality thirty years ahead using a single maximum likelihood regression applied equation-by-equation to a short time series, but the method is the most obvious starting place, and a building block for most other methods described here. We explicate it here for expository purposes, to provide a baseline comparison with other approaches, to introduce some important issues in forecasting we use elsewhere, and as a starting point to introduce our notation. We consider two equation-by-equation approaches, based on the exponential-Poisson and linear-normal (or least squares) specifications, with the latter an approximation to the former.

3.1.1 Poisson Regression

We observe d_{it}, the number of people who die in year t in cross-section i. Because this variable is an event count, a reasonable starting point is to assume that d_{it} is a Poisson process, with unknown mean λ_{it}. (Obviously more sophisticated event count models may be chosen, such as based on the negative binomial or generalized event count distributions (Cameron and Trivedi, 1998; King, 1989*a*; King and Signorino, 1996), but we do not need them for the expository purpose of this chapter.) We summarize this information as follows:

$$d_{it} \sim \text{Poisson}(\lambda_{it}) \ , \quad \text{E}[d_{it}] = \lambda_{it}, \ , \quad \text{Var}[d_{it}] = \lambda_{it} \tag{3.1}$$

At the core of an equation-by-equation Poisson regression lies some specification for the expected value of the dependent variable as a function of some covariates. In our case, we are interested in the mortality rate $M_{it} \equiv d_{it}/p_{it}$, where p_{it} is the population of cross-section i at time t. To define this formally, denote by $\mathbf{Z}_{it}$ a $1 \times k_i$ row vector of covariates and by $\boldsymbol{\beta}_i$ a $k_i \times 1$ column vector of coefficients. (Throughout this book, we use **bold** to indicate vectors or matrices with at least one dimension equal to k_i and Greek to denote unknown parameters.) A common choice is the following log-linear specification:

$$\text{E}[M_{it}] \equiv \exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i). \tag{3.2}$$

Combining specification 3.2 with Equation 3.1 we summarize the equation-by-equation Poisson regression model as follows:

$$p_{it}M_{it} \sim \text{Poisson}\left(p_{it}\exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)\right) \tag{3.3}$$

The log-likelihood function for model 3.3 is then:

$$\ln \mathcal{P}(M \mid \boldsymbol{\beta}) \propto \sum_i \sum_t p_{it}\left[M_{it}\mathbf{Z}_{it}\boldsymbol{\beta}_i - \exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)\right]. \tag{3.4}$$

Here, as in the rest of the book, we have used the convention that when we drop an index of a variable we mean that we are taking the union over all the dropped indices.

The ML estimator of this model, which we denote by $\hat{\boldsymbol{\beta}}_i^{\text{PML}}$, satisfies the following first-order conditions:

$$0 = \sum_t p_{it}\mathbf{Z}_{it}\left[M_{it} - \exp\left(\mathbf{Z}_{it}\hat{\boldsymbol{\beta}}_i^{\text{PML}}\right)\right]. \tag{3.5}$$

3.1.2 Least Squares

Although the Poisson regression outlined in the previous section is appealing because it explicitly recognizes the event count nature of the data, it has a computational disadvantage in that the associated ML estimator requires optimizing non-linear equations for which no analytical solution exists.

It is possible to modify the approach of the previous section in such a way that the resulting ML estimator is a least-square estimator, and therefore can be computed as a solution of a linear system. The idea is to replace the Poisson distribution of Equation 3.1 with the log-normal distribution. These two distributions are obviously very different near the origin, but if the expected value of the number of deaths is not too small then they are fairly similar. The assumption of log-normality implies that the logarithm of the number of deaths is normally distributed. Hence, we define the log-mortality rate as

$$m_{it} = \ln M_{it} = \ln \frac{d_{it}}{p_{it}} \tag{3.6}$$

which would then be normally distributed. Since the logarithmic function has unbounded range, a linear specification for log-mortality can be used, suggesting the following model:

$$\begin{aligned} m_{it} &\sim \mathcal{N}\left(\mu_{it}, \sigma_i^2\right) \\ \mu_{it} &= \mathbf{Z}_{it}\boldsymbol{\beta}_i, \end{aligned} \tag{3.7}$$

where $\mu_{it} \equiv \mathrm{E}[m_{it}]$ is the expected log-mortality rate, and m_{it} is assumed independent over time after conditioning on $\mathbf{Z}$ (where $\mathbf{Z}$ may include lags of m_{it}). This model is also more flexible than Poisson regression, since the variance is not restricted to equal the mean.

Equation 3.7 posits a linear specification for the expected value of log-mortality, $\mathrm{E}[\ln M_{it}] \equiv \mathrm{E}[m_{it}] \equiv \mu_{it}$, whilst the previous section posits a linear specification for the log of the expected value of mortality, $\ln \mathrm{E}[M_{it}] \equiv \lambda_{it}$. The two models are obviously close, but not identical. One key difference between the models is that, in Equation 3.7, log-mortality is not defined when the realized value of mortality is zero. This need not be a problem when the expected value of the number of deaths is large enough, since in this case we will never observe zero deaths in a given year. However, when the expected value of deaths is small (say less than 10), which can happen because the cause of death is rare or the population group is small, some adjustment must be made.

A common and easy solution is to assign to each cross-section an extra 0.5 deaths every year (Plackett, 1981, p.5) before taking the log. The constant added should not be arbitrary, since one can generate almost any coefficient values by tinkering with this constant (King, 1988). Plackett's justification for 0.5 comes from his approximation of the log of expected mortality by expectation of the log of the number of deaths plus a constant. He shows that the approximation is optimized when the constant is 0.5. This issue is strictly related to the fact that the log-normal approximation to the Poisson distribution is not appropriate when the expected value of deaths is small, and so a somewhat better approach, short of moving to a full event count model, would be to replace the zeros with a number obtained by some imputation technique. We analyze this issue formally in Section 6.5 (Page 130). For simplicity, we assume

in this chapter that the data happen to contain no observed zeros, or that adding 0.5 deaths is a satisfactory enough solution.

Under these assumptions the likelihood function for the model in Equation 3.7 is simply:

$$\mathcal{P}(m \mid \boldsymbol{\beta}_i, \sigma_i^2) \propto \prod_i \sigma_i^{-T} \exp\left(-\frac{1}{2\sigma_i^2}\sum_t (m_{it} - \mathbf{Z}_{it}\boldsymbol{\beta}_i)^2\right) \tag{3.8}$$

A more general formulation can be obtained by allowing each observation m_{it} in the equation above to be weighted with some exogenous weight b_{it}. This is formally done by replacing the specification in Equation 3.7 with

$$\begin{aligned} m_{it} &\sim \mathcal{N}\left(\mu_{it}, \frac{\sigma_i^2}{b_{it}}\right) \\ \mu_{it} &= \mathbf{Z}_{it}\boldsymbol{\beta}_i. \end{aligned} \tag{3.9}$$

The weight b_{it} may reflect prior knowledge, such as knowing that an observation in a particular year is very noisy, or knowing that the variance of log mortality is, under certain conditions, inversely proportional to the expected value of the number of deaths. We will explore some of these possibilities in 6.5. It should be noted the parameter σ_i in Equation 3.9 has a different meaning from the parameter σ_i in Equation 3.7.

If the approach of Equation 3.9 is taken, it is convenient to make a change of variable so that the weights b_{it} disappear from the likelihood. Defining $y_{it} \equiv \sqrt{b_{it}} m_{it}$ and $\mathbf{X}_{it} \equiv \sqrt{b_{it}}\mathbf{Z}_{it}$, we rewrite the likelihood as follows:

$$\mathcal{P}(y \mid \boldsymbol{\beta}_i, \sigma_i^2) \propto \prod_i \sigma_i^{-T} \exp\left(-\frac{1}{2\sigma_i^2}\sum_t (y_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_i)^2\right) \tag{3.10}$$

The maximum likelihood estimator is therefore the following computationally fast weighted least squares estimator:

$$\boldsymbol{\beta}_i^{\text{wls}} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i y_i. \tag{3.11}$$

3.1.3 Computing Forecasts

Forecasts can be computed from a given model in at least three ways. These apply to the equation-by-equation model as well as some other models (such as the pooled model described in Section 3.2). They include forecasted covarates, no covariates, or lagged covariates, which we now explain.

Forecasting Covariates The first possibility is to fit the model to the observed data, forecast the covariates using separate models for each, assume that the relationship between m and $\mathbf{Z}$ remains constant between the in and out of sample periods,

and use the forecasted values of $\mathbf{Z}$ to create fitted values for m. This approach works well if indeed the relationship remains constant into the future and if high quality covariate forecasts are available. In practice, the uncertainty in the covariate forecasts propagate into mortality forecasts, greatly increasing the forecast variance. In addition, a valid application of a two-stage procedure like this should ideally be accompanied by forecasts that simultaneously model the full set of covariates without zero correlation restrictions. This is especially important for computing appropriate uncertainty estimates, and also for sensitivity tests based on using ranges of plausible values for each covariate. Unfortunately, generally accepted multivariate forecasts like these are not available.

Autoregressive Models A second method by which one can use a model like this to forecast mortality is to drop all variables from $\mathbf{Z}$ except functions of lags of m and to use a standard multi-step-ahead forecasting method. For example, if $\mathbf{Z}$ only includes a lag of mortality, we forecast for year t by computing $\hat{\mu}_t = \mathbf{Z}_{i,t-1}\boldsymbol{\beta}_i$. Subsequent steps involve using $\hat{\mu}_t$ in place of lagged mortality. This procedure can obviously be generalized to include full autoregressive integrated moving average (ARIMA) models by changing the functional form in Equation 3.7. This procedure works if the model closely approximates the data generation process, but it is heavily model-dependent and so minor deviations from the model get amplified through each step, sometimes driving forecasts far off the mark when many steps are required. The uncertainty in multi-step forecasts amplify especially quickly, and so ARIMA models tend to be most used only for much longer observed time series than are available for this problem. Of course, a potentially more serious problem with this procedure is that information from covariates is ignored. In experiments with our data, we found that even when a time sereis whas highly autocorrelated, an autoregressive model produced inferior out-of-sample forecasts due to high levels of model dependence.

Lagged Covariates A final procedure forecasts t periods ahead by using the one-step-ahead forecasting algorithm applied to data where the covariates are lagged by t periods. The disadvantage of this procedure is that either data are lost at the start of the period (where the values of $\mathbf{Z}_{it}$ for $t < 1$ are required) or earlier values of the covariates must be collected or "backcast." Backcast covariates can be created via a combination of statistical methods and expert judgment and are uncertain but will normally be less uncertain than covariate forecasts. The key disadvantage of this method is that to forecast t periods ahead it posits a relationship between two sets of variables lagged t periods, which may not be reasonable for some applications.

We experimented with all three methods, but found that covariates were often informative and that expert backcasts seemed far more reliable than forecasts. Smoking rates, to take one example, can in principle go in any direction henceforth, but prior to 1950 we have considerable evidence that people smoked less, which at a minimum bounds the estimates between zero and the 1950 value. In practice, even more

information is available, all of which should be used. Although any of the three methods can be used, we routinely find that lagging covariates is the most scientific approach. Unlike autoregressive approaches, the information in the covariates are used. Moreover, although we had hoped to gather the best covariate forecasts from the appropriate experts in fields corresponding to each covariate, we found in practice no less uncertainty in forecasted covariates than in forecasts of mortality.

3.1.4 Summary Evaluation

We note that equation-by-equation weighted least squares forecasts well in some instances, such as with high quality data for causes with many deaths (e.g., U.S. forecasts of cardiovascular disease among older men). But for most countries, which have shorter time series, or for younger age groups or other diseases, this method forecasts quite poorly. The time series on which it is based is normally quite short, and the variance estimates are large. We find that this approach typically overfits the data — resulting in a model that matches the in sample data well, but forecasts out-of-sample poorly. The result is a lot of noise: averaged over cross-sectional units this method may even be nearly unbiased, but the variance is so large for any one as to make it unusable in most circumstances. This method also treats all cross-sectional units as if they are independent, whereas the data and observable prior knowledge strongly suggest otherwise.

This method falls at the low bias, high variance end of the continuum representing bias-variance tradeoffs. Almost all other methods combine different cross-sections in attempts to trade some bias for larger reductions in variance.

3.2 Time-Series-Cross-Sectional Pooling

Whereas equation-by-equation weighted least squares is a low bias, high variance approach, Murray and Lopez's (1996) pooling approach is just the opposite, producing low variance but potentially high bias estimates. The advantages of one are thus the weaknesses of the other. Neither is optimal, but understanding the contributions of each is very helpful in building better models.

3.2.1 The Model

Whereas the equation-by-equation approach pools all time periods for a single age, sex, cause, and country, Murray-Lopez pools all time periods and countries for a given age, sex, and cause. They use a linear-normal model and hence

$$\begin{aligned} m_{ict} &\sim \mathcal{N}\left(\mu_{ict}, \sigma_i^2\right) \\ \mu_{ict} &= \mathbf{Z}_{ict}\boldsymbol{\beta}_i, \end{aligned} \tag{3.12}$$

where quantities are indexed by country c ($c = 1, \ldots, C$), time period t ($t = 1, \ldots, T$), and the fixed cross-sectional unit i ($i = 1, \ldots, n$) that now stands for age, sex, and cause. Thus, the variables in any one regression vary over c and t, and the coefficients are allowed to vary only over i (age, sex, and cause).

Because this model allows each coefficient to be based on many more observations than for equation-by-equation least squares ($C \times T$ at most compared to T at most, respectively), the variance of the quantities estimated is substantially smaller. However, the model makes the implausible assumption that $\boldsymbol{\beta}_i$ is the same for every country. Murray and Lopez did not, and clearly would never, pool age groups, since from their public health perspective we know very well that different age groups die of different causes and at different rates. Like scholars tend to do in many fields, they estimated separate coefficients for cross-sections they understood within the framework of their discipline but pooled over dimensions they were less focused on. From a political science perspective, however, the constancy assumptions are not plausible: the model assumes the direct effect of an extra year of education is same in the U.S. as it is in Tajikistan, and the effect of a given increase in GDP is the same in Benin as it is in Germany. Although we do not have the same strong prior information about the exact levels of these coefficients across countries as we have across age groups, especially since they are not necessarily causal effects, we have no good reason to think that they are the same across all national boarders.

What is the consequence of pooling cross-sections with coefficients that are not the same? At a minimum, parameters that vary more than indicated in the model lead to standard errors that are too small, but if the variation in the parameters is related to the variables in the equation, then a direct application of the model in Equation 3.12 will lead to bias. Indeed, every empirical test we run confirms that the coefficients vary considerably across countries: Pooling often induces high levels of bias.

3.2.2 Post-Estimation Intercept Correction

Murray and Lopez (1996) were of course aware of the problem of pooling coefficients that were not completely the same. The technology to partially pool in an appropriate fashion did not exist, and so they needed to pool something. Pooling countries was certainly the most obvious choice, and probably the best given the constraints available at the time. But since they were aware of the problems they found a (somewhat unconventional) approach to the problem that partially corrects for the difficulties while still allowing for pooling.

Instead of adding fixed effects for countries, which after all would only address unmodeled parameter variation in the intercept, they use a technique known as *intercept correction* (Clements and Hendry, 1998), which does this and a bit more. The idea is to fit the model in Equation 3.12, compute a predicted value to create a "forecast" for the last in-sample observation, calculate the error in fitting this last data point,

and assume the same method will generate the same error for out-of-sample forecasts. Since they use covariate forecasting to construct their forecasts, this method amounts to computing a k-year ahead forecast as the usual regression predicted value minus the last observed residual:

$$\mathbf{Z}_{ic,T+k}\hat{\boldsymbol{\beta}}_i - e_{icT}$$

where $e_{icT} = (m_{icT} - \mathbf{Z}_{icT}\hat{\boldsymbol{\beta}}_i)$ is the residual in country c and the last observed year T. This procedure may seem flawed, but since, under the model, $\mathrm{E}[e_{icT}] = 0$, subtracting e_{icT} from the usual model forecasts does not change the expected value of the forecast and so introduces no bias. However, intercept correction increases the variance of the model-based forecast (by $\mathrm{V}[e_T] = \sigma_i^2$), and so if the model is correct the procedure will be suboptimal. This makes sense of course, since if the series is noisy, the forecast depends on *one* data point that could be far from any underlying systematic pattern.

On the other hand, if the model is wrong due to a structural shift in the underlying parameters near the end of the in-sample period, or large differences across countries, then intercept correction can bring the forecasts back in line by reducing bias. This method cannot correct for bias due to variation in the slope parameters, but it adds some robustness to model violation in short term forecasts (and in long term forecasts biased by a constant shift factor) at the cost of some efficiency. Moreover, the higher variance is not a terrible cost under this approach given how many more observations are available in this pooled method to reduce the variance.

We find that without intercept correction, out-of-sample forecast errors from the Murray-Lopez model are immense: the coefficients are noisly estimated, often near zero, and the levels are often far off. With intercept correction, the forecasts are vastly improved since at least the levels are closer to correct, and when mortality changes slowly over time near-zero coefficients on the covariates and a level correction is not a bad combination. However, the approach still leaves considerable room for improvement: the age profiles are not constrained to be reasonable, most dynamics are usually missed, and the forecasts are typically not close to the mark. Although intercept correction no doubt improves a pooling model with constant coefficients, the large changes it introduces make it hard to imagine a justification for a use of a model that was so biased in the first place. For small to moderate changes in the forecasts, intercept correction is sometimes a reasonable and practical procedure. But when changes due to intercept correction are large and frequent, as for log-mortality, researchers should probably pursue a different approach, even when other features of the forecasts are satisfactory.

3.3 Partially Pooling Cross-Sections via Disturbance Correlations

One (non-Bayesian) way to allow cross-sections to borrow strength partially, without pooling, is to postulate a correlation among the set of disturbances for the separate cross-sectional regressions — as in seemingly unrelated regression models (SURM) (Zellner, 1962). Since the correlation in these models is among scalars (the disturbances), SURM can be made to work even if different covariates are available for different cross-sections, which addresses one of the problems of pooled cross-sectional time-series.

However, SURM was not intended to, and does not, resolve other important difficulties. If the explanatory variables are the same for each cross-section, then SURM reduces to equation-by-equation least squares. In this situation, the technique offers no efficiency advantage. When explanatory variables are highly correlated, even if not the same, SURM gives results that are very similar to least squares. This suggests that a stronger notion of smoothing is required for our problem, and indeed for most applied time-series-cross-sectional problems.

Finally, to make SURM work, one needs knowledge of the disturbances and their correlations with the covariates, information which, if available, should of course be used. However, in dealing with complex multivariate problems with numerous cross-sections it is not clear how confident we can be about knowledge regarding the behavior of these unobservable quantities. For SURM to work well, we must be in the odd situation where the things we *know* about the cross-sections (the covariates) are minimally correlated, but at the same time the things we *do not know* (the disturbances) are maximally correlated. This situation may occur sometimes, such as when different uncorrelated covariates are available for different cross-sections, and when powerful covariates common to all the cross-sections can be clearly identified but not measured (a situation that may also lead to omitted variable bias). However, the SURM framework clearly does not provide the kind of assumptions researchers would want to rely on to build a general method of borrowing strength between related linear regressions.

More generally, using SURM just because it is a convenient mathematical framework with which to introduce correlations among the cross-sections does not seem wise without a good substantive reason. In practice, empirical results using SURM often do not differ much from least squares. Instead, our guiding principle is that we should use the prior knowledge we truly have, rather than knowledge it would be convenient to have but we do not possess. In the next Part, we follow this principle and show how certain types of prior knowledge, that we do have in our application, and many others are likely to have in their's, are easily incorporated in the analysis of cross-sectional time series models using a Bayesian framework.

3.4 Cause-Specific Methods with Micro-level Information

We now discuss several approaches to forecasting mortality that include external sources of information in different ways. All share the feature of decomposing mortality into separate components, each of which is forecast separately to arrive at a mortality forecast.

3.4.1 Direct Decomposition Methods

We illustrate direct decomposition methods and their assumptions by summarizing the PIAMOD (Prevalence, Incidence, Analysis MODel) approach of forecasting cancer mortality rates (Verdecchia, Angelis and Capocaccia, 2002). This method uses cancer registry data that summarize the histories of individual patients to forecast cancer population mortality rates. Cancer registries are available in subsets of highly developed countries, but not in most other parts of the world, and so the technique is not as widely applicable as others.

Modeling

PIAMOD mortality forecasts are based on a deterministic relationship that decomposes mortality at one time into the prevalence and incidence of, and relative survival from, cancer, along with the death rate. Denote the fraction of people who die in age group a from cancer as M_a, and decompose it as

$$M_a = \sum_{a'=0}^{a} (1 - v_{a'}) \pi_{a'} \tau_{aa'} \theta_{aa'} \tag{3.13}$$

where the sum is over all cohorts from birth up to age a, and where $(1 - v_{a'})$ is the fraction of healthy individuals at time a' (one minus the prevalence), $\pi_{a'}$ is the incidence rate (the probability of contracting cancer between ages a' and $a' + 1$), $\tau_{aa'}$ is the relative survival probability at age a for a diagnosis of cancer at age a', and $\theta_{aa'}$ is the crude death rate from cancer from age a' to $a' + 1$ among those diagnosed with the disease.

Then each of the component parts are themselves forecast, and plugged into Equation 3.13 to produce a forecast of M_a. The healthy population is estimated by using Equation 3.13, setting $\theta_{aa'} = 1$, and summed up to year $a - 1$. Incidence and relative survival are then estimated to give an estimate of the healthy population as well as for use directly in Equation 3.13.

Incidence $\pi_{a'}$ is estimated as a logistic or exponential regression, as polynomial functions of age and cohort, respectively:

$$g(\pi_{at}) = \alpha + \sum_{j=1}^{j_a} \eta_j a^j + \sum_{j=1}^{j_c} \omega_j (t-a)^j \tag{3.14}$$

where $g(\cdot)$ is the log or logit link, α, $\eta_1, \ldots, \eta_{j_c}$, and $\omega_1, \ldots, \omega_{j_c}$, and are estimable parameters. Verdecchia, Angelis and Capocaccia (2002) decide on the number of polynomial terms, j_a and j_c, via a practically reasonable but theoretically questionable stepwise procedure that involves a function of likelihood ratios. They report an application where the polynomial for age is of degree eight and for cohort is degree two, which seems to them to be both reasonable and consistent with previous applications. To forecast incidence, the age and cohort effects are assumed to be constant into the future. Sometimes the cohort linear term (which of course is indistinguishable from a period effect) is let to drift (and in which case the logit instead of log link is used to avoid exponential growth).

The relative survival probability, $\tau_{aa'}$, is estimated by a mixture model for age-sex-period stratum i for the time since diagnosis d

$$\tau_a(d) = \alpha_i + (1-\alpha_i)\exp\left(-(\lambda_i d)^{\gamma_i}\right) \tag{3.15}$$

with the weight α_i modeled as a logistic function of time in order to help control for right censoring:

$$\alpha_i(t) = \frac{1}{1+\pi_0 \exp(\pi t)}. \tag{3.16}$$

Survival is forecast by merely choosing different scenarios and computing the mortality forecast for each. The pessimistic scenario of Verdecchia, Angelis and Capocaccia (2002) assumes that survival improvements do not occur. Their optimistic scenario assumes survival rates continue to improve exactly as in previous years. They do not consider a scenario where improvements would occur at an increasing rate.

Finally, PIAMOD models the crude death rate due to cancer, $\theta_{aa'}$, as

$$\theta_{aa'} = \left(1 - \frac{\tau_{aa'}}{\tau_{a,a'+1}}\right)(1-q_a^*)$$

where q_a^* is the probability of death from competing causes (i.e., other than cancer) at age a for a birth cohort surviving to age a. Since q_a^* is normally unobserved, the population crude death rate q_a is often substituted instead.

3.4.2 Microsimulation Methods

Another approach to forecasting based on external, exogenous sources of information is microsimulation. The idea here is to develop a computational (as distinct from

formal or statistical) model of the life course of people with or without particular illnesses, and to cull information from the academic literature to set parameter values.

The canonical microsimulation approach to disease-specific mortality forecasting is Weinstein et al.'s (1987; see also Salomon et al., 2002) model of death from (and incidence and cost of) coronary heart disease. The idea is to set up a flow chart of life paths, including onset of coronary heart disease, paths through treatment, recurrence, preventitive and therapeutic interventions, and death outcomes. The boxes with these (and other) states, and arrows that represent transitions among them, are articulated as much detail as information exists. The key fact about this approach is that no estimation is carried out by the authors. Instead, parameter values, such as the transition probabilities, are gleaned from prior academic studies.

Although Weinstein's specific model has been used for forecasting and compared to actual out-of-sample mortality data, this is not the typical approach take in this literature. The vast majority of these models are capable of producing forecasts, conditional on the assumptions, but they are not really empirical forecasting models and have not been validated on out-of-sample data. Instead, they are highly useful and fairly systematic reviews of the literature, and demonstrations of what the literature, taken as a whole, says about a particular cause of death.

3.4.3 Interpretation

The PIAMOD and microsimulation approaches include many reasonable modeling decisions. Each equation in PIAMOD or box and arrow in the microsimulation approach may lead to insights about a component of mortality. The approaches include additional information, and they use data closer to individual human beings, about which we have more real biological knowledge than we do about population aggregates.

Counterbalancing these advantages, of course, are the detailed data requirements. In part because of these substantial data demands, the approaches have not been widely used in practical forecasting applications. In addition, the detailed individual-level modeling means that more modeling decisions need to be made, and one winds up with a fairly large number of decisions to be made about which we may have little prior knowledge. This can be seen in PIAMOD, from the high dimensional polynomial modeling of incidence, to the mixture model for relative survival, to the assumptions of constant age profiles over time, etc. Mortality forecasts are necessarily highly dependent on many of these modeling decisions.

Similarly, many ways of laying out the boxes and arrows of the microsimulation approach can be seen as equally plausible, and there exist few ways of validating each part of the model. Although this approach remains an extremely creative way to summarize vast tracks of literature, it is difficult to use to produce reliable forecasts for many causes of death.

3.5 Concluding Remarks

The methods of forecasting mortality described in this chapter fall on a dimension from low bias, high variance, for equation-by-equation analyses, to higher bias and lower variance, such as the Murray-Lopez forecasts. Both of these methods include covariates, which are intended to code some of what we know about the patterns of mortality from numerous biological and aggregate empirical studies. Parametric curve-fitting also falls at the higher bias, lower variance end of the continuum, but it excludes knowledge we may have in the form of measured covariates (although it would not be difficult to add covariates to the forecasts of the parameters in Equation 2.3).

Better methods of forecasting will generally come from more information, and so the methods we develop below will all allow the use of covariates, when available. However, the key intuition provided by each method will also be retained. That is, we begin with equation-by-equation analyses and like pooling approaches we give up some bias for larger reductions in variance. Like all three approaches, we also seek to improve forecasts by finding and incorporating additional information wherever possible.

Part II
Statistical Modeling

In Part II, we introduce a class of statistical models that generalize linear regression for time-series-cross-sectional analyses. We also provide new methods for identifying, formalizing, and incorporating prior information in these and other models. Chapter 4 introduces our model and a new framework for generating priors. Chapter 5 extends the framework to grouped continuous variables, like age groups. Chapter 6 explains how to connect all modeling choices to known substantive information. Then Chapter 7 implements our framework for a variety of other variables like those which vary over geographic space, and various types of interactions. Chapter 8 then provides more detailed comparisons between our model and spatial models for priors on coefficients and then extends our key results and conclusions to Bayesian hierarchical modeling.

Chapter 4

The Model

4.1 Overview

The models introduced in this chapter all depend on the specification of priors, which we introduce here and then detail in the rest of Part II. Details of how one can actually compute estimates using this model appear in Part III. Our strategy is to begin with the models in Chapter 3 that use covariates and add in information about the age-mortality profile in a different way than in the models in Chapter 2. After putting all the information from both approaches in a single model, we then add other information not in either, such as the similarity of results from neighboring countries or time periods.

In this Section, we use a general Bayesian hierarchical modeling approach to information pooling.[1] Although developing models within the Bayesian theory of inference is entirely natural from the persepctive of many fields, in several respects it is a departure for demography. It contrasts most strikingly with the string of scholarship extending over most of the last two centuries that seeks to find a low-dimensional parametric form for the mortality age profile (see Section 2.4 and the remarkable list in Tabeau (2001)). It has more in common with principle components approaches like Lee-Carter, in that we also do not attempt to parameterize the age profile with a fixed functional form, but our approach is more flexible and capable of modeling patterns in and out of sample known from prior research in demography or any other chosen by the researcher. Our approach also contrasts with the tendency of demographers to use their detailed knowledge only as an ex post check on their results. We instead try to incorporate as much of this information as possible into the model. Our methods tend to work better only when we incorpoarate information demographers have about observed data or future patterns.

The opposite of course applies too: Researchers forecasting variables for which no

[1] For other approaches to Bayesian hierarchical modeling, and for related ideas, see Blattberg and George (1991), Gelman et al. (1995), Gill (2002), and Western (1998).

prior quantitative or qualitative analyses or knowledge exists will not benefit from the use of our methods. And of course those who use incorrect information may of course degrade their forecasts by adding priors.

Although our approach will work with any relevant probability density for log-mortality, including those based on event count models discussed in Section 3.1.1, we fix ideas by developing our model by building on the equation-by-equation least squares (LS) model, described in Section 3.1.2. This model is

$$\begin{aligned} m_{it} &\sim \mathcal{N}\left(\mu_{it}, \frac{\sigma_i^2}{b_{it}}\right) \qquad i = 1, \ldots, N \ , \quad t = 1, \ldots T \\ \mu_{it} &= \mathbf{Z}_{it}\boldsymbol{\beta}_i, \end{aligned} \tag{4.1}$$

where as before m_{it} is the log-mortality rate (or a generic dependent variable) with mean μ_{it} and variance σ_i^2/b_{it}, b_{it} is some exogenous weight, and $\mathbf{Z}_{it}$ is a vector of exogenous covariates. We are not concerned with the choices of the weights b_{it} and the covariates $\mathbf{Z}_{it}$ here: We discuss these modeling choices in Chapter 6; although they are crucial, their specifics have no effect on the overall structure of our model.

The specification in Equation 4.1 forms the basic building block of our hierarchical Bayesian approach, and so we now interpret the coefficients $\boldsymbol{\beta}_i$ and the standard deviations σ_i as random variables, with their own prior distributions. We denote the prior for the variables σ_i generically as $\mathcal{P}(\sigma)$. The prior for the coefficients $\boldsymbol{\beta}$, which usually depends on one or more hyper-parameters θ, we denote by $\mathcal{P}(\boldsymbol{\beta} \mid \theta)$. The hyper-parameters θ also have a prior distribution $\mathcal{P}(\theta)$. (By our notation conventions, $\mathcal{P}(\theta)$ and $\mathcal{P}(\sigma)$ are different mathematical expressions; see Appendix A.)

We choose the specific functional form of the priors $\mathcal{P}(\sigma)$ and $\mathcal{P}(\theta)$ to make the computations simple (usually a Gamma or inverse-Gamma density), with the mean and variance set using genuine prior knowledge. However, our central arguments in this chapter, and most of the rest of this book, are about the specification for the prior for the coefficients, $\mathcal{P}(\boldsymbol{\beta} \mid \theta)$. This prior will be taken as highly informative, reflecting considerable prior knowledge. The issue at hand is deciding precisely how to formalize this prior knowledge in this density.

Using the likelihood function $\mathcal{P}(m \mid \boldsymbol{\beta}, \sigma)$ from Equation 3.8 (Page 54), and assuming that σ is a priori independent of $\boldsymbol{\beta}$ and θ, we express the posterior distribution of $\boldsymbol{\beta}$, σ and θ conditional on the data m as:

$$\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m) \propto \mathcal{P}(m \mid \boldsymbol{\beta}, \sigma) \left[\mathcal{P}(\boldsymbol{\beta} \mid \theta)\mathcal{P}(\theta)\mathcal{P}(\sigma)\right]. \tag{4.2}$$

where $\mathcal{P}(\boldsymbol{\beta}, \theta, \sigma) \equiv \mathcal{P}(\boldsymbol{\beta} \mid \theta)\mathcal{P}(\theta)\mathcal{P}(\sigma)$ is the prior. Once the prior densities have been specified, we usually summarize the posterior density of $\boldsymbol{\beta}$ with its mean,

$$\boldsymbol{\beta}^{\text{Bayes}} \equiv \int \boldsymbol{\beta}\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m)\, d\boldsymbol{\beta} d\theta d\sigma. \tag{4.3}$$

(or sometimes the mode) and can then easily compute forecasts using one of the three methods described in Section 3.1.3.

This section provides a framework for the information pooling problem: By choosing a suitable prior density for $\boldsymbol{\beta}$ we summarize and formalize prior qualitative knowledge about how the coefficients $\boldsymbol{\beta}_i$ are related to each other, so that information is shared among cross-sections. If the prior for $\boldsymbol{\beta}$ is specified appropriately, the information content of our estimates of $\boldsymbol{\beta}$ will increase considerably. This, in turn, can result in more informative and more highly accurate forecasts.

4.2 Priors on Coefficients

As we have described, a common way to derive a prior for $\boldsymbol{\beta}$ is to use the following kind of prior knowledge: "similar" cross-sections should have "similar" coefficients. The most common approach is to use a class of Markov random field priors, which are an example of an intrinsic autoregressive prior. These models are closely related to the autoregressive priors pioneered by Besag and his colleagues (Besag, 1974, 1975; Besag and Kooperberg, 1995) in that they allow spatial smoothing for units like age groups that vary over nongeographical space. The priors formalize this knowledge by introducing the following density:

$$\mathcal{P}(\boldsymbol{\beta} \mid \Phi) \propto \exp\left(-\frac{1}{2}H^{\beta}[\boldsymbol{\beta}, \Phi]\right), \tag{4.4}$$

where,

$$H^{\beta}[\boldsymbol{\beta}, \Phi] \equiv \frac{1}{2}\sum_{ij} s_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2_{\Phi} \tag{4.5}$$

where we use the notation $\|\mathbf{x}\|^2_{\Phi}$ to refer to the weighted Euclidean norm $\mathbf{x}'\Phi\mathbf{x}$ and where the symmetric matrix s is known as the *adjacency matrix*, and its elements can be thought of as the inverse of the "distance," or the proximity, between cross-section i and cross-section j.[2] It is useful, for future reference, to write Equation 4.5 in an alternative way:

$$H^{\beta}[\boldsymbol{\beta}, \Phi] = \sum_{ij} W_{ij}\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j \ . \tag{4.6}$$

where $W = s^{+} - s$ is a positive semi-definite symmetric matrix whose rows sum to 1 (see Appendix B.2.6, Page 253). The matrix Φ is a generic symmetric, positive definite matrix of parameters, which help summarize the distance between vectors of coefficients. Since it is usually unknown the matrix is considered to be a set of hyperparameters to be estimated, with its own prior distribution. In practice

[2] Although constraining the elements of this matrix to be positive is consistent with their interpretation as proximities, the constraint is not necessary mathematically. The only constraint on s is that the quadratic form defined by Equation 4.5 be positive definite.

this matrix is likely to be taken to be diagonal in order to limit the number of the unknowns in the model, although this implies the strong assumption that elements of the coefficient differences $(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$ are a priori independent of each other. This specification also implies that Φ is constant over i, which we show later is highly improbable in many applications.

The function $H^{\beta}[\boldsymbol{\beta}, \Phi]$ assumes large values when similar cross-sections (i.e., s_{ij} "large") have coefficients far apart (i.e,. $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}$ is also "large"). Therefore Equation 4.4 simply says that, a priori, the most likely configurations of coefficients $\boldsymbol{\beta}$ are those in which similar cross-sections have similar coefficients, or, in other words, those in which the coefficients $\boldsymbol{\beta}$ vary smoothly across the cross-sections.

A key point is that the prior defined by Equations 4.4 and 4.5 is *improper* (which means that the probability density in Equation 4.4 integrates to infinity; see Appendix C). The improperness stems from the fact that the function $H^{\beta}[\boldsymbol{\beta}, \Phi]$ is constant and equal to 0 whenever $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are equal, regardless of the levels at which the equality occurs (i.e., in the subspace $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \forall i, j = 1, \ldots, N$). This causes no statistical difficulties: Since the likelihood is proper (normal), the posterior is always proper. Indeed, an improper prior is a highly desirable feature in most applications, since it does not constrain the regression coefficients to be close to any particular value (which would normally be too hard to specify from prior knowledge), but rather only to be similar to each other. In other words, the prior density in Equation 4.4 is not sensitive to the absolute values or levels of the coefficients, only to their relative values.

Example 1 Suppose the cross-sections are labeled only by countries (with no age group subclassification). Then s could be a symmetric matrix of zeros and ones, where $s_{ij} = 1$ indicates that country i and country j are "neighbors", in the sense that we believe they should have similar regression coefficients. Neighbors could also be coded on the basis of physical contiguity, proximity of major population areas, or frequency of travel or trade between the countries. In practice, the matrix s would need to be constructed by hand by a group of experts on the basis of their expectations of which countries should have similar coefficients, which of course requires the experts to understand the meaning of all the regression coefficients and how they are supposed to vary across countries. ⊠

Example 2 Suppose cross-sections are labeled by age groups, or by a similar variable with a natural ordering (with no country-level subclassification). Then s could be a tridiagonal matrix of ones, so that every age group has as neighbors its two adjacent age groups. A more general choice is a band matrix, with the size of elements decaying as a function of the distance from the diagonal. Although the general pattern desired may be clear from prior knowledge, choosing the particular values of the elements of s in this situation would be difficult, since they do not directly relate to known facts or observed quantities. ⊠

4.3 Problems with Priors on Coefficients

In Bayesian modeling, we summarize, formalize, and incorporate nonsample knowledge in our inferences by specifying a prior density. This prior must be specified for all unknown parameters. Thus, for our problem, no matter what arguments one might make, we will ultimately need a prior for $\boldsymbol{\beta}$. The issue is how to turn qualitative and impressionistic knowledge into a specific mathematical form. But herein lies a well-known disconnect in Bayesian theory: Because the prior knowledge we have is typically in a very different form than the probability density we ultimately need, the task of choosing a density often requires as much artistic choice as scientific analysis. In many Bayesian models, this disconnect is spanned with a density that forms a reasonable approximation to prior knowledge.

Unfortunately, in the case of Bayesian models with covariates, the disconnect can be massive and, we demonstrate here, the resulting density chosen is often inappropriate. Our critique applies to many Bayesian spatial or hierarchical models that put a prior on a vector of coefficients and where the prior is intended to convey information. The problem here is that the jump from qualitative knowledge to prior density is too large, and some steps are skipped or intuited incorrectly. This argument applies to many models with spatial smoothing like that in Equation 4.4 and more generally to hierarchical models with clusters of exchangeable units that include covariates. We describe these problems here with spatial smoothing and make the extension to hierarchical models, featuring clusters of exchangeable units, in Section 8.2.

4.3.1 Little Direct Prior Knowledge Exists About Coefficients

To put a prior on the vector $\boldsymbol{\beta}$, we need to be in possession of nonsample knowledge about it. When an element of $\boldsymbol{\beta}$ coincides with a specific causal effect, the claim to nonsample knowledge is complicated, but sometimes plausible. For example, we know that twenty-five years of tobacco consumption causes a large increase in the probability of death (from lung cancer, heart disease, and other causes) in humans. However, $\boldsymbol{\beta}$ in our models are at the population level, and so they are not necessarily causal effects. For example, if we observe that tobacco consumption is positively related to lung cancer mortality across countries, it may be that smokers are getting lung cancer, but it could also be true — on the basis of the same patterns in the aggregate data — that it is the nonsmokers who happen to live in countries with high tobacco consumption who are dying at increasing rates. Whether we can imagine the reason for such a pattern is unimportant. It can occur and if it does the connection from the aggregate level relationship to the individual level at which the biological causal effect is known may be severed. This of course is an example of the well-known ecological inference problem (Goodman, 1953; King, 1997), the point being that without special models to deal with the problem $\boldsymbol{\beta}$ may not be a causal effect

even for a covariate as apparently obvious as tobacco consumption.

A better case for the claim of prior knowledge about $\boldsymbol{\beta}$ may be variables where the causal effect operates at the country level. For example, a democratic electoral system, or a comprehensive health care system, may lead to lower mortality from a variety of causes. Although these effects would also operate at the individual level, the causal effect could plausibly occur at the societal level. In that situation, no ecological inference problem exists and the case that we may really possess prior knowledge about at least this coefficient is more plausible.

However, even when one coefficient is truly causal, its interpretation is clear, and much prior knowledge exists about its likely direction and magnitude, it is typically not the only coefficient. Normally, we have a set of control variables with corresponding coefficients. The problem is that coefficients on control variables are treated as nuisance parameters, are typically not the subject of study, and are rarely of any direct interest. As such, even for regressions that include well-specified causal effects, we will likely have very little direct prior knowledge about most of the coefficients.

A final point is that $\boldsymbol{\beta}$ is obviously not scale invariant with respect to $\mathbf{Z}$: if we double $\mathbf{Z}$ we are also halving $\boldsymbol{\beta}$. This is not a problem if we truly understand the coefficients, since we would merely scale everything appropriately and set the prior to suit. However, when the coefficients' values are not fully understood, several problems can ensue. The main problem here is that the whole model requires that $\boldsymbol{\beta}$ take on the same meaning for all cross-sectional units. However, if the meaning or scale of $\mathbf{Z}$ changes at all, then our prior should change. Yet, the parameters Φ in Equation 4.4 have no subscript and are assumed constant over all units. In some situations, this is plausible but, even for variables like GDP, we expect some changes in scale over the units, even after attempts to convert currencies and costs of living.

This problem is sometimes addressed by standardizing $\mathbf{Z}$ in some way, such as by subtracting its sample mean and dividing by its sample standard deviation. This undoubtedly helps in some situations, but it just as certainly does not solve the problem. For a simple example, suppose one covariate is GDP per capita and another real disposable income per 100 population. Suppose that the right normalization here is to multiply GDP by 100 (even though this assumes away a host of other potential problems). Now suppose that for whatever reason GDP per capita varies very little over countries in some data set, but real disposable income varies enormously. In that situation, standardization would exacerbate the problem rather than solve it. The general problem here is that the sample does not necessarily contain sufficient information with which to normalize the covariates. Some exogenous information is typically needed.

4.3.2 Normalization Factors Cannot Be Estimated

Whether the coefficients are meaningful or not, the prior in Equation 4.4 contains the expression $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_\Phi$, which implies that the coefficients can all be made comparable.

In particular, it assumes that we can translate the coefficient on one variable in a single cross-sectional regression to the scale of a coefficient on another variable in that cross-section or some other cross-section. Indeed, this prior specifies a particular metric for translation, governed by the hyperprior parameter matrix Φ.

To be more specific, we denote individual explanatory variables by the index v, and rewrite Equation 4.5 (Page 69) as

$$\begin{aligned} H^{\beta}[\boldsymbol{\beta}, \Phi] &\equiv \frac{1}{2} \sum_{ij} s_{ij} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}^2 \\ &= \sum_{ijv} W_{ij} \boldsymbol{\beta}_i^v b_j^v \end{aligned} \tag{4.7}$$

where W is a function of s defined in Appendix B.2.6 (Page 253) and, most importantly,

$$b_j^v \equiv \sum_{v'} \Phi_{vv'} \boldsymbol{\beta}_j^{v'} \tag{4.8}$$

is the translation of coefficient v' in cross section j, into the same scale as that for coefficient v in cross-section i.

As is more obvious in this formulation, Φ serves the critical role of normalization constants, making it possible to translate from one scale to another. For example, if we multiply degrees Celsius by 9/5 and add 32, we get degrees Fahrenheit, where the numbers 9/5 and 32 are the normalization constants. The translations that Φ must be able to perform include normalizing to the same scale (1) coefficients from different covariates in the same cross-sectional regression, (2) coefficients from the same covariate in different cross-sectional regressions, and (3) coefficients from different covariates in different cross-sectional regressions. Each of these three cases must be made equivalent via normalization, and all this prior knowledge about normalization must be known ex ante and coded in Φ.

The role of the normalization can be seen even more clearly by simplifying the problem to one where Φ is diagonal. In this situation, the normalization factor is especially simple:

$$b_j^v = \Phi_{vv} \boldsymbol{\beta}_j^{v'}.$$

and so Φ_{vv} simply provides the weights to multiply into the coefficient vector in one cross-section to get the coefficient vector in another cross-section.

This alternative formulation is appropriate only if we have prior knowledge that different components of $\boldsymbol{\beta}$ are independent. In other words, although Equation 4.8 contains the correct normalization, regardless of independence assumptions, the prior in Equation 4.7 allows us to use only those parts of the normalization that are relevant to forming the posterior, and independence among components of $\boldsymbol{\beta}$ means that the cross-product terms (i.e., when $v \neq v'$) in Equation 4.8 would not be needed.

Although assuming that elements of a prior are independent is common in Bayesian modeling, the assumption of independence is far from innocuous here, since the result can greatly affect the comparability of coefficients from different cross-sections or variables, and thus can enormously influence the final result.

The key to putting priors on coefficients is knowing Φ. Without this knowledge, the translation from one scale to another will be wrong, the prior will not accurately convey knowledge, and our estimates and forecasts would suffer. Unfortunately, since we often know little about many of the $\boldsymbol{\beta}$ coefficients, researchers usually know even less about the values in Φ. Any attempt within the Bayesian theory of inference to bring the data to bear on the prior parameter values will fail, which is easy to see by trying to estimate Φ by maximizing the posterior: Since Φ does not appear in the likelihood, the entire likelihood becomes an arbitrary constant and can be dropped. As such, under Bayes, the data play no role in helping us learn about Φ; all information about it must come from prior knowledge, which of course is the problem.

Some scholars try to respond to the lack of knowledge of Φ as good Bayesians by adding an extra layer to the modeling hierarchy and putting a proper hyperprior on Φ. Ultimately, however, we always need to choose a mean for the distribution of Φ. And that deeply substantive choice will be critical. Adding variance around the mean does not help much in this situation since it merely records the degree to which the smoothing prior on $\boldsymbol{\beta}$ (and our knowledge of the normalization factor) is irrelevant in forming the model posterior: If a prior on the coefficients is to do any good, one must know the normalization factor, Φ, or choose a sufficiently narrow variance for the prior on it. Otherwise, no Bayesian shrinkage occurs and the original motivation for using the model vanishes.

The fact is that Φ is inestimable from the given data and must be imposed a priori with exogenous knowledge. Adding a prior so that Φ is identified does not help unless that prior is also meaningful since the estimates will be the results of prior specification rather than empirical information. If prior knowledge about the normalization factor does not exist, then the model cannot be meaningfully specified.

4.3.3 We Know about the Dependent Variable, not the Coefficients

When experts say that neighboring countries or a set of regressions are all "similar" they are not usually talking about the similarity of the coefficients. It is true that in Bayesian analysis, we need a prior on coefficients, and so it may seem reasonable to attach the qualitative notion of similarity to the formal Bayesian prior density in Equation 4.4. But reasonable it is not, at least not generally. In most situations, it seems that "similarity" refers to the dependent variable or the expected value of the dependent variable, not the coefficients, and assuming that similarity in the expected value of the dependent variable applies to similarity in the coefficients turns out to be a serious flaw.

Even if experts from public health and demography are willing to accept the linear functional form we typically specify, $\mu_{it} \equiv \mathbf{Z}_{it}\boldsymbol{\beta}_i$, they do not normally observe the coefficients, $\boldsymbol{\beta}$ or even any direct implications of them. Many of them are not quantities of interest in their research, since most do not directly coincide with causal effects. Instead, the only outcome of the data generation process that researchers get to observe is the log-mortality rate, m_t, and the log-mortality rate, or at least the average of multiple observations of it, serves as an excellent estimate of the expected log-mortality rate. As such, it is reasonable to think that analysts might have sufficient knowledge with which to form priors about the expected mortality rate, even if most of the coefficients are noncausal and on different scales.

Indeed, we find that when asking substantive experts for their opinion about what countries (or age groups, etc.) are alike, they are much more comfortable offering opinions about the similarity of expected mortality than regression coefficients. In fact, on detailed questioning, they have few real opinions on the coefficients even considered separately. This point thus follows the spirit of Kadane's focus on prior elicitation methods that are "predictive" (focusing on the dependent variable) rather than "structural" (focusing on the coefficients) (Kadane and Wolfson, 1998; Kadane et al., 1980; Kadane, 1980).

To see why priors on μ do not translate automatically into priors on $\boldsymbol{\beta}$ without further analysis, consider a simple version of the cross-sectional variation in the expected value of the dependent variable, $\mu_{it} \equiv \mathbf{Z}_{it}\boldsymbol{\beta}_i$, at one point in time t. This version is merely the difference between two cross-sections i and j, assuming that the covariates in cross-section i and j are of the same type

$$\begin{aligned} \mu_{it} - \mu_{jt} &= \mathbf{Z}_{it}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) + (\mathbf{Z}_{it} - \mathbf{Z}_{jt})\boldsymbol{\beta}_j \\ &= \text{Coefficient variation} + \text{Covariate variation} \end{aligned} \tag{4.9}$$

This expression decomposes the difference (or variation in) the expected value of the dependent variable into coefficient variation and covariate variation. A prior on variation in the expected value does not translate directly into coefficient variation because it ignores covariate variation. In other words, this expression demonstrates that having $\boldsymbol{\beta}_i \approx \boldsymbol{\beta}_j$ does *not* guarantee that the expected value of the dependent variable assumes similar values in cross-section i and j, because of the term $(\mathbf{Z}_{it} - \mathbf{Z}_{jt})\boldsymbol{\beta}_j$, which is not necessarily small. Obviously the more similar $\mathbf{Z}_{it}$ is to $\mathbf{Z}_{jt}$ the smaller is this term. However, there is no reason, a priori, for which two cross-sections with similar patterns of mortality should have similar patterns of the *observed* covariates: Some of the similarity may arise from patterns of the unobservables, or, when some of the covariates are "substitutes" of each other, from a different mix. For example, two countries might achieve similar patterns of mortality due to cardiovascular disease by different means: one could have first class surgical and pharmaceutical interventions that keep people alive but very poor public health and education facilities in preventing people from getting sick in the first place, and the

other could have the opposite pattern. In this situation, we would observe differences in covariates and their coefficients.

4.3.4 Difficulties with Incomparable Covariates

But even when the covariates behave in such a way that this extra source of variation is not an issue, another problem may surface. In the previous section, we implicitly assumed that all cross-sections share the same "type" of covariates and the same specification. However, the dependent variable may have different determinants in different cross-sections, and some covariates may be relevant in some cross-sections but not in others. For example, in forecasting mortality rates, we know that fat and cigarette consumption are important determinants of mortality, but these covariates are observed only in a non-random subset of countries. Similarly, we would not expect the availability of clean water to explain much variation in mortality rates in most of the developed world. In this situation, we could pool the corresponding coefficients only in the cross-sections for which these covariates are observed, but then we might introduce unpredictable levels of pooling bias. In general, pooling coefficients is not a viable option when we have different covariates in different cross-sections.

Moreover, even when we have the same type of covariates in all cross-sections, pooling coefficients makes sense only if the covariates are directly comparable. A simple example is the case of GDP: if we want to pool the coefficients on GDP not only will this covariate have to be expressed in the same currency (say US 1990 dollars), but also subjected to further adjustments such as purchasing power parity (PPP), which are not trivial matters. Having a variable with the same name in different countries does not guarantee that it means the same thing. If it does not, substance matter experts would have no particular reason to believe that the coefficients in one cross-section would be similar to that in another country, because the coefficients themselves would mean entirely different things.

4.4 Priors on the Expected Value of the Dependent Variable

In this section we show how to address the issues from Section 4.3 using the simple idea of focusing attention on the expected value of the dependent variable, rather than on the coefficients. Researchers may know fairly precisely how the expected value is supposed to vary across cross-sections, or something about its behavior over time, or interactions among these or other variables.

However, to get the usual Bayesian modeling technology to work, we ultimately need priors expressed in terms of the coefficients since they are the parameters to be estimated. We therefore propose the following two-step strategy, aimed at deriving a prior density on the regression coefficients, but constructed from knowledge of priors

specified on the expected value of the dependent variable. First, we specify the prior in terms of the expected value and then we add information about the functional form and translate it into a prior on the coefficients.

4.4.1 Step 1: Specify a Prior for the Dependent Variable

We begin by thinking non-parametrically (i.e., qualitatively, before entertaining a specific parametric functional form) about the expected value of the dependent variable as a function μ_{it} of the cross-sectional index i ($i = 1, \ldots, N$) and time t ($t = 1, \ldots, T$). Although μ is naturally thought of as an $N \times T$ matrix, it will be more convenient to think of it as the column vector in $\mathbb{R}^{T \times N}$ obtained by concatenating the N time series corresponding to the different cross-sections one after the other. The experts' knowledge can be seen as a set of L statements about properties of μ, and we assume that it is possible to translate them into L formulas of the form:

$$H_l[\mu] \quad \text{should be small} \quad l = 1, \ldots L \tag{4.10}$$

where H_l are functionals of μ (a functional is a map from a set of functions into the set of real numbers).[3]

Example 1 A simple example is the case where we believe μ varies very little over time in each cross-section but the different cross-sections have no necessary relationships. In this situation, we could have $L = N$, and $H_i[\mu]$ could be the average rate of temporal variation of μ in cross-section i. If we think that this set of constraints is too restrictive we may want to enforce our statements only on average, and then replace the N functionals H_i with the single functional $\sum_i H_i[\mu]$. ⊠

Example 2 Suppose we know that, for any given year t, the cross-sectional profile of μ_{it} should be not too far from some specified profile g_i over cross-sections, but we have no prior beliefs about time series patterns. Then we could set $L = T$ and $H_t[\mu] = \sum_i (\mu_{it} - g_i)^2$, for all t. Alternatively we may want to enforce our statements only on average, and then replace the T functionals H_t with the single functional $\sum_t H_t[\mu]$. ⊠

[3]The idea that prior knowledge can be represented in statements like in Equation 4.10 is very old. In its simplest form we see it in the method of graduation (Whittaker, 1923; Henderson, 1924), but its broad and full formalization was given by Tikhonov in the framework of regularization theory (Tikhonov, 1963; Tikhonov and Arsenin, 1977; Morozov, 1984), where the functionals H_l are usually called *stabilizers* or *regularization functionals*. Similar ideas appear in the theory of splines, starting with the seminal work of Schoenberg (Schoenberg, 1946; De Boor, 1978, Wahba, 1990) where the functionals H_l are usually called *smoothness functionals*. The fact that we build a prior density starting from a stabilizer is not by chance: there is a deep connection between regularization theory and Bayesian estimation, which was first unveiled by Kimeldorf and Wahba (Kimeldorf and Wahba, 1971; Wahba, 1990), as well as one between the method of graduation and Bayesian estimation (Taylor, 1992; Verrall, 1993).

We now put the statements above in a probabilistic form. Think of μ as a random variable, and define a normal probability density on μ as

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}\sum_l \theta_l H_l[\mu]\right) \equiv \exp\left(-\frac{1}{2}H[\mu,\theta]\right) \ , \ \mu \in \mathbb{R}^{T\times N} \tag{4.11}$$

where $\theta = (\theta_1, \ldots, \theta_l)$ is a set of positive parameters (often called hyper-parameters, regularization parameters or smoothing parameters). The choice of the exponential function in the equation above is a matter of convenience at this point and is in line with practical applications, while the factor $\frac{1}{2}$ is there only to simplify future calculations. The probability density in Equation 4.11 assigns high probability only to those configurations such that $\theta_l H_l[\mu]$ is small for all l, which is precisely what we want in a formal version of Equation 4.10. The parameters θ_l control how small we want $H_l[\mu]$ to be. A simple but important observation is that they control all the moments of the prior density. Therefore if additional information is available about some moments (for example we may have an idea of what the variance of μ might be) then these parameters could be determined. In general they will be known with some uncertainty, and therefore they are usually taken to be random variables with known prior distributions. Since their precise value is not relevant in this section we take them as user-specified for the moment, and will consider the problems of their choice in Chapter 6 and estimation in Chapter 9.

4.4.2 Step 2: Translate to a Prior on the Coefficients

Equation 4.11 is a convenient and flexible way to summarize prior knowledge but it only tells half of the story, the other half being told by the covariates. In fact, the density in Equation 4.11 is defined over the entire space $\mathbb{R}^{T\times N}$, and assigns positive probability to *any* realization of the vector μ. However, this does not take into account the linear specification in Equation 4.1, which says that only the values of μ explained by the covariates $\mathbf{Z}$ can be realized, that is μ must lie in some subspace $\mathbb{S}_{\mathbf{Z}} \subset \mathbb{R}^{T\times N}$. Therefore the prior 4.11 is only valid in the subspace $\mathbb{S}_{\mathbf{Z}}$, and it should be set to 0 outside it.[4]

To formalize this result, we rewrite the specification in Equation 4.1 in matrix form as $\mu = \mathbf{Z}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^{\sum_i k_i}$ is the column vector obtained by concatenating the N vectors $\boldsymbol{\beta}_i$, k_i is the number of covariates in cross-section i, and $\mathbf{Z}$ is a diagonal, block matrix with the data matrices $\mathbf{Z}_i$ as the blocks on the diagonal. Then the set $\mathbb{S}_{\mathbf{Z}}$ is formally defined as $\mathbb{S}_{\mathbf{Z}} \equiv \{\mu \in \mathbb{R}^{T\times N} \mid \mu = \mathbf{Z}\boldsymbol{\beta} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^{\sum_i k_i}\}$. We summarize this information by writing:

$$\mathcal{P}(\mu \mid \theta) \propto \begin{cases} \exp\left(-\frac{1}{2}H[\mu,\theta]\right) & \text{if } \mu \in \mathbb{S}_{\mathbf{Z}} \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

[4]Methods related in spirit to the idea described here can be found in analyses of prior elicitation in model selection problems (see Ibrahim and Chen, 1997; Laud and Ibrahim, 1996, 1995; Oman, 1985; Weiss, Wang and Ibrahim, 1997).

It is now clear that *on the subspace* $\mathbb{S}_{\mathbf{Z}}$*, that is on the support of the prior, the relationship* $\mu = \mathbf{Z}\boldsymbol{\beta}$ *is invertible* by the usual formula $\boldsymbol{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mu$ (assuming only that $\mathbf{Z}$ is of full rank). This result implies that we can use Equation 4.12 to derive a probability distribution for $\boldsymbol{\beta}$. Since the transformation $\mu = \mathbf{Z}\boldsymbol{\beta}$ is linear its Jacobian is an irrelevant constant (the tools to compute it are presented in Appendix C), and we obtain the probability density for $\boldsymbol{\beta}$ by simply plugging $\mu = \mathbf{Z}\boldsymbol{\beta}$ in Equation 4.12. Therefore, we write Equation 4.12 in terms of the coefficients $\boldsymbol{\beta}$ as

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \begin{cases} \exp\left(-\frac{1}{2}H^{\mu}[\boldsymbol{\beta}, \theta]\right) & \text{if } \mu = \mathbf{Z}\boldsymbol{\beta} \in \mathbb{S}_{\mathbf{Z}} \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

where

$$H^{\mu}[\boldsymbol{\beta}, \theta] \equiv H[\mathbf{Z}\boldsymbol{\beta}, \theta] \tag{4.14}$$

Equation 4.13 contains exactly the same amount of information as contained in Equation 4.12: The fact that $\mathcal{P}(\mu \mid \theta)$ is 0 outside of $\mathbb{S}_{\mathbf{Z}}$ is expressed by the expression $\mu = \mathbf{Z}\boldsymbol{\beta}$, and the density of μ on $\mathbb{S}_{\mathbf{Z}}$ implied by the prior on the coefficients in Equation 4.13 is, by construction, the one of Equation 4.12.

For clarity of notation, we use the superscript μ in $H^{\mu}[\boldsymbol{\beta}, \theta]$ to remind us that, unlike the version in Equation 4.4 used to smooth directly based on the coefficients, this density has been derived using knowledge about μ. (Since in our formulation the prior densities are always of the form 4.13 for some appropriate choice of the function in the exponent, we will often refer to the function in the exponent as "the prior," without the risk of confusion.)

The important fact about Equation 4.13 is that it has the desired property: It assigns high probability only to configurations of the coefficients $\boldsymbol{\beta}$ such that the corresponding predicted values of the dependent variable $\mu = \mathbf{Z}\boldsymbol{\beta}$ have high probability, which conforms to our prior knowledge. Notice that the specification $\mu = \mathbf{Z}\boldsymbol{\beta}$ holds for the years for which we have observations as well as for the years for which we need to make a forecast: this implies that the covariates in the expressions above have a temporal range which extends into the future, and that prior knowledge is enforced on both the past and the future of the expected value of the dependent variable.[5]

After these steps have been performed, the rest is standard Bayesian analysis: We plug the prior of Equation 4.13 into the expression for the posterior distribution of Equation 4.2, leaving to be solved only the computational problem of calculating Equation 4.3, which we address in Chapter 9.

Chapters 5 and 7 are devoted to showing how our approach works in practice by deriving explicit expressions for priors on $\boldsymbol{\beta}$ in a diverse variety of important cases. The pleasant surprise is that, for a wide choice of smoothness functionals $H[\mu, \theta]$, the implied prior for $\boldsymbol{\beta}$ turns out to have a mathematical form which is very similar to the one in Equation 4.5, which is well understood, but without its shortcomings (discussed in Section 4.3).

[5]This implies that the notation $\sum_t$ has different meanings when it appears in the likelihood and in the prior, but for notational simplicity, we do not distinguish between the two.

4.5 A Basic Prior for Smoothing over Age Groups

In order to keep the mathematical level at a minimum we start with a very simple prior on μ, which, although not recommended in most applications, generates a prior for the coefficients with all the relevant characteristics.

4.5.1 Step 1: a Prior for μ

We assume for the moment that there is only one country and A age groups, so that the expected value of the dependent variable in age group a at time t is μ_{at}, with $a = 1, \dots A$ and $t = 1, \dots, T$. We consider a very simple form of prior knowledge: At any point in time *nearby age groups have similar values of* μ. A simple way to represent this kind of knowledge is based on the average squared difference expected log-mortality in adjacent age groups (averaged over time periods):

$$H[\mu, \theta] \equiv \frac{\theta}{T} \sum_t \sum_{a=1}^{A-1} (\mu_{at} - \mu_{a+1,t})^2 \ , \quad \text{should be small.}$$

This smoothness functional has two important properties:

1. It takes on small values only when nearby age groups have similar values of μ;
2. It is indifferent to arbitrary, time dependent, shifts constant across the age profiles. More precisely, it is invariant with respect to the transformation:

$$\mu_{at} \rightsquigarrow \mu_{at} + f_t \qquad\qquad \forall f_t \in \mathbb{R}$$

   In other words, for any given year, the prior associated with the functional above suggests we are ignorant with respect to the level of the dependent variable. (We formalize and generalize this notion of prior indifference in Section 5.1.)

We now rewrite the functional above in a form more amenable to generalization. First define the matrix s such that $s_{aa'} = 1$ if and only if $|a - a'| = 1$, and 0 otherwise. Using this notation the functional above can be written as:

$$H[\mu, \theta] \equiv \frac{\theta}{2T} \sum_t \sum_{aa'} s_{aa'} (\mu_{at} - \mu_{a't})^2 = \frac{\theta}{T} \sum_t \sum_{aa'} W_{aa'} \mu_{at} \mu_{a't} \tag{4.15}$$

where we have defined the matrix $W = s^+ - s$ (see Appendix B, Page 253). Therefore the prior density for μ has the form

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{\theta}{2T} \sum_t \sum_{aa'} W_{aa'} \mu_{at} \mu_{a't} \right). \tag{4.16}$$

One feature of the prior density in Equation 4.16 is that it has zero mean and is symmetric around the origin, so that the probability of an age profile and its negative are the same. Depending on the application this may or may not be appropriate. For example this is not realistic when analyzing logged mortality rates: In this case, we know that, in any given year, the age profiles will look, on average, like some cause-specific "typical" age profile $\bar{\mu} \in \mathbb{R}^A$. Fortunately, it is easy to modify the prior to take this information in account, by letting the prior have mean $\bar{\mu}$ in any year:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{\theta}{2T}\sum_{t}\sum_{aa'} W_{aa'}(\mu_{at} - \bar{\mu}_a)(\mu_{a't} - \bar{\mu}_{a'})\right) \tag{4.17}$$

One issue is where $\bar{\mu}$ comes from. One productive procedure is to have an expert draw pictures of his or her expectations for average log-mortality age profile, $\bar{\mu}$. Alternatively, a reasonable typical age profile $\bar{\mu}$ could be synthesized from the data, or borrowed from other countries not in the analysis, possibly after some preprocessing and smoothing, and subject to the "approval" of some expert. In this case it looks like we have a data dependent prior, which of course would not seem "prior" and so would not appear appropriate. However, this problem is easily solved by noticing that using a prior which is not mean-zero is equivalent to using a mean-zero prior in which we have replaced the dependent variable m_{at} with $m_{at} - \bar{\mu}_a$. In the following, therefore, we keep using Equation 4.16 rather than the more cumbersome Equation 4.17, where we keep in mind that, depending on the application, μ may either be the expected value of the dependent variable or its deviation from the typical age profile $\bar{\mu}$.

4.5.2 Step 2: from the Prior on μ to the Prior on $\boldsymbol{\beta}$

We now proceed to the second step, and substitute our specification in the functional in Equation 4.15. In this case the specification is simply $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$, and substituting it into Equation 4.15 we obtain:

$$\begin{aligned} H^{\mu}[\boldsymbol{\beta}, \theta] &\equiv \frac{\theta}{T}\sum_{aa't} W_{aa'}(\mathbf{Z}_{at}\boldsymbol{\beta}_a)(\mathbf{Z}_{a't}\boldsymbol{\beta}_{a'}) \\ &= \theta \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'} \end{aligned} \tag{4.18}$$

where the second line uses the fact that the coefficients $\boldsymbol{\beta}$ do not depend on time and so the sum over time can be performed once for all, and where we have defined the matrix:

$$\mathbf{C}_{aa'} \equiv \frac{1}{T}\mathbf{Z}'_a\mathbf{Z}_{a'}$$

so that $\mathbf{Z}_a$ is the usual data matrix of the covariates in cross-section a, which has $\mathbf{Z}_{at}$ for each row. Hence, the prior for $\boldsymbol{\beta}$, conditional on the parameter θ, is now simply

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}\right) \tag{4.19}$$

This is the long-sought-for prior over the coefficients, which we have built using only prior knowledge on the expected value of the dependent variable μ. Its most remarkable characteristic is the fact that, *since the covariates* $\mathbf{Z}_{at}$ *and* $\mathbf{Z}_{a't}$ *are of dimensions* k_a *and* $k_{a'}$*, respectively, then* $\mathbf{C}_{aa'}$ *is a rectangular* $k_a \times k_{a'}$ *matrix, and it does not matter whether we have the same number or type of covariates in the two cross-sections* a *and* a'*.*

4.5.3 Interpretation

While we postpone to Chapter 8 a thorough comparison between this prior and the prior we would have obtained by imposing smoothness of the coefficients over age groups, as described in Section 4.2, it is useful to write them side by side as follows:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}\right) \quad \Leftrightarrow \quad \mathcal{P}(\boldsymbol{\beta} \mid \Phi) \propto \exp\left(-\frac{1}{2} \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \Phi \boldsymbol{\beta}_{a'}\right)$$

This comparison helps us emphasize that focusing on the expected value of the dependent variable allows us to solve two problems at the same time: (1) we replaced an entire unknown matrix Φ with the quantities $\theta \mathbf{C}_{aa'}$, which are known up to a single scalar parameter; and (2) our formulation allows each cross-section to have its own specification and therefore for different covariates in different cross-sections. In addition, the new prior, although conceptually very different from the usual prior over the coefficients, is computationally similar, and does not imply any additional difficulties from the point of view of the implementation.

In addition, while having real prior knowledge about Φ is extremely rare, the one remaining parameter in our formulation, θ, is directly linked to quantities which we can directly interpret and on which we are likely to have prior information. Although this will be explored in detail in Section 6.2, we report here a key result. If we let $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$, the quantity

$$\frac{1}{T} \sum_t \sum_{aa'} s_{aa'} (\mu_{at} - \mu_{a't})^2 = \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'},$$

which appears in the exponent of our prior, represents, an average over time and in a mean square sense, *how much the expected value of the dependent variable varies from one age group to the next.* Postponing some technicalities related to the fact that the prior in Equation 4.19 is improper, the expected value of the quantity above under the prior in Equation 4.19 is:

$$E\left[\sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'}\right] = \frac{K}{\theta},$$

where K is a number that depends on W and the matrices $\mathbf{C}_{aa'}$ and that can be easily computed. In most applications, we will have an idea of how much the expected log-mortality rate varies between adjacent age groups, and so we could easily specify a range of values for the left side of this equation. This then immediately leads to a range of values for θ, and therefore to a prior density $\mathcal{P}(\theta)$. *This result underlines the main philosophy of our approach: We write priors using only quantities we understand and for which we have genuine prior knowledge.*

Before we proceed to analyze more sophisticated priors, two key remarks are in order:

- As pointed out above, this prior is invariant with respect to the transformation $\mu_{at} \rightsquigarrow \mu_{at} + f_t \ , \ \ \forall f_t \in \mathbb{R}$. This implies that the prior is constant over an entire subspace of its domain, and therefore its integral over the domain is infinite: in other words the prior is improper. This causes no difficulties since our likelihood, and therefore our posterior, are proper. It is also a highly desirable feature of a prior, smoothing expected values of the dependent variable toward each other but without requiring one to specify at what specific level they smooth toward. Note that this feature is distinct from (proper) priors that are merely diffuse with respect to a parameter. This prior will have *no* effect on constant shifts in the posterior, no matter how much weight is put on it (or, correspondingly, how small we make its variance).

- We mentioned above that the prior presented is not necessarily what we recommend in applications. The reason for this can be seen by rewriting it as:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{\theta}{2T}\sum_t \sum_{a=1}^{A-1}(\mu_{at} - \mu_{a+1,t})^2\right) \tag{4.20}$$

  This version illuminates the feature of this prior that increments in log-mortality between adjacent age groups are independent, and therefore samples from this prior will tend to look as a random walk (as a function of *age*, not time), and therefore it will not be particularly smooth. In Figure 4.1, we present samples from the prior above (with each draw an $A \times 1$ vector represented by one line), for a fixed time t. There are 17 age groups, at 5 years intervals, so that $A = 17$. The left graph gives the case of a zero mean, while in the right graph, we add a mean to the prior, where the mean is a typical age profile for all-cause male mortality. Each age profile is quite jagged over age groups, indicating that it is not a very good representation of our prior knowledge. The reason for the lack

of smoothness is that the increments between nearby age groups are specified to be independent. We will see in Chapter 5 that by introducing correlations between the increments, much smoother and hence much more appropriate age profiles can be obtained.

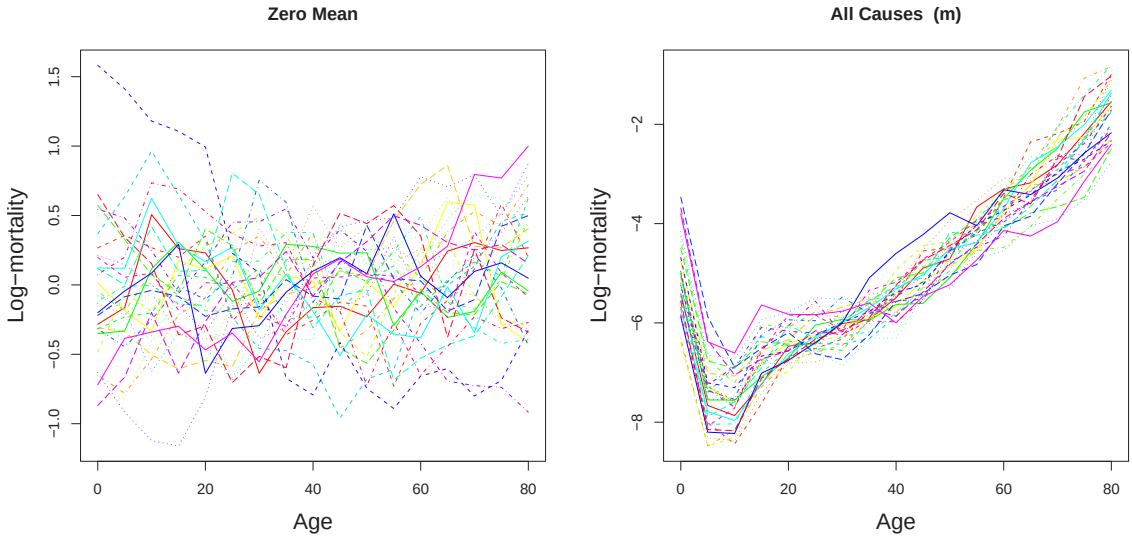

Figure 4.1: Samples from the prior in Equation 4.20 with zero mean (for the left graph) and non-zero mean (for the right graph). There are 17 age groups, at 5 years intervals. For the right graph the mean $\bar{\mu}$ is the average age profile for all cause mortality in males. The average has been computed over all the countries with more than 20 observations and over all the available years. The value of θ has been chosen so that the standard deviation of μ_a is 0.5, on average over the age groups. Notice the jagged behavior of each line in the two graphs, making these priors undesirable for most applications.

4.6 Concluding Remark

In the rest of this book, we develop specific models within the class of models defined in this Chapter. In particular, Chapters 5 and 7 repeatedly use the two steps offered in Section 4.4 for each of a variety of different data types to derive new models under the framework of this chapter.

Although we use the simple linear-normal likelihood throughout our work, nothing in our approach would necessarily constrain one to that model. Our methods for specifying the priors on the expected value of the dependent variable need not be changed at all for other likelihood models, such as nonlinear functional forms or non-normal densities. There would be different computational considerations of course,

but the general approach offered here still applies.

Chapter 5

Priors Over Grouped Continuous Variables

In this chapter, we define a prior for the similarity of a set of cross-sections ordered by a discretized continuous variable, such as a set of age groups. This extends our version of Bayesian spatial models that smooth on the dependent variable rather than coefficients to quantities that vary over conceptual, rather than geographic, space. From a practical point of view, a key result given here is a method of defining the entire spatial contiguity matrix that is a function of only a single adjustable parameter. We begin in Section 5.1 by developing a specific notion of prior indifference that we use in the rest of the chapter and book. A detailed example of smoothing over age groups then appears in Section 5.2. Chapter 6 follows this analysis with practical methods for making the various necessary choices in using these priors, and Chapter 7 develops priors for the similarity of vectors defined over time and geographical space, as well as combinations of these dimensions.

5.1 Definition and Analysis of Prior Indifference

The task of choosing a prior density for a Bayesian model involves clarifying and formalizing one's knowledge about the likely values of and patterns in a set of parameters, but it also involves specifying what one is indifferent to. A good prior must obviously be informative about the former but not about the latter. For example, demographers and public health experts are normally confident that the expected log-mortality rate varies smoothly over age groups and is likely to stay that way in the future, but they are normally less willing to offer an opinion about precisely what level the rate will be at for any particular year, country, cause, or sex group.

A huge literature in statistics attempts to formalize what information, or lack of it, is represented by a prior, and especially how we can write priors that are minimally informative. Deep philosophical issues arise, primarily around how to represent com-

plete ignorance in the form of a specific probability density — a philosophical stance sometimes known as "logical Bayesianism". Numerous creative ideas have been suggested to try to achieve this goal in some part, such as making the prior invariant to reparameterization (Jeffreys, 1961), but as is recognized the task is ultimately impossible: Here, as everywhere, counting on scientific progress about purely philosophical issues would not be prudent. As Dawid (1983, p.235) writes, "The formalization of ignorance thus remains the central object of a continuing quest by the knights of the Bayesian round table: inspiring them to imaginative feats of daring, while remaining, perhaps, forever unattainable."

The problem, from the perspective of the philosophy of inference, is that a prior density is a specific probabilistic statement and thus represents considerable knowledge, even if the density is described as "flat," "diffuse," or "noninformative." Following a detailed review of the practical choices offered in this literature, Kass and Wasserman (1996, p.1343) recommend the choice of "reference priors" for standard problems, but nevertheless conclude that "it is dangerous to put faith in any 'default' solution" unless the prior is dominated by the data. Of course, if the prior is dominated by the data, then likelihood inference will normally work well, the special features of Bayesian inference vanish, and of course no prior needs to be specified in the first place (except sometimes for computational reasons like MCMC algorithms).

In our work, we see no reason to subscribe to the Bayesian religion as a way to make all inferences, but we do find its associated technology to be exceptionally useful when prior knowledge is available, especially in complicated models. When prior knowledge is not available, the likelihood theory of inference is a perfectly adequate approach (Edwards, 1972; King, 1989*b*). We thus feel no driving normative need to state a philosophical view on representing ignorance from a purely Bayesian perspective. If we have a philosophical viewpoint, it is utilitarianism (or consequentialism), which in our view is almost by definition the appropriate philosophy when the only relevant normative criterion is creating something useful. Utilitarianism may not answer the desire of philosophers of science for a self-contained, logically consistent, and normatively satisfying theory of inference, but it works.

The main problem we tackle here is that we often know some things and not others about the same set of parameters, and wish to write informative priors only for the things we know. For the quantities we do not know we cannot write a proper prior, and so we use a flat, constant (improper) prior. We see no need to justify the constant prior by appeal to the "principle of insufficient reason" (Laplace, 1951, original: 1820) or other such concepts (that themselves are based on insufficient reason!). Instead, we view this choice as a simple combination of likelihood and Bayesian inference: When we have information we use it and the likelihood; when we have no such information, we use only the likelihood. Our approach has much in common with the spirit of "robust Bayesian analysis" (Berger, 1994; King and Zeng, 2002), although the technology is very different. More relevant to our particular technical approach the pioneering work of Julian Besag and the literature on spatial smoothing (Besag,

1974, 1975; Besag and Kooperberg, 1995), as well as the work of Speckman and Sun (2001), and the whole literature on non-parametric regression and in particular the seminal work of Wahba (1978).

In the following, we define what we call *prior indifference*, or the indifference of a prior density to a specific set of patterns or values of a set of parameters. We borrow freely from the authors cited above, and others, and combine and extend strands of literature from a diverse set of scholarly fields in order to present a simple but coherent approach. (This chapter requires only some linear algebra and basic mathematical concepts. For readers not familiar with the mathematical concepts we use, such as vector spaces, subspaces, orthogonality, and null spaces, we provide a self-contained review in Appendix B, and a glossary of our notation in Appendix A.)

We begin with an elementary observation about the simplest possible case and build from there.

5.1.1 A Simple Special Case

Consider the problem of writing a prior for the d components $\mu_1, \ldots, \mu_d$ of some vector μ. If we know something about the first r components, but not about the last $n = d - r$ components, then we would write a prior which simply does not depend on, or is *indifferent* to, the last n components. This prior would have the property:

$$\mathcal{P}(\mu_1, \ldots, \mu_r, \mu_{r+1}, \ldots, \mu_d) = \mathcal{P}(\mu_1, \ldots, \mu_r, \mu'_{r+1}, \ldots, \mu'_d) \ , \quad \forall \mu_i, \mu'_i \in \mathbb{R} \ , \quad i = 1, \ldots, d \tag{5.1}$$

The dependency on the first r variables would be determined by what we know about them. Notice that this prior is obviously improper, since the integral over the last n variables is infinity. This will never be a problem in our applications, since (because our likelihood is proper) our posteriors will always be proper. Improperness therefore is relevant only as a side-effect of the assumption of indifference to some of the parameter space.

A good way to understand prior indifference in this simple special case, and indeed in any more general specification, is to imagine weighting the prior as heavily as possible (or, equivalently, letting its variance tend toward zero). Even in this extreme situation, our prior will have absolutely no influence over the last n parameters. In contrast, a proper prior in this situation would cause the estimation procedure to ignore what the data (and likelihood) have to say about the parameters; it would force the posterior to degenerate to a spike over each parameter, thus allowing the posterior to reflect only the *single value* for each parameter chosen by the investigator in setting the hyperparameters. In contrast, our improper priors, when maximally weighted, only constrain the posterior to a *subset* of the parameter space, known as the null space. The null space in this example is a subset of parameters; in our other more general priors, the null space reflects particular patterns in the parameters.

Simple as it is, the formula above can take us very far if properly applied. The problem with it is that it *seems* unlikely that in our applications we can partition our set of parameters in two nonoverlapping subsets, one over which we have knowledge, and one over which we do not. We emphasize *seems* because, as we will see shortly, it is indeed the case that such a partition is always possible, although it may become apparent only after an appropriate linear change of variables.

5.1.2 General Expressions for Prior Indifference

In order to understand prior indifference better we first rewrite Equation 5.1 in a more abstract way. First define the following r-dimensional subspace of $\mathbb{R}^d$:

$$\mathbb{S}_\circ \equiv \{\mu \in \mathbb{R}^d \mid \mu = (0, 0, \ldots, 0, \mu_{r+1}, \ldots \mu_d)\} \tag{5.2}$$

Its r-dimensional orthogonal complement, that is the set of vectors in $\mathbb{R}^d$ which are orthogonal to all the elements of $\mathbb{S}_\circ$ (See Appendix B.1.11, Page 239), is then

$$\mathbb{S}_\perp \equiv \{\mu \in \mathbb{R}^d \mid \mu = (\mu_1, \mu_2, \ldots, \mu_r, 0, \ldots, 0)\}.$$

Clearly any vector $\mu \in \mathbb{R}^d$ can be uniquely decomposed into the sum of two vectors, one in $\mathbb{S}_\circ$, which we denote by $\mu_\circ$, and one in $\mathbb{S}_\perp$, which we denote by $\mu_\perp$. The vectors $\mu_\circ$ and $\mu_\perp$ can be obtained as linear transformations of the vector μ, that is $\mu_\circ = P_\circ \mu$ and $\mu_\perp = P_\perp \mu$, where the matrices $P_\circ$ and $P_\perp$ are called the *projectors* onto $\mathbb{S}_\circ$ and $\mathbb{S}_\perp$ respectively. The projector onto a subspace is uniquely determined by the subspace; that is for any given subspace we can easily derive the corresponding projector (as described in Appendix B.1.13, Page 240). Thus, in the present case, it is easy to see that:

$$P_\circ = \begin{pmatrix} 0_{r\times r} & 0_{r\times d} \\ 0_{d\times r} & I_{d\times d} \end{pmatrix} , \quad P_\perp = \begin{pmatrix} I_{r\times r} & 0_{r\times d} \\ 0_{d\times r} & 0_{d\times d} \end{pmatrix} .$$

Using this notation we rewrite our expression of prior indifference in Equation 5.1 as

$$\mathcal{P}(\mu) = \mathcal{P}(\mu + \mu*) , \quad \forall \mu \in \mathbb{R}^d , \quad \forall \mu^* \in \mathbb{S}_\circ. \tag{5.3}$$

We read this equation by saying that the prior $\mathcal{P}$ is constant over the subspace $\mathbb{S}_\circ$, or is indifferent to $\mathbb{S}_\circ$. Another way of rewriting this equation is as follows:

$$\mathcal{P}(\mu) = \mathcal{P}^*(P_\perp \mu) , \quad \text{for some probability density } \mathcal{P}^* \tag{5.4}$$

The last equation makes clear that $\mathcal{P}(\mu)$ does not depend on $\mu_\circ$, the part of the vector μ which is in the subspace $\mathbb{S}_\circ$.

5.1.3 Interpretation

The reason for rewriting Equation 5.1 as Equation 5.3 or 5.4 is that the latter two hold independently of the particular choice of coordinate system, and even if μ cannot be uniquely partitioned into nonoverlapping subsets. In fact, suppose we want to describe our system in terms of the random variable $\nu = B\mu$, for some invertible matrix B: The prior density of ν will not in general satisfy any equation of the form 5.1. However, an equation of the type 5.3 will still hold, where μ has been replaced by ν and $\mathbb{S}_\circ$ has been replaced with its image under the transformation B.

While it is rarely the case that we can naturally express our ignorance in the form of Equation 5.1 at first, it happens often that we can express it in the form 5.3, for appropriate choices of the subspace $\mathbb{S}_\circ$. In general the subspace $\mathbb{S}_\circ$ will be written in a different form from Equation 5.2, but this is irrelevant: all that matters is that *any vector $\mu \in \mathbb{R}^d$ can be written as the sum of two orthogonal parts, $\mu_\circ$ and $\mu_\perp$, such that we only have knowledge about $\mu_\perp$.*

Since $P_\perp \mu$ is a linear combination of the elements of μ, one way to interpret Equation 5.4 (and therefore Equation 5.3) is by saying that we have prior knowledge only about some particular linear combinations of the elements of μ. Notice also that, given any subspace $\mathbb{S}_\circ$, we could always find a change of variable $\nu = B\mu$ such that our notion of indifference, expressed in terms of ν, will take a form like the one of Equation 5.1. However, although it is good to know that this can be done, since it helps to clarify the fact that all we are doing is making separate lists of things we know and do not know, this exercise is not of practical interest, since Equations 5.3 and 5.4 can be used directly.

Example 1 Let $\mu \in \mathbb{R}^d$ be a vector of random variables. Assume, for example, that they represent the expected values of log-mortality in d different countries, for a given year. We refer to the set of d countries as the world. Suppose we have knowledge about some properties of μ but not about others. For example, we may not have any idea of what the world mean $\bar{\mu} \equiv d^{-1} \sum_i \mu_i$ of log-mortality should be, since data in some countries have never been collected. Hence, given two configurations whose elements differ by the same constant c we cannot say which one is most likely. This is equivalent to saying that, whatever prior density for μ we choose, it should have the property that

$$\mathcal{P}(\mu_1, \mu_2, \ldots, \mu_d) = \mathcal{P}(\mu_1 + c, \mu_2 + c, \ldots, \mu_d + c) \ , \quad \forall c \in \mathbb{R}$$

We now rewrite this expression by introducing the one-dimensional subspace

$$\mathbb{S}_\circ \equiv \{\mu \in \mathbb{R}^d \mid \mu = (c, c, \ldots, c) \ , \quad c \in \mathbb{R}\} \tag{5.5}$$

Thus, the equation above is equivalent to:

$$\mathcal{P}(\mu) = \mathcal{P}(\mu + \mu^*) \ , \quad \forall \mu^* \in \mathbb{S}_\circ \tag{5.6}$$

This suggests that the prior density should only be a function of $\mu_\perp = P_\perp \mu$, where $P_\perp$ is the projector onto the subspace of Equation 5.5.

What are the orthogonal complements, $\mu_\perp$ and $\mu_\circ$, in this case? It is easy to see that:

$$\mu_\circ = \bar{\mu}\ (1, 1, \ldots, 1)\ , \qquad \mu_\perp = (\mu_1 - \bar{\mu}, \mu_2 - \bar{\mu}, \ldots, \mu_d - \bar{\mu})$$

This result is intuitive: $\mu_\circ$ contains all the information about the global mean of μ, while $\mu_\perp$ contains all the remaining information, but no information about $\bar{\mu}$. In other words, if we are given $\mu_\perp$ we can reconstruct μ up to an additive constant, while if we are given $\mu_\circ$ we can only reconstruct its global mean. Given our (lack of) knowledge it is therefore to be expected that the prior should only depend on $\mu_\perp$.

Now that we have identified the subspace $\mathbb{S}_\circ$, and we know that the prior should be a function of $P_\perp \mu$, we could proceed to use additional pieces of information to constrain the prior further. A typical step would be to assume that it is normal and write,

$$\mathcal{P}(\mu) \propto \exp\left(-\frac{1}{2}\theta(P_\perp \mu)' B (P_\perp \mu)\right)$$

for some positive definite matrix B and some positive parameter θ which controls the size of the overall variance. In this expression, B represents our knowledge, and $P_\perp$ our ignorance. That is, $P_\perp$, when multiplied into μ, wipes out the piece of μ about which we wish to profess our ignorance (i.e., the null space). This expression can be rewritten as

$$\mathcal{P}(\mu) \propto \exp\left(-\frac{1}{2}\theta \mu' W \mu\right), \tag{5.7}$$

where we have defined the matrix $W \equiv P_\perp B P_\perp$ (remember that $P_\perp$ is symmetric). The only difference between this prior and a regular normal prior is that here, since $P_\perp$ is singular, the matrix W is singular and admits a non-trivial null space $\mathfrak{N}(W)$. Recall that the null space of a matrix W is the set of vectors μ such that $W\mu = 0$ (see Appendix B.2.1, Page 243 for more detail). In this case the null space $\mathfrak{N}(W)$ coincides with $\mathbb{S}_\circ$, from Equation 5.5. Since W is singular, the prior is improper, as expected. The improperness comes only from the existence of the null space, which is the set of vectors "invisible" to W; that is, $W\mu = W(\mu + \mu^*)$ for any $\mu^* \in \mathfrak{N}(W)$.

Although improper, the prior is still meaningful, as long as we think of a vector μ not as an individual element of $\mathbb{R}^d$, but as an equivalence class, obtained by adding to μ the arbitrary constants c to all its elements. Under this view, all the usual operations performed on prior densities, such as computation of the moments, can be performed on this prior (see Appendix C for details). For example, when we say that the prior above has zero mean what we are really saying is that the mean of the prior is known to be zero up to the addition of an arbitrary element of $\mathfrak{N}(W)$. ⊠

Although the example presented above is very simple, the final form in Equation 5.7, with W singular and positive semi-definite, closely represents all the priors we consider in this book.[1] The advantage of priors of this form is that the matrix W not only encodes information about the quantities we know (their correlations, for example), but also, through its null space, defines the subspace to which the prior is indifferent.

Our approach then follows two steps.

- First, we use the concept of *the null space of a matrix* to analyze W, the advantage of which is that the tools in linear algebra to characterize and analyze null spaces are well developed.

- Second, we note, that when a pattern in or values of the parameters μ are in the null space of W, then the prior in Equation 5.7 has the property of prior indifference given in Equation 5.6.

To understand prior indifference, then, we only need to understand the null space of W.

We also add slightly novel terminology by referring to the null space $\mathfrak{N}$ of the matrix W in Equation 5.7 as *the null space of the prior*. Since the expression $\mu' W \mu$, with W singular and positive semi-definite, defines a semi-norm (see Appendix B, page 233), it would be more appropriate, and more in line with some literature, to refer to $\mathfrak{N}$ as "the null space of the semi-norm associated with the prior," but this terminology seems unnecessarily complicated and so we do not adopt it here.

Before proceeding to a full analysis of several classes of priors, we point out that partially informative priors can also be used to force the random variables to assume a certain *class* of configurations with high probability, without requiring them to take on any *one* configuration as would be the case with a proper prior. To see this, consider the prior in Equation 5.7, with its null space $\mathfrak{N}(W)$, and let θ assume larger and larger values. This will force the proper part of the prior to become increasing concentrated around $\mu_\perp = 0$, but still leave $\mu_\circ$ unaffected. Plugging such a prior in the posterior is then equivalent to constraining the solution to the entire null space rather than to a point, as would be the case for a proper prior. Which element of the null space corresponds to the solution will then be determined by the data through the likelihood. If we built the prior in such way that the null space is a set of configurations with "desirable" properties, then we will have found the configuration

[1] Priors similar to that in Equation 5.7, often defined over an infinite set of random variables, are commonly called "partially improper" or "partially informative" priors. They play a fundamental role in nonparametric regression (Speckman and Sun, 2001; Wahba, 1975, 1978, 1990), where, among other things, they provide the link, originally unearthed by Kimeldorf and Wahba (1970), between spline theory and Bayesian estimation. The importance and usefulness of the prior being improper was stressed by Wahba (1978), who pointed out that it can be used as a mechanism to guard against model errors. Priors similar to this form also appear in the spatial statistics literature often under the name of "(conditionally) intrinsic autoregressive" priors.

with these properties that best fit the data. We explore this observation more in detail in the following example.

Example 2 Let μ_t be a random variable representing the expected value of log-mortality in a given cross-section. We take t to be a continuous variable for the purpose of explanation, and discretize it later. Consider the case in which we know that the time series describes a seasonal phenomenon and must have (approximately) the following form:

$$\mu_t = \gamma_1 \sin(\omega t + \gamma_2)$$

where we know ω but we have no idea about the parameters γ_1 and γ_2 (higher frequencies could be included, but exclude them for simplicity). Now notice that the time series above satisfies the following differential equation, independently of the value of the parameters γ_1 and γ_2:

$$\left(\frac{d^2}{dt^2} - \omega^2\right)\mu_t \equiv L\mu_t = 0$$

where the differential operator L is defined by the parenthetical term on the left side of the equation. The set of solutions of this differential equation is obviously a subspace of the set of all possible time series, defined as the null space $\mathfrak{N}(L)$ of the operator L, in analogy with the definition of the null space for matrices. Our ignorance over the possible values of γ_1 and γ_2 implies that we are indifferent over the null space of L. However, we also know that the time series must lie, approximately, in $\mathfrak{N}$, since it must have that particular form.

Now discretize the time series so that it has length T and replace the differential operator L with the corresponding $T \times T$ matrix L. An appropriate prior for this problem could have the following form:

$$\mathcal{P}(\mu) \propto \exp\left(-\theta \|L\mu\|^2\right)$$

where θ is some large number. This prior will assign high probability only to those configurations such that $L\mu$ is approximately 0, but will not specify, among those, which are the most likely. Notice that since $L\mu = L\mu_\perp$ this prior is written as a function of $\mu_\perp$ only, as expected. ⊠

5.2 Step 1: A Prior for μ

In this section we consider the case in which the cross-sectional index is a variable like age, which is intrinsically continuous, although it is discretized in practice. We proceed in two distinct steps, as outlined in Section 4.4: In this Section, we build a

non-parametric (qualitative) prior for the expected value of the dependent variable, and then, in Section 5.3, use it along with an assumption about the functional form to derive a prior for the regression coefficients $\boldsymbol{\beta}$.

Begin by setting the index $i = a$ and think of age as a continuous variable for the moment, so that the expected value of the dependent variable is a function $\mu(a, t) : [0, A] \times [0, T] \to \mathbb{R}$. The reason for starting from a continuous variable formulation is that in so doing we can borrow from the huge literature on non-parametric regression and splines, where smoothness functionals are commonly used. A potential problem of such an approach is that when μ is a function it is more difficult to give rigorous meaning to expressions such as $\mathcal{P}(\mu \mid \theta) \propto \exp(-H[\mu, \theta])$, since there are some non-trivial mathematical technicalities involved in defining probabilities over sets of functions. We need not to worry about this issue, though, since we will discretize the function μ and the smoothness functional $H[\mu, \theta]$ before defining any probability density, which will therefore always defined in terms of a finite number of variables.

We assume that we have the following prior knowledge: At any point in time the expected value of the dependent variable μ is a smooth function of age. By this we mean that *nearby age groups have similar values of* μ. We now formalize this idea.

5.2.1 Measuring Smoothness

Our immediate goal is to find functionals $H_t[\mu]$ defined for any time t which are small when μ is a smooth function of a (remember that a functional is a map from a set of functions to the set of real numbers; See Appendix B.1.6, Page 236). Functionals of this type are easily constructed using the observation that the oscillating behavior of a function is amplified by the application of any differential operator to it. Therefore an initial candidate for $H_t[\mu]$ could be:

$$H_t[\mu] \equiv \int_0^A da \left(\frac{d^{\mathtt{m}} \mu(a, t)}{da^{\mathtt{m}}} \right)^2 \quad \text{should be small } \forall t \in [0, T] \tag{5.8}$$

where $\mathtt{m}$ is an arbitrary integer which will be referred to as *the degree of smoothness*, for reasons which will become clear shortly. The parenthetical measures the slope (or higher derivatives) of $\mu(a, t)$ as a function of age for any time t. The squared term recognizes that "smoothness" is unaffected by the sign of the slope. Finally, the integral computes the average (of the squared slope) over different ages. In other words, Equation 5.8 is the expected value of the squared derivative, taken with respect to the uniform probability density (with the constant factor $1/A$ representing the uniform density omitted for simplicity). For related ideas in spline theory, see Schoenberg (1946), de Boor (1978), and Wahba (1975, 1990).

Example In order to convince ourselves that this functional does what we expect it to do we compute it for a family of wiggly functions and check that it gets larger if we make the functions more wiggly. Fix $t = 1$ and take the family $\mu_k(a, 1) = \sin(\frac{2\pi k a}{A})$,

indexed by the integer k. These are sin waves of frequency proportional to k, so k is a measure of how wiggly these functions are. First notice that, taking $\mathbb{m}$ even for simplicity, we have:

$$\frac{d^{\mathbb{m}}\mu_k(a,1)}{da^{\mathbb{m}}} = \left(\frac{2\pi k}{A}\right)^{\mathbb{m}} \sin\left(\frac{2\pi k a}{A}\right)$$

Therefore taking the derivative of order $\mathbb{m}$ amplifies the magnitude of the function of a factor $k^{\mathbb{m}}$. This is easily seen in Figure 5.1, where we show the function above and its derivative of order $\mathbb{m} = 2$ for the values $k = 2$ and $k = 5$: While the amplitude of the sin function is independent of k, the derivative corresponding to $k = 5$ has much larger amplitude than the derivative corresponding to the value $k = 2$. Then a simple computation shows that:

$$H_1[\mu] \equiv \frac{A}{4}\left(\frac{2\pi k}{A}\right)^{2\mathbb{m}}$$

Now it is clear that as k increases the functions μ_k oscillate more and the smoothness functional gets larger, as desired. ⊠

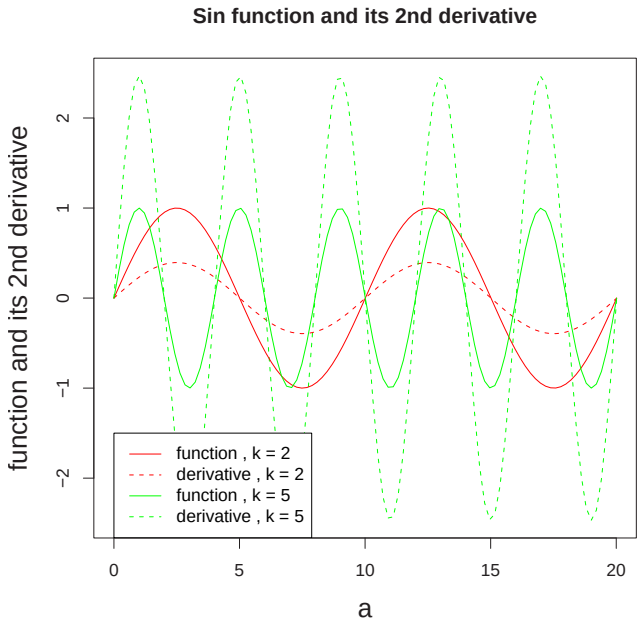

Figure 5.1: The sin function and its 2nd derivative, for different frequencies. On the vertical axis we have both the function $\mu_k(a) = \sin(\frac{2\pi k a}{A})$ and its 2nd derivative. Here $A = 20$ and k takes on the values 2 and 5.

For a given k the value of the functional is increasing with $\mathbb{m}$. Therefore large values of $\mathbb{m}$ correspond to functionals which are very restrictive, since in these cases

even small values of k can lead to a large value for the smoothness functional. This justifies calling $\mathtt{n}$ the degree of smoothness (other justifications for this terminology lie in spline theory and in some other important properties of the smoothness functional above but we will not discuss them here). For further information see Wahba (1990), DeBoor (1978*a*), Schumaker (1981), or Eubank (1988).

Since every function on $[0, A]$ can be written as a linear superposition of sine and cosine waves, this example turns out to be completely general, and shows that the functionals defined above are indeed *always* a measure of smoothness. We shall therefore use them and our method of deriving them quite generally, even, in Chapter 7, for priors defined over units that are not inherently continuous.

5.2.2 Varying the Degree of Smoothness over Age Groups

Before proceeding we point out that the smoothness functional in Equation 5.8 can be generalized in a way that can be very useful in practice. It is often the case that, while $\mu(a, t)$ is a smooth function of a, it can be smoother in certain regions of its domain than in others. For example, if $\mu(a, t)$ is the expected value of log-mortality from all causes we know that it will have a fairly sharp minimum at younger ages (less smooth), but it will be almost a straight line at older ages (more smooth). (For example, see Figure 2.1, Page 27.) Therefore, penalizing the lack of smoothness uniformly across age groups would misrepresent our prior knowledge: Younger ages should be penalized relatively less than older ages. This problem is easily fixed by replacing the Lebesgue measure da in the integral in Equation 5.8 with a more general measure $dw^{\text{age}}(a)$.[2] For example we may set $dw^{\text{age}}(a) = a^l da$ for some $l > 0$ in order to penalize older ages more.

Thus, we represent prior knowledge about smoothness of the expected value of the dependent variable over ages, with the additional information about where in the age profile different levels of smoothness will occur, as follows:

$$H_t[\mu] \equiv \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathtt{n}}\mu(a,t)}{da^{\mathtt{n}}} \right)^2 \quad \text{should be small } \forall t \in [0, T]. \tag{5.9}$$

However, enforcing this constraint above for every time t can be difficult or unrealistic, and therefore it may be preferable to have a slightly different formulation, where the functionals $H_t[\mu]$ are averaged over time according to a measure $dw^{\text{time}}(t)$. If the "small" in Equation 5.9 is the same for all times t the measure $dw^{\text{time}}(t)$ will be the uniform Lebesgue measure, otherwise it can be chosen to enforce the constraint more in certain years than in others. Therefore instead of Equation 5.9 we consider smoothing on average over time, and represent our prior knowledge as follows:

$$H[\mu, \theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathtt{n}}\mu(a,t)}{da^{\mathtt{n}}} \right)^2 \quad \text{should be small} \tag{5.10}$$

[2]The "Lebesgue measure" means that the integral is performing an average of a quantity over the uniform density.

where we have also added the fixed positive parameter θ, which controls how small the functional should be. It is important to notice that the integration interval $[0, T]$ can include future values as well as past ones, allowing one to impose prior knowledge on in and out of sample predictions. Obviously more complicated choices then Equation 5.10 can be made, and we will indeed discuss some of them in Section 6.1, but for the moment Equation 5.10 is sufficient to explain both the idea and the formalism.

5.2.3 Null Space and Prior Indifference

Putting aside for the moment technical issues involved in giving a precise meaning to a probability density defined over a function space, we define from Equation 5.10 a prior density over μ as follows:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{\theta}{2}\int_0^T dw^{\text{time}}(t)\int_0^A dw^{\text{age}}(a)\left(\frac{d^{\mathbb{n}}\mu(a,t)}{da^{\mathbb{n}}}\right)^2\right)$$

One reason for which such a prior is useful is that it is indifferent to a specific rich class of patterns of the expected value of the dependent variable. The key observation is that the derivative of order $\mathbb{n}$ is an operator whose null space is the set of polynomials of degree $\mathbb{n}-1$. To clarify, denote by $p_{\mathbb{n}}$ the set of polynomials in a of degree at most $\mathbb{n}-1$, that is the set of functions of the form:

$$f(a,t) = \sum_{k=0}^{\mathbb{n}-1} b_k(t)a^k.$$

These functions have the property that

$$\frac{d^{\mathbb{n}}}{da^{\mathbb{n}}}f(a,t) = 0\ , \quad \forall a,t \in \mathbb{R}$$

Therefore the prior above has the indifference property:

$$\mathcal{P}(\mu \mid \theta) = \mathcal{P}(\mu + f \mid \theta)\ , \quad \forall f \in p_{\mathbb{n}}$$

This implies that, at any point in time t, we have no preference between two functions that differ by a polynomial of degree $\mathbb{n}$ in age, or, in other words, we consider the two functions equiprobable. The polynomials we are indifferent to have coefficients that are arbitrary functions of time.

Example: $\mathbb{n} = 1$ Consider the simplest case, in which $\mathbb{n} = 1$. The first derivative is indifferent to any constant function, and therefore our notion of prior indifference here is expressed by saying that:

$$\mathcal{P}(\mu \mid \theta) = \mathcal{P}(\mu + f(t) \mid \theta)\ , \quad \text{for any function } f(t)$$

Therefore, while we know something about how the dependent variable μ varies from one age group to the next, we declare ourselves totally ignorant about the absolute levels it may take. ⊠

Example: $\mathtt{n} = 2$ The second derivative is indifferent to constant and linear functions, and therefore our version of prior indifference is expressed by saying that:

$$\mathcal{P}(\mu \mid \theta) = \mathcal{P}(\mu + f(t) + g(t)a \mid \theta) \ , \quad \text{for any function } f(t), g(t)$$

In this case we are indifference to a larger class of patterns then the one in Example 1: Not only do we have no preference over two age profiles that differ by a constant, we also we do not distinguish between age profiles differing by a linear function of age. Put differently, we declare ourselves ignorant of the mean and age-trend of the age profiles.⊠

In both of these examples we impose no constraints on the functions $f(t)$ and $g(t)$ which appear in the null space of the prior. In real applications, this is of course unrealistic, since although we are ignorant about the levels of the age profiles we expect them to move smoothly as a function of time. (We address this issue below by using another smoothness functional, which explicitly encourages the expected value of the dependent variable to vary smoothly over time.)

5.2.4 Nonzero Mean Smoothness Functional

As pointed out in Section 4.5.1, the functional in Equation 5.10 is symmetric around the origin: It assigns the same value, and therefore the same probability, to μ and $-\mu$, which may be undesirable. If a "typical" age profile $\bar{\mu}(a)$ is available it may be more appropriate to use the following smoothness functional instead:

$$H[\mu, \theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathtt{n}}}{da^{\mathtt{n}}} (\mu(a,t) - \bar{\mu}(a)) \right)^2 \tag{5.11}$$

This smoothness functional represents a different kind of prior information: The *deviation* of the dependent variable from the mean age profile varies smoothly across age groups.

This is an important distinction: It may happen that the age profiles themselves are not particularly smooth (for example there may be a huge variation in log-mortality from age group 0-4 to age group 5-9) and therefore it would not be appropriate to use the prior associated with Equation 5.10. However, we may still expect them to look like "smooth variations" of the typical age profile $\bar{\mu}$, and therefore the smoothness functional in Equation 5.11 may be more appropriate. Since using the smoothness functional in Equation 5.11 is equivalent to that in Equation 5.10 in which the dependent variable has been redefined as $\mu \rightsquigarrow \mu - \bar{\mu}$, we only use, unless otherwise noted, the simpler form 5.10 in the following. This implies that when we refer to the dependent variable as "log-mortality", we might also be referring to its deviation from $\bar{\mu}$, depending on whether we have set $\bar{\mu} = 0$ or not.

5.2.5 Discretizing: From Age to Age Groups

Now that we have a generic smoothness functional given by Equation 5.10 the next step is computational: both the age and time variable are discrete in practice, so that the function $\mu(a,t)$ should be replaced by an $A \times T$ matrix with elements μ_{at}, the $\mathtt{n}$-th derivative should also be replaced by a matrix, and the integral by a weighted sum. We develop discrete versions of the $\mathtt{n}$-th derivative in Appendix D; for the moment all we need to know is that this appendix provides well-defined matrices $\mathrm{D}^{\mathrm{age},\mathtt{n}}$ which approximate the derivative of order $\mathtt{n}$ with respect to age. Therefore we should make in Equation 5.10 the replacements:

$$\mu(a,t) \rightsquigarrow \mu_{at} \qquad \frac{d^{\mathtt{n}}\mu(a,t)}{da^{\mathtt{n}}} \rightsquigarrow \sum_{a'} \mathrm{D}^{\mathrm{age},\mathtt{n}}_{aa'}\mu_{a't} \qquad \int_T dw^{\mathrm{time}}(t) \int_A dw^{\mathrm{age}}(a) \rightsquigarrow \sum_{at} w^{\mathrm{time}}_t w^{\mathrm{age}}_a$$

where w^{time}_t and w^{age}_a are vectors of positive weights, summing up to 1, that correspond to the measures $dw^{\mathrm{time}}(t)$ and $dw^{\mathrm{age}}(a)$. In order to keep the notation simple we assume here that $dw^{\mathrm{time}}(t)$ and $dw^{\mathrm{age}}(a)$ are simply normalized Lebesgue (uniform) measures, and therefore we set $w^{\mathrm{time}}_t = T^{-1}$ and $w^{\mathrm{age}}_a = A^{-1}$. The smoothness functional above can now be redefined in its discretized form:

$$H[\mu,\theta] \equiv \frac{\theta}{TA} \sum_{at} \left(\sum_{a'} \mathrm{D}^{\mathrm{age},\mathtt{n}}_{aa'}\mu_{a't} \right)^2 .$$

Introducing the matrix $W^{\mathrm{age},\mathtt{n}} \equiv A^{-1}(\mathrm{D}^{\mathrm{age},\mathtt{n}})'\mathrm{D}^{\mathrm{age},\mathtt{n}}$ we rewrite the expression above in simpler form as

$$H[\mu,\theta] = \frac{\theta}{T} \sum_{aa't} W^{\mathrm{age},\mathtt{n}}_{aa'}\mu_{at}\mu_{a't} \equiv \frac{\theta}{T} \sum_{t} \mu'_t W^{\mathrm{age},\mathtt{n}}\mu_t \tag{5.12}$$

where μ_t is an $A \times 1$ vector whose elements are μ_{at}, also referred to as the time series age profile at time t. This implies that the prior for μ has the form:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{\theta}{2} \sum_{t} \mu'_t W^{\mathrm{age},\mathtt{n}}\mu_t \right). \tag{5.13}$$

5.2.6 Interpretation

We now further interpret the smoothness functional in Equation 5.12. First, we have seen in Section 5.2.3 that the prior associated with the smoothness functional in Equation 5.10 is indifferent to polynomials of degree $\mathtt{n}-1$ in age, with time-dependent coefficients. This important and useful property was derived in the continuous setting, and it also holds in the discretized setting if the derivative operator is discretized properly. In fact, any discretized form of the derivative of order $\mathtt{n}$ should have the property that $\mathrm{D}^{\mathrm{age},\mathtt{n}}\nu = 0$ for any vector ν of the form $\nu_a = a^k$, $k = 0, 1, \ldots, \mathtt{n}-1$

(and any linear combination of such vectors). This means that the matrix $\mathrm{D}^{\mathrm{age},\mathtt{n}}$ has nullity equal to $\mathtt{n}$ and rank equal to $A-\mathtt{n}$. Since the matrix $W^{\mathrm{age},\mathtt{n}}$ is proportional to $(\mathrm{D}^{\mathrm{age},\mathtt{n}})'\mathrm{D}^{\mathrm{age},\mathtt{n}}$, its eigenvalues are simply the squares of the singular values of $\mathrm{D}^{\mathrm{age},\mathtt{n}}$ (see Section B.2.4, Page 248). As a result, $W^{\mathrm{age},\mathtt{n}}$ has the same rank and nullity as $\mathrm{D}^{\mathrm{age},\mathtt{n}}$:

$$\mathrm{rank}(W^{\mathrm{age},\mathtt{n}}) = A - \mathtt{n} \ , \quad \mathrm{nullity}(W^{\mathrm{age},\mathtt{n}}) = \mathtt{n}.$$

Therefore the prior specified by Equation 5.13 is improper, since $W^{\mathrm{age},\mathtt{n}}$ is singular. The improperness comes from the fact that we do not want to commit ourselves to specify a preference over some properties of the age profiles, such as the mean (when $\mathtt{n} = 1$) or the mean and trend over ages (when $\mathtt{n} = 2$). However the prior, unlike improper flat priors, does represent some genuine knowledge: In fact, the prior is proper and informative once we restrict ourselves to the age profiles that lie in the subspace orthogonal to the null space.

Take for example $\mathtt{n} = 1$, so that the null space is the set of constant age profiles. The space of age profiles orthogonal to the null space is the space of age profiles with zero mean. In this space, the prior is proper, and we can, for example, draw samples from it. In general the prior in Equation 5.13 is proper once we restrict ourselves to age profiles whose moments of order up to $\mathtt{n}-1$ are zero (ensuring that they are orthogonal to the null space of the prior). (The technical details of how to draw from prior 5.13 and compute associated expected values are described in Appendix C.) Obviously once we have a sample from the prior we can add arbitrary elements of the null space and obtain random draws that have exactly the same probability as the original sample under the improper prior. Thus, one question is which samples should we show? We adopt the convention that when we sample from an improper prior, we only show the samples whose projection on the null space is 0 (because this is actually how the samples are obtained), and leave to our imagination the task of adding arbitrary elements of the null space in order to visualize the prior indifference. This is usually easy when the null space consist of constant or linear functions. However, in order to aid this process, before showing samples from different kind of priors, we now show how samples may look like when we add an arbitrary member of the null space.

Consider the cases $\mathtt{n} = 1$ and $\mathtt{n} = 2$, with $\bar{\mu} = 0$ (zero mean) and $\bar{\mu}$ set to some typical age profile (in this case the one for all-cause male log-mortality). For Figure 5.2 we drew three samples from the proper portion of the prior in Equation 5.13. Then we added to each an arbitrary element of the null space. Notice that we say an "arbitrary" and not a "random" element of the null space because we cannot draw at random from the null space since the density over it is improper. Hence, we selected the particular elements here for visual clarity. In the top left panel we have set $\mathtt{n} = 1$ and $\mu = 0$: The prior has zero mean and the null space is the space of constant functions. Each of the three random samples from the proper portion of the prior is color coded (red, green, or blue). We then repeat each of the three samples three

times by adding to the sample three arbitrary elements of the null space. Hence, in this graph, we can see three red curves, that differ only by a constant. Our prior is indifferent to the choice among these three red curves, in that they have identical prior probabilities. The same holds for the three green curves and three blue curves in the top left graph of the figure. (The samples originally produced by the algorithm which samples from the proper prior are the ones in the middle, which have zero mean over age groups, but this of course is a minor technical point about how we happen to choose to draw priors.)

In the top right panel we show a similar graph, but with a nonzero mean prior, for which we set $\bar{\mu}$ to some typical shape for the age profile of male all-cause log-mortality. In the bottom left panel we give samples from a zero-mean prior with $\mathtt{n} = 2$, whose null space consists of the space of linear functions. The original samples are again the ones in the middle (zero mean and zero trend). We have added constant positive and negative shifts and linear terms with positive and negative slopes to form the other two sets of three curves in this graph. The bottom right panel has been obtained in the same manner of the bottom left panel, but with a non-zero mean prior (same $\bar{\mu}$ as in the top right panel).

Of course, whenever we talk about null space and prior indifference we are always idealizing the situation somewhat: obviously it is not true that we are totally ignorant about the levels of the age profiles (for example we have the constraint that log-mortality is a negative number, and some of the values in the figure have positive numbers!). What we mean by "ignorant" is that we think that the prior knowledge is sufficiently less important than the knowledge contained in the data that it should be ignored. In this situation, we might as well pretend we do not have any such prior knowledge and take advantage of the nice properties of the null space of the prior.

We now proceed to analyze in more detail what samples from the improper priors described in this section look like, ignoring the null space. To this end we show in Figure 5.3, samples from the proper part of the prior in Equation 5.13 for $\mathtt{n} = 1, 2, 3, 4$.

An obvious feature of these graphs is that the samples become less and less "jagged" (or locally smooth) as $\mathtt{n}$ increases: This is what we should expect, since the smoothness functional is built to penalize an average measure of local "jaggedness". (The same pattern can also be seen in Figure 5.2, which we constructed to focus on the null space.) Another way to say this is that as $\mathtt{n}$ increases the values of μ at different ages become more and more correlated with each other.

Another very evident feature of these graphs is that, as $\mathtt{n}$ increases, the samples acquire more and more large "bumps" (or global changes in direction). If we think of the number of bumps as a measure of oscillation then this implies that as $\mathtt{n}$ increases the samples oscillate more. This sounds like a contradiction: the point of building smoothness functionals is to penalize functions which oscillate too much, and we have been claiming all along that as $\mathtt{n}$ increases, the functionals become more restrictive and therefore their samples should oscillate less. The contradiction is only apparent,

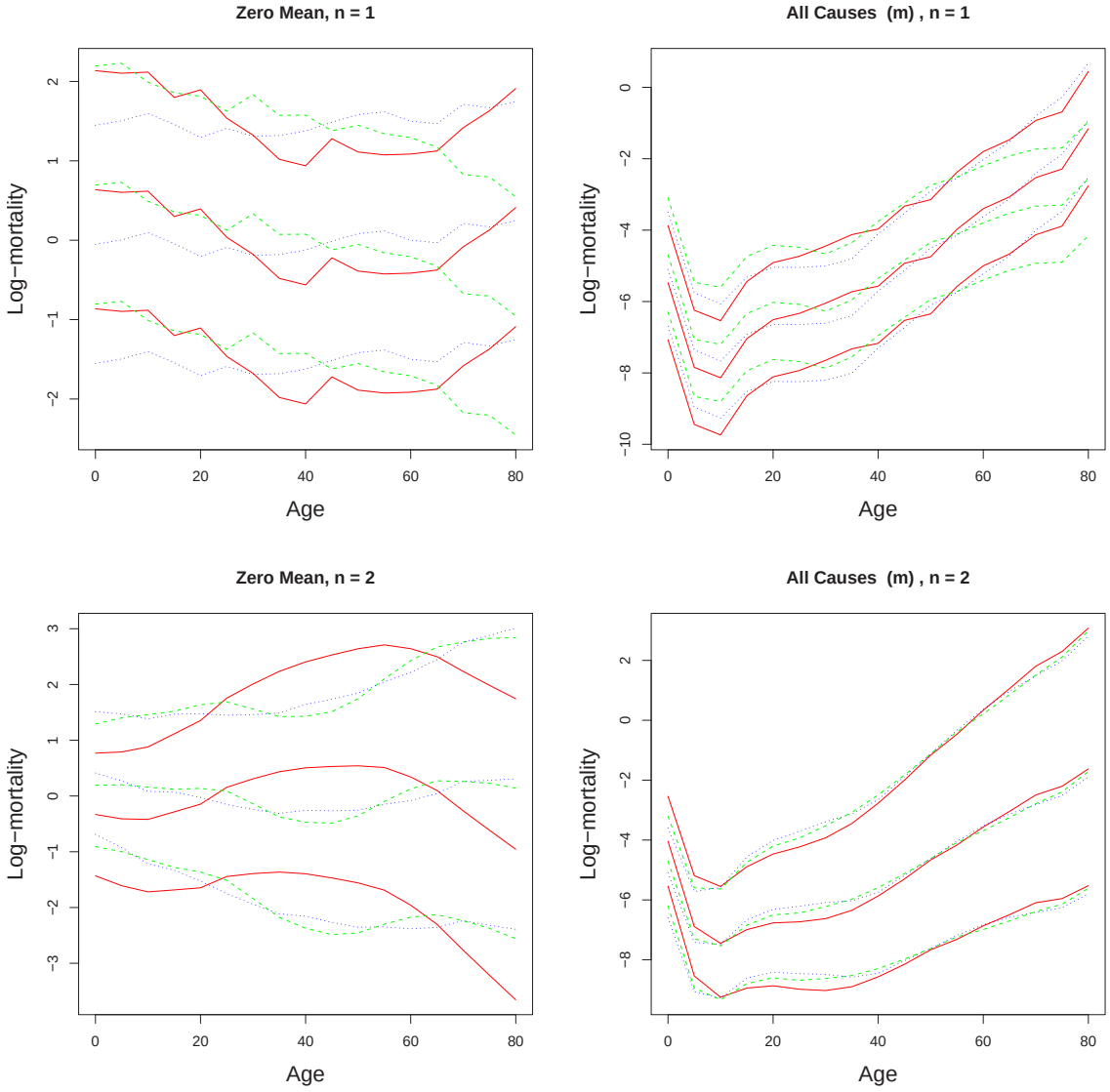

Figure 5.2: Age profile samples from smoothness priors with added arbitrary elements of the null space. For each panel, different colors correspond to different samples, while curves of the same color differ by an element of the null space. Top left: $\mathtt{n} = 1$ and $\bar{\mu} = 0$; Top Right: $\mathtt{n} = 1$ and $\bar{\mu} \neq 0$; Bottom left: $\mathtt{n} = 2$ and $\bar{\mu} = 0$; Bottom right: $\mathtt{n} = 2$ and $\bar{\mu} \neq 0$. These graphs have data with 17 age groups, at 5 years intervals, labeled $0, 5, \ldots, 80$. The value of θ has been chosen so that the standard deviation of μ_a is 0.3, on average over the age groups, and the scale is the same in all graphs.

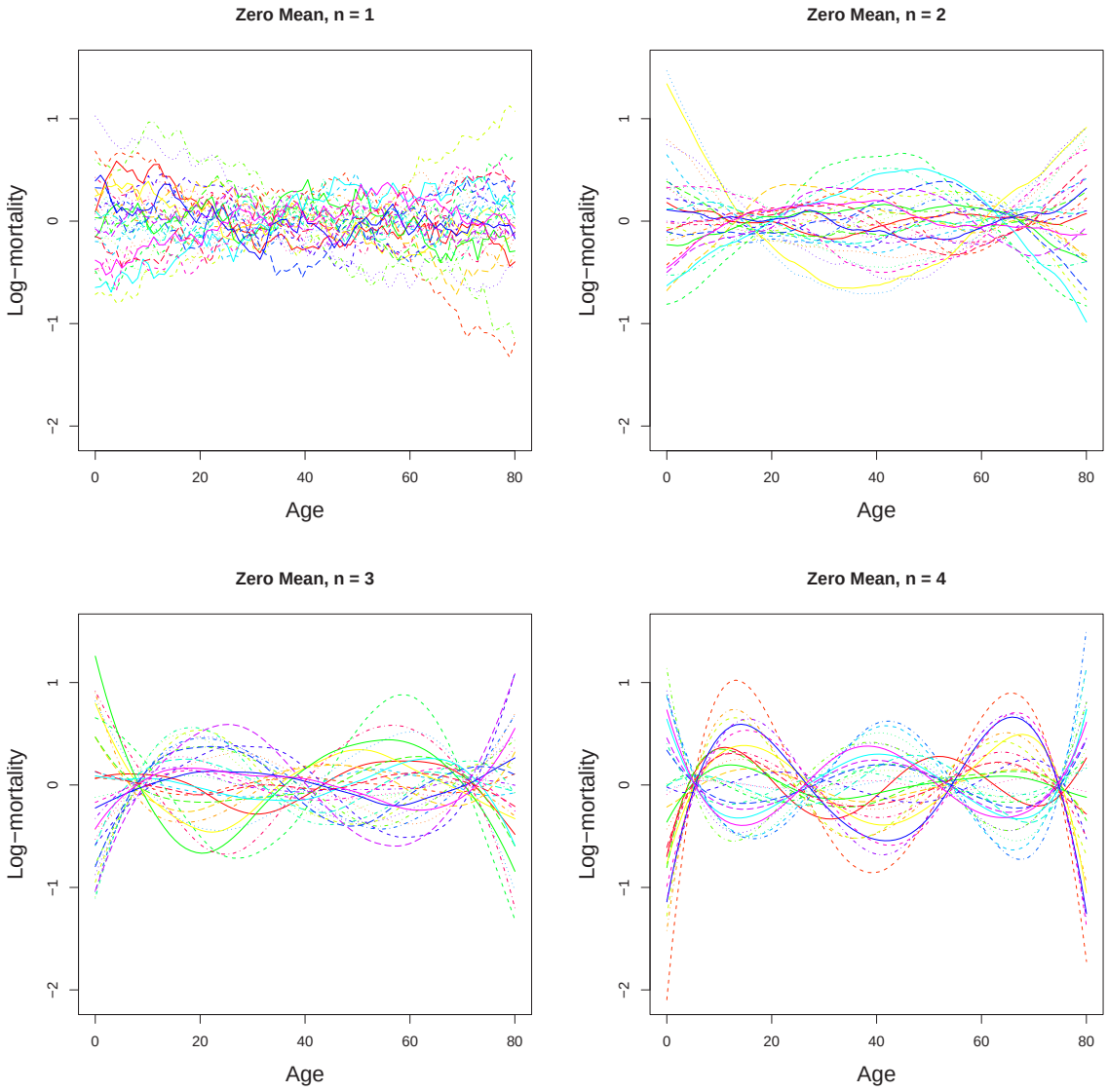

Figure 5.3: Age profile samples from the smoothness prior 5.13 for $\mathtt{n} = 1, 2, 3, 4$. Here $A = 80$, and there are 81 age groups, from 0 to 80. The value of θ has been chosen so that the standard deviation of μ_a is 0.3, on average over the age groups, and the scale is the same in all graphs.

however, because we have been mixing two kinds of oscillations: one is *local oscillation*, measured locally by the derivative of order $\mathtt{n}$, and the other is *global oscillation*, measured by the number of bumps, or, better, by the number of zero-crossings. The smoothness functional in Equation 5.10 is built to penalize the local amount of oscillation, on average, and it does not care about the global shape of a function. In fact, we can dramatically alter the global shape of a function by adding to it a polynomial of degree $\mathtt{n}-1$ without changing the value of the smoothness functional at all. Take for example $\mathtt{n}=4$, so that the null space is the 4-dimensional space of polynomials of degree 3. Polynomials of degree 3 can have 2 "bumps", but they are the smoothest possible curve according to this smoothness functional. Therefore, it should not be surprising that samples from the prior often have one more bump, as is the case for most of the samples in the bottom/right panel of Figure 5.3.[3]

The samples we show in Figure 5.3 are all for the zero-mean prior. In order to give an idea of how the samples look like when the prior is not zero mean, like in Equation 5.11, we repeated the same experiment centering the prior around $\bar{\mu}$, where $\bar{\mu}$ has been chosen as the average age profile of all-causes log-mortality in males (the average is over all years and all 67 countries with more than 20 observations). The results are reported in Figure 5.4. The most "reasonable" age profiles are those obtained with $\mathtt{n}=2$, for which the null space is the set of linear functions of age. If this null space is too large we can combine the priors for $\mathtt{n}=1$ and $\mathtt{n}=2$ in order to reduce the size of the null space but retain smooth age profiles. We address this issue in more detail in Section 6.1.

Finally, we compare the smoothness functionals in Equation 5.12 derived in this section with the "bare bones" smoothness functional in Equation 4.15 (Page 80). Since $\mathtt{n}\geq 1$, the constant vector $\nu=(1,1,\ldots,1)$ is always in the null space of $W^{\text{age},\mathtt{n}}$, implying that the rows and columns of $W^{\text{age},\mathtt{n}}$ always sum to 0. In turn, this implies (via the result in Appendix B.2.6, Page 253) that it is always possible to find a matrix $s^{\text{age},\mathtt{n}}$ such that we can write the smoothness functional in Equation 5.12 (Page 100) in the same form as Equation 4.15 (Page 80):

$$H[\mu,\theta]=\frac{\theta}{T}\sum_{t}\sum_{aa'}s^{\text{age},\mathtt{n}}_{aa'}(\mu_{at}-\mu_{a't})^2. \tag{5.14}$$

Since the derivative is a local operator, the matrix $W^{\text{age},\mathtt{n}}$ will usually have a "band" structure, such that $W^{\text{age},\mathtt{n}}_{aa'}$ is different from 0 only if a and a' are "close" to each other (although not necessarily first neighbors). This structure is reflected in the matrix $s^{\text{age},\mathtt{n}}$, which makes clear that the smoothness functional in Equation 5.12 is a sum of "local" contributions, obtained by comparing the value of μ in a certain age group with the values in nearby age groups. In this respect the smoothness

[3] Another noticeable feature of these graphs is that the variance of the samples for the first and last age group becomes larger with $\mathtt{n}$. This is partly due to the difficult of writing a good discretization of the derivative operator near the edges of the domain (for age group 0 and 80 only "one-sided" information can be used), and it is sensitive to choices we make in this regard.

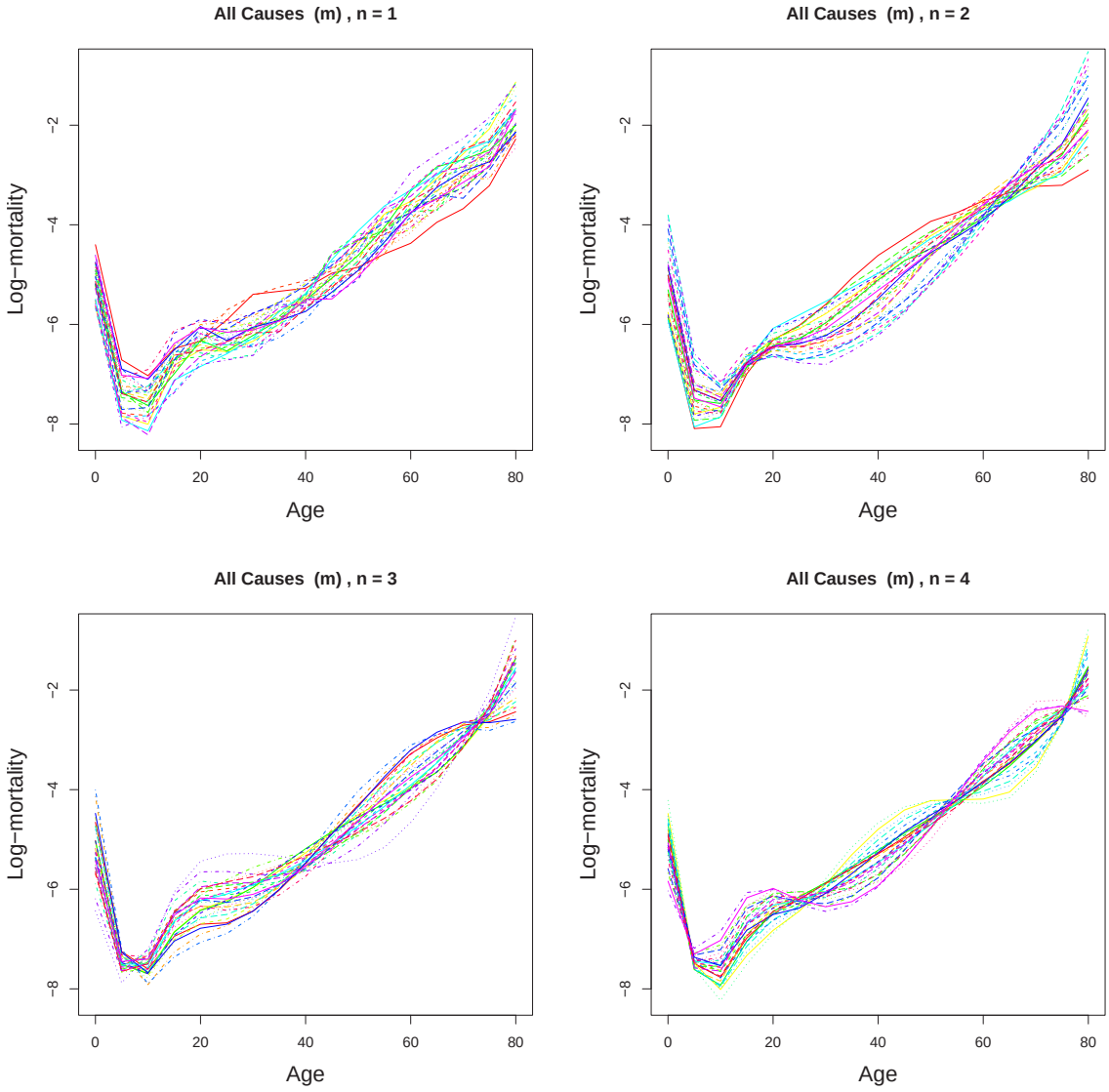

Figure 5.4: Age profile samples from the smoothness prior 5.13 for $\mathtt{n} = 1, 2, 3, 4$ and a typical age profile for all-causes log-mortality in males. There are 17 age groups ($A = 17$), at 5 years intervals, labeled $0, 5, \ldots, 80$. The value of θ has been chosen so that the standard deviation of μ_a is 0.3, on average over the age groups, and the scale is the same in all graphs.

functional in Equation 5.12 is like the one in the previous chapter, which resulted from pairwise comparisons between neighboring age groups. An important difference between Equations 5.14 and 4.15, however, is that in Equation 4.15 the "weights" $s_{aa'}$ were chosen to be positive: This allows us to interpret the smoothness functional as a way to penalize configurations in which similar age groups do not have similar values of μ. However, as soon as $\mathtt{n}$ becomes larger than 1, many of the elements of $s^{\text{age},\mathtt{n}}$ become negative, which may appear counterintutitive. However, this must be the case once we realize that if the elements of $s^{\text{age},\mathtt{n}}$ were all positive then the null space of the functional could only be the set of constants, independent of the values of $s^{\text{age},\mathtt{n}}$, while we know that the size of the null space increases with $\mathtt{n}$. In other words, the elements of $s^{\text{age},\mathtt{n}}$ become negative in order for some cancellations to occur, cancellations necessary to ensure that the null space has the correct structure.

Thus, it may be tempting to build priors "by hand" starting from the intuitive formula in Equation 5.14, where the elements of $s^{\text{age},\mathtt{n}}$ are chosen to be positive, since it easy to understand its meaning in this case. In some cases this is appropriate, and we shall do so when we will consider smoothness functionals over discrete variables, such as countries, in Chapter 7. In other cases, however, following the approach of Equation 5.14 would probably cause us to miss a richer class of smoothness functionals, and so it is more appropriate to start from the more formal notions of smoothness we offer here, such as the one expressed by Equation 5.10. Such an approach has a tremendous practical advantage in that we do not have to choose the elements of the matrix $s^{\text{age},\mathtt{n}}$: They are provided to us from the discretization of the derivative operator, so that the only choice we have to make is about the parameter θ and the degree of smoothness $\mathtt{n}$.

5.3 Step 2: From the Prior on μ to the Prior on $\boldsymbol{\beta}$

5.3.1 Analysis

Now that we have a better understanding of the meaning of the prior on μ in Equation 5.12, we proceed to Step 2 of our strategy and derive a meaningful prior in terms of $\boldsymbol{\beta}$ by using our prior for μ constrained to fit the specification $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$. One way to think about this procedure is as another way to add information, by restricting ourselves to patterns for the expected value of the dependent variable that can be explained by a set of covariates. Formally this is done by projecting the prior implied by Equation 5.12 on the subspace spanned by the covariates. Substituting $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$ into Equation 5.12 we obtain:

$$\begin{aligned} H^{\mu}[\boldsymbol{\beta},\theta] &\equiv \frac{\theta}{T}\sum_{aa't} W^{\text{age},\mathtt{n}}_{aa'}(\mathbf{Z}_{at}\boldsymbol{\beta}_a)(\mathbf{Z}_{a't}\boldsymbol{\beta}_{a'}) \\ &= \theta\sum_{aa'} W^{\text{age},\mathtt{n}}_{aa'}\boldsymbol{\beta}'_a\mathbf{C}_{aa'}\boldsymbol{\beta}_{a'} \end{aligned} \tag{5.15}$$

where the second line uses the fact that the coefficients $\boldsymbol{\beta}$ do not depend on time and so the sum over time can be performed once for all, and where we have defined the matrix:

$$\mathbf{C}_{aa'} \equiv \frac{1}{T}\mathbf{Z}_a'\mathbf{Z}_{a'}$$

so that $\mathbf{Z}_a$ is the usual data matrix of the covariates in cross-section a, which has $\mathbf{Z}_{at}$ for each row. Hence, the prior for $\boldsymbol{\beta}$, conditional on the parameter θ, is now simply

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'}^{\text{age},\mathbb{m}} \boldsymbol{\beta}_a' \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}\right) \tag{5.16}$$

5.3.2 Interpretation

We now make three brief but critical observations. First, the vectors of covariates $\mathbf{Z}_{at}$ and $\mathbf{Z}_{a't}$ are of dimensions k_a and $k_{a'}$, respectively, and so $\mathbf{C}_{aa'}$ is a rectangular $k_a \times k_{a'}$ matrix, and it does not matter whether we have same number or type of covariates in the two cross-sections.[4] That is, this result enables us to include all available covariates in the time series regression in each cross-section, even if they differ from cross-section to cross-section in number, content, or meaning.

Second, the weights $W_{aa'}^{\text{age},\mathbb{m}}$ in Equation 5.16 are fully specified once we choose, from prior information, the order $\mathbb{m}$ of the smoothness functional in Equation 5.10 (see Section 6.1). That is, all A^2 elements of this matrix — all elements of which, under previous approaches, would need to be specified by hand — are uniquely determined by the single scalar $\mathbb{m}$.

Third, the form of the prior in Equation 5.16 depends on the fact that the cross-sectional index can be thought of as a (possibly discretized) continuous variable, so that we can define the fundamental notion of smoothness with respect to a continuous variable in terms of derivatives. We show in Section 7.2 that when the cross-sectional index is a label, like a country name, a formally similar approach is viable and leads to a prior of the same form as the one in this section.

[4] This last statement is true even if the age groups indexed by a are not equally spaced. In this case it is somewhat more complicated to build the matrix $W_{aa'}^{\text{age},\mathbb{m}}$, since one is required to approximate the $\mathbb{m}$-derivative of μ using unequally spaced points: This task goes beyond the simple rules explained in Appendix D, but straightforward methods can be found in standard numerical analysis textbooks.

Chapter 6

Model Selection

Like any statistical method, and especially any Bayesian method, our approach comes with a variety of adjustable settings and choices. As discussed in Chapter 1, however, we needed to make so many forecasts for our application that we had to limit these choices as much as possible. As it turned out, limiting choices is almost the same process as ensuring that the choices made were based on readily available knowledge. This is often not the case in Bayesian modeling, where hyperparameter values and other choices are based on guesses, default settings, "reference priors," or trial and error.

Thus, in this chapter we try to connect every adjustable setting to some specific piece of knowledge that demographers others have readily available from their empirical analyses. For example, at no point in using our methods do users need to set the value of some hyperparameter that has no obvious meaning or connection to empirical reality. Our job in developing these methods, as we see it, is to bring new information to bear on the problem of demographic forecasting, and so we put in considerable effort into taking existing qualitative and quantitative knowledge bases in demography and the related social sciences and providing easy and direct connections to the choices necessary in using our methods.

In this chapter, we discuss choices involving the degree of smoothness (Section 6.1), the prior for the smoothing parameter (Section 6.2), where in the function to smooth (Section 6.3), covariate specification (Section 6.4), and variance function specification (Section 6.5). The results of this chapter made it possible for us to design easy-to-use software that implements our methods, since we were able to translate apparently arcane choices into substantively meaningful decisions about which much is known.

6.1 Choosing the Smoothness Functional

In Chapter 5, we considered a family of smoothness functionals based on the derivative of order $\mathtt{m}$. The parameter $\mathtt{m}$ plays multiple roles in the smoothness functionals of

the type of Equation 5.10:

1. It determines the *local* behavior of the function, that is how the functions looks on a small interval of the domain. Increasing values of $\mathtt{m}$ correspond to samples that are locally more and more smooth.

2. It determines the size of the null space of the functional, which consists of the $\mathtt{m}$-dimensional space of polynomials of degree $\mathtt{m} - 1$.

3. It also determines the *global* shape of the samples, that is how many global "bumps" (or changes of direction) it has, and also how many zeros. This is a side effect of controlling the size of the null space: When $\mathtt{m}$ increases, the polynomials in the null space have more and more bumps. Therefore we can have functions with many bumps, similar to polynomials, with very small values of the smoothness functional: these functions will appear with high probability, which explains why samples from the prior with large $\mathtt{m}$ display, over the whole domain, a high degree of oscillation even if they are locally very smooth.

Reducing an entire proximity matrix to one parameter $\mathtt{m}$ is tremendously convenient, but the fact that only one parameter controls different characteristics of the samples drawn can sometimes be a disadvantage in practical applications. We show how to avoid this disadvantage now. Suppose, for example, that we wish our samples to be locally very smooth, with the kind of local smoothness associated with $\mathtt{m} = 4$. However, we may not want the global behavior associated with $\mathtt{m} = 4$ (see Figure 5.3, Page 104), because it has many bumps, and we may not want as a null space the space of polynomials of degree three, because it is too large (i.e., it may exclude constrants we wish to impose). Suppose instead we want the null space to consist of the space of linear functions. As it turns out, we can have the best of both worlds by considering a larger class of smoothness functionals that includes mixtures of those we have considered until now:

$$H[\mu, \theta] \equiv \sum_{i=1}^{K} \theta_i \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathtt{m}_i} \mu(a,t)}{da^{\mathtt{m}_i}} \right)^2, \tag{6.1}$$

where $\theta_i \geq 0$ and we use the convention that the numbers $\mathtt{m}_i$ are listed in ascending order. We refer to smoothness functionals of this form as *mixed smoothness functionals*, while we refer to a smoothness functional of the form 5.10 as a *standard smoothness functional*.

The reason mixed smoothness functionals are useful is that they enable separate control over the size of the null space and the degree of local smoothness. The size of the null space is controlled by $\mathtt{m}_1$, the lowest order of derivative in the functional, since in order for the smoothness functional to be zero all the terms of the sum in Equation 6.1 must be 0. The degree of local smoothness is controlled by $\mathtt{m}_K$, the highest order of derivative in the functional. In order for the smoothness functional to have small

values, all the individual smoothness functionals in the sum must assume small values, and if the term with $\mathtt{m}_K$ does not assign a small value the smoothness functional will not assume a small value. This makes clear that what is really important in the mixed smoothness functional is the choice of $\mathtt{m}_1$ and $\mathtt{m}_K$, which suggests that we can probably limit ourselves in most applications to the case $K = 2$.

Example We now return to the example at the beginning of this Section where we desire a smoothness functional with samples that look locally very smooth, do not have many global bumps, and have a null space consisting of the set of linear functions. The local smoothness could be obtained from a standard smoothness functional with $\mathtt{m} = 4$, but this will have a null space that is too large and will also have samples with many bumps. If we use a standard smoothness functional with $\mathtt{m} = 2$ we get the right null space, but the samples might not be smooth enough for our purposes.

Therefore we use a mixed smoothness functional with $K = 2$. The fact that the null space must be the set of linear functions immediately determines that $\mathtt{m}_1 = 2$. Since we want samples which look locally very smooth we choose $\mathtt{m}_2 = 4$. The last thing we need to choose is the size of the parameters θ_1 and θ_2. For the purpose of this illustration, all that matters is the relative size of these two numbers, so we fix θ_1 to an arbitrary number, say 1. Obviously if we want the samples to look very smooth we should give much more importance to the prior with the highest derivative. Figure 6.1 gives samples from the "mixed" prior for $\theta_1 = 1$ and for 4 different values of θ_2: 0, 1, 100, 1000.

Notice how increasing the value of θ_2 leaves the "global" shape of the samples and the number of bumps unchanged (since they do not depend on the part of the prior with higher derivative), while the local smoothness steadily increases.⊠

Thus, in a mixed smoothness functional, what determines the qualitative behavior of the samples from the prior are the lowest and highest degrees of the derivative, $\mathtt{m}_1$ and $\mathtt{m}_K$. Thus, in practice we usually limit ourselves to $K = 2$. Although results are sensitive to the choice of the prior, it is unlikely that they are that sensitive: If we also added to Figure 6.1 a smoothness functional with $\mathtt{m} = 3$, we would not see major changes in the samples from the prior.

In practice many users will never need a mixed smoothness functional, and in fact a standard smoothness functional with $\mathtt{m} = 2$ will give reasonable results in most cases. This is consistent with experience from the literature on non-parametric regression, where cubic splines, or thin plate splines (which are related to our prior with $\mathtt{m} = 2$) are used most of times. The point of this section is that if we need something more sophisticated, it is readily available, and no additional concepts are required. From a computational point of view, since mixed functionals are sums of standard functionals, and since a linear combination of quadratic forms is also a quadratic form, the priors determined by both standard and mixed smoothness all have exactly the same general mathematical form as Equation 5.12.

It is true that mixed smoothness functionals have more parameters than stan-

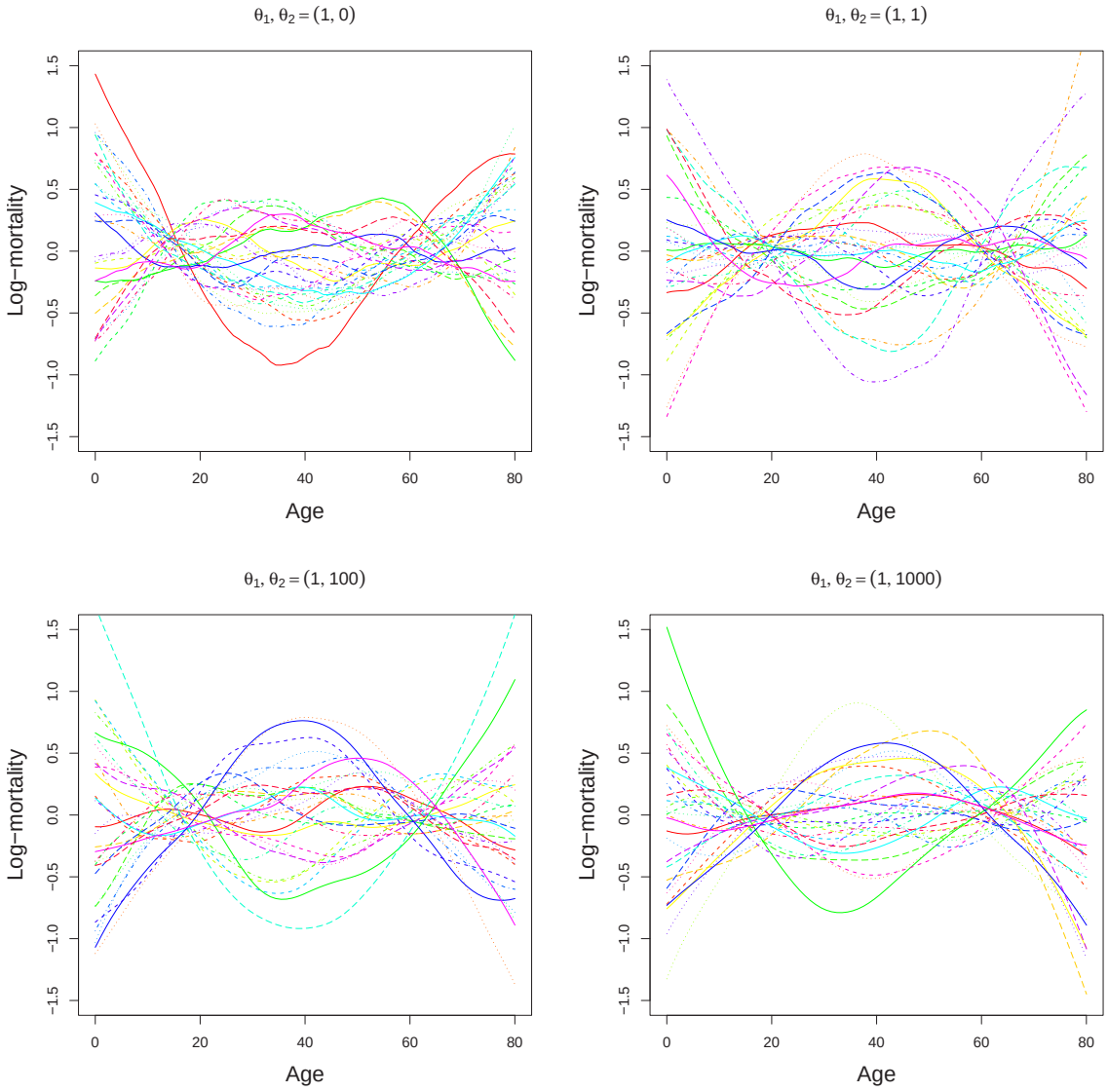

Figure 6.1: Age profile samples from "mixed" smoothness priors with $K = 2$, for different values of θ_2. Here $A = 80$, and there are 81 age groups, from 0 to 80. The graphs are on the same scale. In order to make the graphs comparable the values of θ_1 and θ_2 have been scaled, in each graph, by a common factor, so that the standard deviation of μ_a is 0.3, on average over the age groups.

dard smoothness functionals, but the relative size between the parameters can be determined in advance and kept fixed, leaving only one global hyperparameter. For example, with $K = 2$, we suggest parameterizing the prior as follows:

$$H[\mu, \theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left[\left(\frac{d^{\mathfrak{m}_1} \mu(a,t)}{da^{\mathfrak{m}_1}} \right)^2 + \lambda \left(\frac{d^{\mathfrak{m}_2} \mu(a,t)}{da^{\mathfrak{m}_2}} \right)^2 \right]$$

where the parameter λ is easily chosen by drawing from the mixed prior and selecting the value which leads to the "best" looking samples (that is by making graphs like the ones in Figure 6.1, a step our software automates; see Appendix F). With this parametrization everything we say about the smoothing parameter for standard smoothness functionals in the next Section also applies to θ in the mixed functional above.

6.2 Choosing a Prior for the Smoothing Parameter

All priors introduced thus far come with a smoothness parameter, which we denote by θ. If we assume θ is known, then its effect on the forecast is qualitatively clear: Larger values of θ correspond to smoother predictions, which pay correspondingly less attention to the data. In the limit, with θ going to infinity, our Bayesian estimate leads to a projection which lies in the null space of the prior — the specific element of which is chosen by the likelihood — since the only configuration of coefficients with nonzero probability are those for which the smoothness functional is zero. (This contrasts with proper priors which return a single point, such as the prior mean, when θ goes to infinity.) If we treat θ as a random variable then this point is still valid when we replace θ with its expected value.

The parameter θ can be viewed in two ways, each leading to a different type of procedure for choosing it. The first is to consider θ as a free parameter of the theory, for which we happen to have no direct information. In this situation, we could use relatively automated algorithms to choose θ's optimal value. In our case the optimal value is the one that minimizes an estimate of the forecast error. Usually these algorithms rely on the idea of cross-validation: one or more data points are left out of the data set and used as a "test set" to estimate the accuracy of the forecast on new, unseen data. By repeating this procedure many times over different definitions of the test set, one can construct reasonable estimates of the forecast error, and choose the value of θ that minimizes it. One method based on this idea is Generalized Cross-Validation, pioneered by Grace Wahba and her associates (See especially Golub, Heath and Wahba, 1979, and Wahba, 1980, as well as further discussion in the more recent monograph, Wahba, 1990). Another set of techniques based on a similar idea goes under the generic name of "bootstrapping" (see Efron, 1979, 1982, and the lengthy review in Efron and Tibshirani, 1993). A third approach to the choice of the optimal smoothness parameter is Structural Risk Minimization, which is a very

general approach to statistical inference and model selection (Vapnik, 1998; Hastie, Tibshirani and Friedman, 2001).

A second way to look at the smoothness parameter is to consider it at the same level of other quantities, such as $\boldsymbol{\beta}$ and σ, and treat it as a random variable with its own *proper* prior distribution $\mathcal{P}(\theta)$. (However, unlike $\boldsymbol{\beta}$ and σ, the prior for θ must be proper since the likelihood contains no information about it.) This implies that we must have an idea of what the mean and the variance of $\mathcal{P}(\theta)$ should be. While such information is often not available in many applications where smoothness functionals are typically used, as in pattern recognition, it *is* usually available for the applications described in this book. The main observation is that, although demographers do not have direct prior knowledge about θ in those terms, they typically do have knowledge about quantities determined by θ. Therefore, if the relationship between these quantities and θ can be inverted, knowledge about these quantities translates into knowledge about θ.

We formalize this idea in two stages. First we consider a non-parametric prior for the age profiles, of the type discussed in the previous chapters, ignoring the covariates and the time dimension. Then we introduce covariates and show how the non-parametric approach can be modified in order to be used in practical applications.

6.2.1 Smoothness Parameter for a Non-Parametric Prior

In this section we disregard our linear specification and the time dimensions, and simply the following smoothness prior for the age profiles:

$$\mathcal{P}(\mu) \propto \exp\left(-\frac{1}{2}\theta(\mu-\bar{\mu})'W^{\text{age},2}(\mu-\bar{\mu})\right), \tag{6.2}$$

where μ is an $A \times 1$ age profile vector, $W^{\text{age},2}$ is the matrix which corresponds to a smoothness functional of the form in Equation 5.8 (Page 95), with a derivative of order $\mathbb{m} = 2$, and $\bar{\mu}$ is a "typical" age profile for other infectious diseases in males. The question we address here is: what is a reasonable value for θ? The answer to this question is, simply put, "a value which produces reasonable sample age profiles". An example of reasonable and unreasonable age profiles is shown in Figure 6.2, where we plot two sets of twenty age profiles, each sampled from the prior 6.2. The only difference between the left and right graphs (other than the randomness of the sampling process), is the value of the smoothness parameter θ, which is larger in the right than left graph. Even to the untrained eye, the graph on the right would seem to correspond to an unlikely choice for the smoothness parameter, since each age group, except maybe age group 70, exhibits an unacceptably large variance. Those who study mortality age profiles typically have exactly this type of information. Therefore if we had to choose between the two figures we could easily choose the one on the left. The fact that we are not indifferent between these two figures shows that we are not indifferent to different values of θ, and therefore that prior knowledge about θ exists, and could be used to define at least a range in which θ should lie.

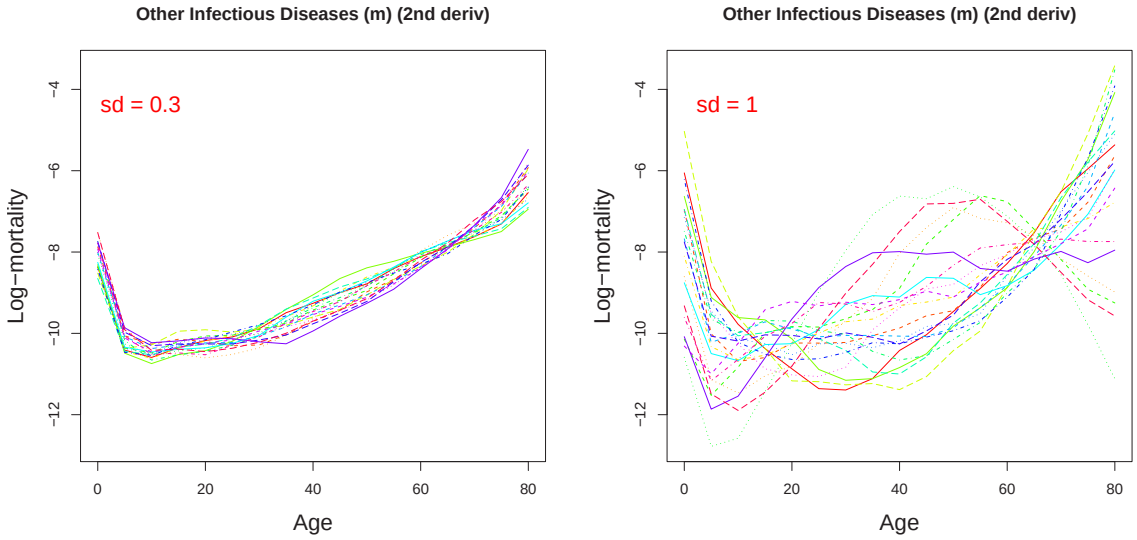

Figure 6.2: Samples from the prior in Equation 6.2 corresponding to different values of θ. In the text, we show how the standard deviation ("sd" in the figure) of log-mortality across samples at each age determines θ.

In principle, we could apply this strategy as follows: For many values of θ, from 0 to $+\infty$, sample from the prior and draw graphs like the ones in Figure 6.2. Then let subject matter experts choose a prior representative of their prior knowledge. The range of smoothness parameters θ associated with the selected figures will give us a range of acceptable values of θ, which we use to build a reasonable probability distribution $\mathcal{P}(\theta)$.

Of course, this approach in practice is too time consuming, and fortunately easier ways to achieve the same result exist. As it turns out, the only fact we need to know is the basis of the experts judgment about the samples. For example, we may postulate that experts judge the different figures according to the overall degree of variability of the age profiles, measured by the average of the age-specific variance of the age profiles. Alternatively, it is possible that experts are confident that, in a certain age range, say 20 to 75, log-mortality does not increase more (or less) than a certain amount going from one age group to the next, and therefore judge the age profiles accordingly.

We now formalize this approach. We assume that experts' opinions can be summarized by a statement of the following form: "on average, the value of $F(\mu)$ is $\bar{F}$", where $F(\mu)$ is the function of the age profiles that experts implicitly use as basis for their judgment. For example, if experts judge samples from the priors by the overall degree of variability of the age profiles, we could set $F(\mu)$ to be the average (over age groups) of the age-specific variance of the prior:

$$F(\mu) \equiv \frac{1}{A}\sum_{a=1}^{A}(\mu_a - \bar{\mu}_a)^2, \tag{6.3}$$

The expected value of $F(\mu)$ is clearly a function of the parameter θ, and therefore setting the expected value $F(\mu)$ to the experts' value $\bar{F}$ uniquely determines θ, by the following equation:

$$\mathrm{E}_{\perp}[F(\mu)|\theta] = \bar{F} \tag{6.4}$$

where the subscript $\perp$ reminds us that the prior in Equation 6.2 is improper and all expected values must be taken with respect to the subspace orthogonal to the null space of the prior (see Appendix C). Equation 6.4 can be solved for θ either numerically or analytically, depending on the choice of F, and therefore used to set the mean of the prior $\mathcal{P}(\theta)$. Since the value $\bar{F}$ will always be provided with an uncertainty interval, the uncertainty on $\bar{F}$ can be easily translated in uncertainty on θ, and therefore used to estimate the variance of $\mathcal{P}(\theta)$.

The relationship between $\bar{F}$ and θ is easily seen for the choice of $F(\mu)$ shown in Equation 6.3, which corresponds to the average of the age-specific variance of the prior. Experts are more likely to think in terms of standard deviations, rather than variance, and therefore it is convenient to define $\bar{F}$ in this case as σ_{age}, and refer to σ_{age} as the *standard deviation of the prior*. Using this definition, and using the formulas of Appendix C to perform the calculation in Equation 6.4, we obtain the following:

$$\mathrm{E}_{\perp}[F(\mu)|\theta] \equiv \frac{1}{A}\sum_{a=1}^{A}\mathrm{E}_{\perp}[(\mu_a - \bar{\mu}_a)^2] = \frac{1}{A\theta}\mathrm{tr}\left(W^{\text{age},2}\right)^{+} = \bar{F} = \sigma^2_{\text{age}} \tag{6.5}$$

where the superscript $+$ stands for the generalized inverse (Appendix B.2.5, Page 251), and the matrix $W^{\text{age},2}$ is known (see Section 5.2.5). Equation 6.5 can now be solved for θ as:

$$\theta = \frac{\mathrm{tr}\left(W^{\text{age},2}\right)^{+}}{A\sigma^2_{\text{age}}} \tag{6.6}$$

In fact, this is the expression we used to produce Figure 6.2: The graph on the left used a value of $\sigma_{\text{age}} = 0.3$, which seems, a priori, an empirically reasonable number (this is the quantity "sd" reported in the top left corner of the figure). Plugging this number in the equation above we obtained a value of θ to use in our simulation. Similarly, for the graph on the right, we choose θ such that $\sigma_{\text{age}} = 1$, which based on our experience with age profiles seems unrealistically large.

6.2.2 Smoothness Parameter for the Prior over the Coefficients

The formulas reported in the previous section are not what we use for forecasting (although they would be appropriate for simple univariate smoothing). In fact, in practical applications our age profiles can only lie in the span of the covariates, and the priors we use are defined over the set of coefficients $\boldsymbol{\beta}$. In order to simplify some of the formulas, it is convenient to re-write our linear specification as $\mu = \bar{\mu} + \mathbf{Z}\boldsymbol{\beta}$, where we have defined:

$$\boldsymbol{\beta} \equiv \begin{vmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_A \end{vmatrix} , \quad \mathbf{Z} \equiv \begin{vmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{Z}_A \end{vmatrix} , \quad \tilde{\mu}_a \equiv \mathbf{Z}_a \boldsymbol{\beta}_a , \quad \tilde{\mu} \equiv \mathbf{Z}\boldsymbol{\beta} = \begin{vmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_A \end{vmatrix} \tag{6.7}$$

In other words, in this section the specification $\mathbf{Z}\boldsymbol{\beta}$ refers to the *centered* age profiles, $\tilde{\mu} = \mu - \bar{\mu}$. Under this definition the prior over the coefficients $\boldsymbol{\beta}$ corresponding to the non-parametric prior in Equation 6.2 has zero mean and can be written as:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'}^{\text{age},2} \boldsymbol{\beta}_a' \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'} \right) \tag{6.8}$$

Introducing the matrix D_{age}:

$$D_{\text{age}} \equiv \begin{vmatrix} W_{1,1}^{\text{age}}\mathbf{C}_{1,1} & W_{1,2}^{\text{age}}\mathbf{C}_{1,2} & \dots & W_{1,A}^{\text{age}}\mathbf{C}_{1,A} \\ W_{2,1}^{\text{age}}\mathbf{C}_{2,1} & W_{2,2}^{\text{age}}\mathbf{C}_{2,2} & \dots & W_{2,A}^{\text{age}}\mathbf{C}_{2,A} \\ \dots & \dots & \dots & \dots \\ W_{A,1}^{\text{age}}\mathbf{C}_{A,1} & W_{A,2}^{\text{age}}\mathbf{C}_{A,2} & \dots & W_{A,A}^{\text{age}}\mathbf{C}_{A,A} \end{vmatrix} \tag{6.9}$$

the prior for the coefficients takes the form:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \boldsymbol{\beta}' D_{\text{age}} \boldsymbol{\beta} \right) \tag{6.10}$$

For a given set of covariates $\mathbf{Z}$, a sample of log-mortality age profiles is obtained by sampling the prior in Equation 6.10, obtaining a random set of coefficients $\boldsymbol{\beta}$ and plugging $\boldsymbol{\beta}$ in the specification $\mu = \bar{\mu} + \mathbf{Z}\boldsymbol{\beta}$. Notice that this procedure generates T age profiles, which are linked over the time-dimensions by the time variation of the covariates $\mathbf{Z}$.

The discussion of section 6.2.1 still applies, except for the fact that the functions $F(\mu)$ will contain an average over time, in addition to the average over age groups. For example, if we think that experts have knowledge about the standard deviation of the prior, we should set:

$$F(\mu) \equiv \frac{1}{AT} \sum_{a=1}^{A} \sum_{t=1}^{T} (\mu_{at} - \bar{\mu}_a)^2, \tag{6.11}$$

where $\mu = \bar{\mu} + \mathbf{Z}\beta$. Equation 6.4 is therefore replaced by:

$$\mathrm{E}_{\perp}[F(\mu)|\theta] = \mathrm{E}_{\perp}[F(\bar{\mu} + \mathbf{Z}\beta)|\theta] = \bar{F} \tag{6.12}$$

where the expected value is now taken over $\boldsymbol{\beta}$ using the prior 6.8. In order to see how this can be used in practice we explicitly perform the calculation in Equation 6.12 with $F(\mu)$ given by Equation 6.11. Using the formulas of Appendix C we show that

$$\mathrm{E}_{\perp}[F(\mu)](\theta) \equiv= \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T}\mathrm{E}_{\perp}[(\mathbf{Z}_{at}\boldsymbol{\beta}_a)^2] = \frac{1}{AT}\mathrm{E}_{\perp}[\boldsymbol{\beta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\beta}] = \frac{1}{AT\theta}\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right)$$

Therefore Equation 6.12 now reads:

$$\frac{1}{AT\theta}\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right) = \bar{F} = \sigma^2_{\text{age}} \tag{6.13}$$

and solving for θ we obtain:

$$\theta = \frac{\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right)}{AT\sigma^2_{\text{age}}} \tag{6.14}$$

Equation 6.14 is ready to be used in practical applications: It says that all we need to know in order to set a reasonable value for θ is an estimate of the standard deviation of the prior σ_{age}, a quantity which is easily interpretable, and in our experience is easy to elicit from subject matter experts.

But what specific numbers for σ_{age} should one choose? In our application, the scale of log-mortality means that a standard deviation of about 0.1 is usually a reasonable starting point, and it implies excursions of about plus or minus 0.3 (three standard deviations) around the prior mean. Obviously each case is different and there is no substitute for good judgment. Therefore our general recommendation at least in our mortality data is to start with 0.1 and try other values near 0.1. For example, since the effect of the smoothing parameter operates on a logarithmic scale, we usually also try values 0.05 and 0.2 (or sometimes as high as 0.3, as in the figure). If the highest value gives reasonable results there is no reason to move to lower values, with the risk of introducing unnecessary bias by restricting likelihood too much. In many examples, a range of values for σ gives similar results.

A key point is that setting θ, or equivalently the standard deviation of the prior σ_{age}, sets *all* the other "summary measures" of the prior. Therefore, if we have knowledge of several summary measures, a good strategy is usually to study the behavior of all of them as a function of the smoothness parameter of the prior. If each summary measure suggests different values for θ, then our prior "knowledge" may well be logically inconsistent and should be reconsidered.

Consider for example the case in which experts know that, in a certain age range $\mathcal{A}$, the change in log-mortality from one age group to the next is expected to remain around a certain level, or that, on average over time, is not expected to exceed a

certain level. This information could be captured respectively by the following 2 summary measures:

$$F_1(\mu) \equiv \frac{1}{T\#\mathcal{A}} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} |\mu_{at} - \mu_{a-1,t}| \tag{6.15}$$

$$F_2(\mu) \equiv \frac{1}{T} \sum_{t=1}^{T} \max_{a \in \mathcal{A}} |\mu_{at} - \mu_{a-1,t}| \tag{6.16}$$

$$\tag{6.17}$$

where $\mathcal{A}$ could be, for example, age range 20 to 80. The reason for restricting age groups to the range $\mathcal{A}$ could be that for many causes of death log-mortality varies much more in younger age groups, and deviates from the common, almost linear, pattern observed in older age groups.

The expected values of F_1 and F_2 depend on θ, and therefore on σ_{age}. In Figure 6.3 we report the expected values of F_1 and F_2 as a function of σ_{age} for a wide range of values of σ_{age}.

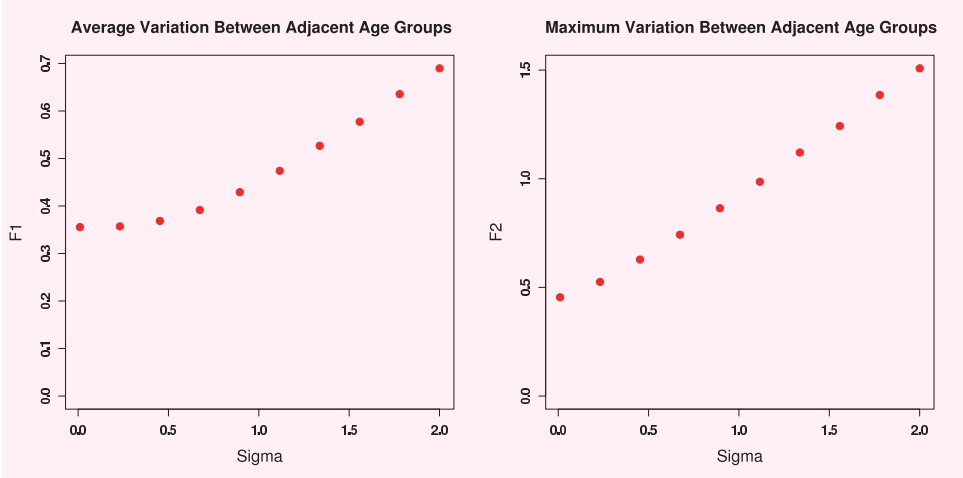

Figure 6.3: The expected value of the summary measures F_1 and F_2 defined in Equation 6.17 as a function of σ_{age}. This example refers to mortality from all causes in males, and the typical age profile μ was obtained by averaging the age profile of log-mortality of all countries in our data base with more than 15 observations. The country of interest is the US, and the covariates are a linear trend and GDP.

For very small values of σ_{age}, the prior is concentrated around the typical age profile $\bar{\mu}$, and therefore the expected values of F_1 and F_2 reflect the properties of μ. As expected, the expected value of both F_1 and F_2 increase with σ_{age}. Especially interesting in Figure 6.3 is that for both summary measures the expected value does not depend strongly on σ_{age} for small values of σ_{age} (say $\sigma_{\text{age}} < 0.5$): In this region the properties of

the samples from the prior, measured by F_1 and F_2, reflect quite closely the properties of the typical age profile $\bar{\mu}$, and are therefore within very reasonable limits. Only after a threshold is reached does increases in σ_{age} begin to produce noticable increases in the summary measures.

The shape of the curve corresponding to the summary measure F_1 is quite typical, in our experience. In fact this summary measure is closely related to another summary measure:

$$F_3(\mu) \equiv \frac{1}{T\#\mathcal{A}} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\mu_{at} - \mu_{a-1,t})^2$$

and we expect that

$$\sqrt{\mathrm{E}_{\perp}[F_3(\mu)|\theta]} \approx \mathrm{E}_{\perp}[F_1(\mu)|\theta]$$

Note that F_3 is a quadratic in μ, and therefore in $\boldsymbol{\beta}$. Furthermore, using the formulas of Appendix C the expected value of any quadratic form in $\boldsymbol{\beta}$ is a linear function of $\frac{1}{\theta}$, and therefore a linear function of σ^2_{age}. This implies the following dependency of $\mathrm{E}_{\perp}[F_1(\mu)|\theta]$ on σ_{age}:

$$\mathrm{E}_{\perp}[F_1(\mu)|\theta] \approx \sqrt{k_1 + k_2 \sigma^2_{\text{age}}}$$

which is precisely the behavior shown in the left panel of figure 6.3.

The fact that a single parameter, σ_{age} determines all the properties of the samples of the prior is both an advantage and a disadvantage. It certainly simplifies our task of chosing θ, but it may also create problems: In the example shown above the summary measures F_1 and F_2 assumed reasonable values, but what if the samples from the prior have the desired standard deviation but not the desired values of other summary measures? This is possible, especially if the summary measures involve dimensions on which the prior does not have much control, such as the time behavior. Therefore in these cases additional priors must be used, increasing the number of smoothness parameters and therefore the number of summary measures over which we have control. This issue will be discussed in chapter 7.

6.3 Choosing Where to Smooth

In Chapter 5, we considered smoothness functionals of the form

$$H[\mu, \theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\text{n}} \mu(a,t)}{da^{\text{n}}} \right)^2,$$

where the measure $dw^{\text{age}}(a)$ allows us to enforce a varying degree of smoothness for different age groups. In this section, we elaborate on this point and offer some practical examples of the effect of making different choices for $dw^{\text{age}}(a)$.

The choice of $dw^{\text{age}}(a)$ is related to the meaning of the dependent variable μ. As mentioned in Chapter 5, μ could be the expected value of log-mortality, or its deviation from some "typical" age profile $\bar{\mu}$. To clarify, we use μ to refer to the expected value of log-mortality, and introduce $\bar{\mu}$ explicitly for prior's mean for expected mortality. Since we are interested in the behavior over age groups, at any fixed point in time, we drop the time variable, so that μ is only a function of age, and the smoothness functional becomes:

$$H[\mu, \theta] \equiv \theta \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathbb{n}}}{da^{\mathbb{n}}} (\mu(a) - \bar{\mu}(a)) \right)^2 \tag{6.18}$$

If we choose $\bar{\mu} = 0$, and therefore a prior with zero mean, a non-uniform measure $dw^{\text{age}}(a)$ penalizes the variation in log-mortality from one age group to the next more in certain age groups and less in others. For example, suppose $dw^{\text{age}}(a)$ is such that it weights older age groups more than younger ones. Samples from such a prior will oscillate more and exhibit more variation at younger age groups, while they will be "stiffer" at older age groups.

We illustrate this idea in Figure 6.4, where each graph displays 100 samples from the prior associated to the smoothness functional in Equation 6.18. For these figures, we have chosen $\bar{\mu} = 0$, $\mathbb{n} = 2$, and a measure of the form $dw^{\text{age}}(a) = a^l da$, $l \geq 0$. The graph on the left corresponds to $l = 0$, that is to a uniform measure, and the one on the right to $l = 3$. Notice that the graphs differ in two respects. First, when $l = 3$ the variation in log-mortality from one age group to the next, in each sample, is much larger for younger age groups than for older ones, as expected. And second, within each age group, the variance of log-mortality is, on average, higher in younger age groups (compare the huge variation observed at age 0 with the smaller variation observed at age 80).

In order to get an idea of how the parameter l can affect the result we introduce a different way of looking at a smoothness functional: Instead of looking at samples from the prior we use it to smooth the data in a classic non-parametric framework, and plot the results for different values of θ which indexes how much effect the prior has. A large class of non-parametric smoothers is the one defined by the following minimization problem:

$$\min_{\mu} \ \|\mu - m\|^2 + \theta \mu' W \mu \tag{6.19}$$

where m is an observed age profile $m = (m_1, \ldots, m_A)$ and $\mu' W \mu$ is the discretized version of the smoothness functional 6.18. Smoothers of this type are common in regularization and spline theory, and can be interpreted as Bayes estimates, as shown by Kimeldorf and Wahba (1970). For our purposes the only thing we need to know about the problem 6.19 is that its solution is given by:

$$\mu(\theta) = (I + \theta W)^{-1} m \tag{6.20}$$

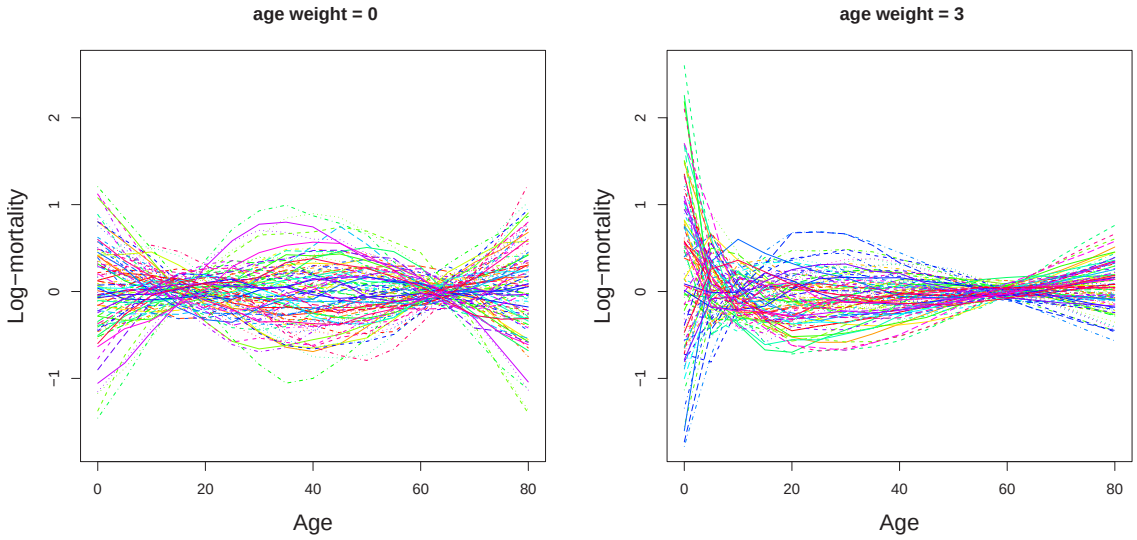

Figure 6.4: 100 random draws from the smoothness functional in Equation 6.18, with $\mu = 0$, $\mathbb{n} = 2$, and measure $dw^{\text{age}}(a) = a^l da$. For the graph on the left $l = 0$ (a uniform measure) and for the graph on the right $l = 3$. The standard deviation, averaged over age groups is the same in both graphs and equal to 0.3.

and that it has the following special cases:

$$\mu(0) = m \ , \quad \lim_{\theta \to \infty} \mu(\theta) = P_\circ m$$

where $P_\circ$ is the projector onto the null space of the smoothness functional in Equation 6.18. The last identity is easily derived using the eigenvector/eigenvalue decomposition of W and partitioning the eigenvectors into a basis for the null space and a basis for its orthogonal complement. Its interpretation is simple: When θ goes to infinity the smoothed version of m is its best approximation from the null space of W. For example if we choose $\mathbb{n} = 2$, so that the null space consists of the linear functions, when θ goes to infinity the smoothed version of the data is the straight line which best fits m.

In Figure 6.5 we report the result of the smoother in Equation 6.20 for different values of θ and for different choices of l. The data in the two graphs are the same (the red dots), and correspond to the age profile of log-mortality for respiratory infectious disease in Sri Lankan males in year 2000. In each graph we have plotted the smoothed version of the data for 12 values of θ from 1,000 to 0 (the value 1,000 is for all practical purpose equal to infinity). The smoothed curves are color-coded along the rainbow colors: the lines in red to yellow correspond to very large values of θ, while those in blu-violet to very small values of θ. The only difference between the graphs is the value of l, which is 0 in the left graph and $l = 4$ in the right graph. Notice that when $l = 0$ too much smoothing occurs at younger age groups, and it is difficult to find a

value of θ which smooths the data well. When $l = 4$, instead, the smoothed curves are allowed to remain fairly steep and to "bend" of a considerable amount at young ages, even for relatively large values of θ, producing a better range of smoothing curves.

Notice also how the smoothed curves become a straight line when θ becomes very large: This is an undesirable feature of this smoothness functional, and a consequence of having chosen $\bar{\mu} = 0$ and a null space consisting of straight lines. If a "typical" age profile $\bar{\mu}$ is available much better results can be obtained, as we will now show. For the purpose of this experiment we have synthesized a profile $\bar{\mu}$ by averaging the age profiles of the 50 countries (not including Sri Lanka) for which at least 20 observations are available (this criterion aims to select countries with "good" time series). Alternatively, we could have chosen a to average only the countries which are "neighbor" of Sri Lanka.

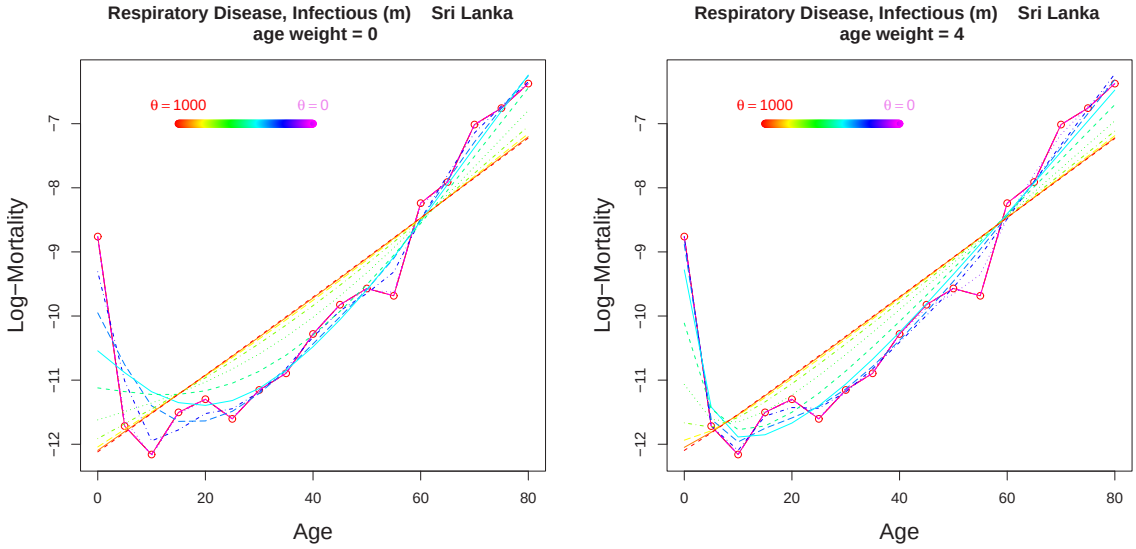

Figure 6.5: Smoothed versions of the age profiles of respiratory infectious disease in Sri Lankan Males (data from year 2000). Here we use prior 6.18 with zero mean, and report in each figure the results for θ from 0 to 1000. For the graph on the left $l = 0$ (uniform measure) and for the graph on the right $l = 4$.

We now show in Figure 6.6 the same kind of graphs we showed in figure 6.5, with the only difference being that $\bar{\mu}$ is no longer zero. Now there is much less difference between the two graphs, since the mean of the prior is responsible for explaining most of the large variation between age groups at younger ages. This could be expected: using a prior with non-zero mean is in fact equivalent to use a prior with zero mean where the dependent variable is the deviation of log-mortality from the typical age profile. While for log-mortality there is a strong argument for penalizing non-smooth behavior less at younger age groups, a similar argument is less clear if the dependent variable is the deviation of log-mortality from the typical age profile. In this case

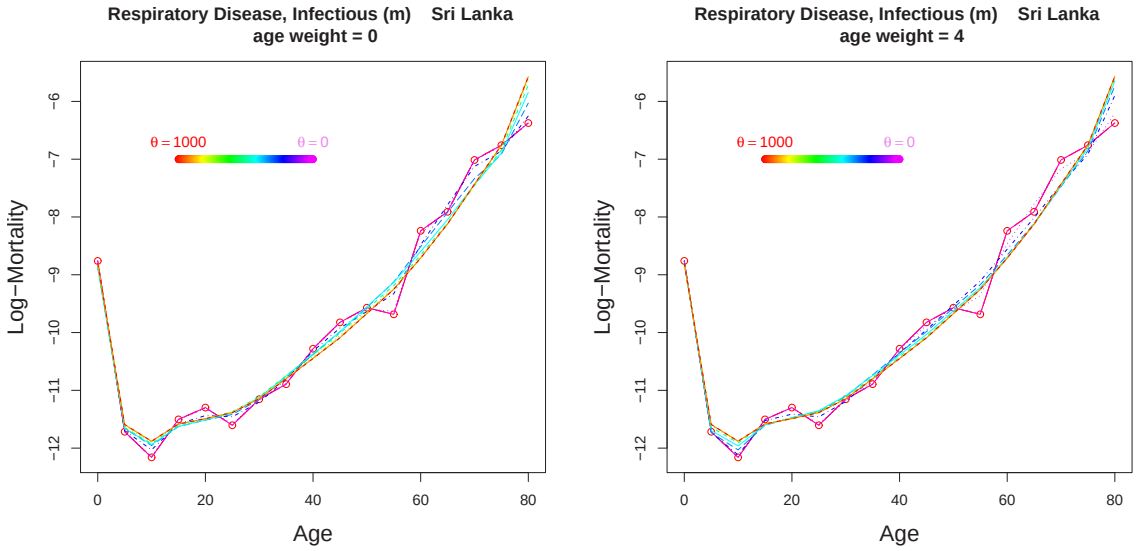

Figure 6.6: Smoothed versions of the age profiles of respiratory infectious disease in Sri Lankan Males (data from year 2000). Here we use prior 6.18 with mean $\bar{\mu}$ different from 0, and report in each figure the results for θ from 0 to 1000. For the graph on the left $l = 0$ (uniform measure) and for the graph on the right $l = 4$.

the reason for having $l \neq 0$ is slightly different: in some cases knowledge about the shape of the age profiles could be more accurate in older age groups than in younger age groups, and therefore we would like to have a prior whose variance is higher in younger age groups. By smoothing less at younger age groups we are also allowing the variance within each of the young age groups to be higher. Therefore, even if the prior has non-zero mean we may still want to use a value of l different from 0. We illustrate this effect in figure 6.7, which is the counterpart of figure 6.4, but with $\bar{\mu}$ chosen as in figure 6.6. Now the samples from the prior are quite similar: the main difference between the two graphs is that, within each group, the variance of the samples is higher at younger age groups. Since the graphs in figure 6.6 correspond to the priors whose samples are represented in figure 6.7, it is not surprising that the results with $l = 0$ and $l = 3$ are fairly similar.

Obviously other forms of prior knowledge on the shape of age profiles could be available, which may not be represented by the simple choice $dw^{\text{age}}(a) = a^l da$ (for example, one may want to allow more variation both in young and old age groups, but not in the middle ones). We suggest that, in any case, researchers study graphs of the kind we have produced here in order to understand what is the prior which best represent their knowledge. (Our software produces these graphs automatically; see Appendix F.)

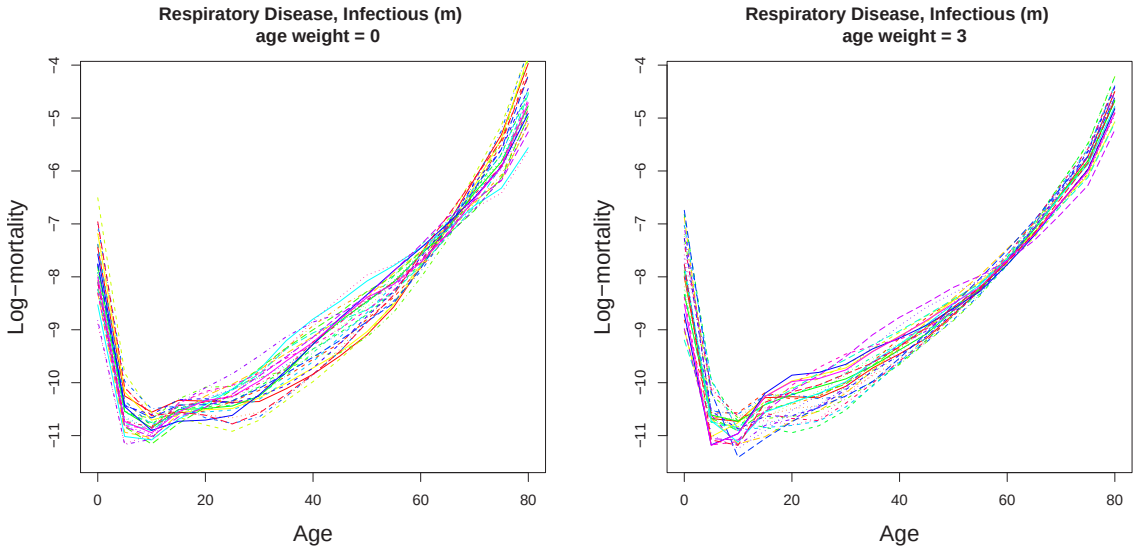

Figure 6.7: 100 random draws from the smoothness functional in Equation 6.18, with a non-zero mean $\bar{\mu}$, $\mathtt{n} = 2$, and measure $dw^{\text{age}}(a) = a^l da$. For the graph on the left $l = 0$ (a uniform measure) and for the graph on the right $l = 3$. The standard deviation, averaged over age groups is the same in both graphs and equal to 0.3.

6.4 Choosing Covariates

The choice of covariates in regression models is normally a major decision, as or more important than most of the other statistical issues that often arise. And indeed, the same rules apply in forecasting with our models as with any other use of regression for forecasting: Choose covariates that pick up on systematic patterns that are likely to persist, rather than idiosyncratic features likely to overfit in-sample data only. Reduce the chances of overfitting by using priors to reduce the effective sample space, or if necessary drop covariates. Et cetera. The importance of these usual cautions are hard to overestimate, since even well-designed priors will not always avoid the bias induced by misspecifying covariates. But our procedure involves an additional factor that is implied by everything that has come before in this book and that we now make explicit.

The second step of our two-step procedure in Section 5.2 is to project the prior specified in terms of the expected value of the dependent variable μ on the subspace spanned by the covariates $\mathbf{Z}_{at}$ into the lower-dimensional vector of coefficients $\boldsymbol{\beta}$. Effectively, we are able to invert what would be a non-invertible (many-to-one) relationship by restricting the full prior on μ to the subspace that spams $\mathbf{Z}$, for which the equation $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$, is invertible. The key to the whole procedure, however, is having a set of covariates that makes it possible to express the relationships of interest specified under the prior for μ. The danger is that a rich prior for μ could

be matched with an impoverished set of covariates such that the resulting prior restricted to the subspace spanned by the covariates is not able to reflect most of the interesting patterns allowed under the original unrestricted prior for μ.

Thus, we now provide tools with which one can check to see that important characteristics of the prior are not lost in when we take the projection. We focus in particular on the null space, since when we project the non-parametric prior on the space spanned by the covariates we also project its null space, and in principle it is even possible for this operation to cause the null space to disappear or at least to be greatly reduced. Since this would be an undesirable feature, we need to understand the conditions under which this could happen and how to avoid it.

6.4.1 Size of the Null Space

For simplicity, we drop the indexes over countries, and denote by $\mu_t \in \mathbb{R}^A$ an age profile in year t. The non-parametric prior on μ can then be written as

$$H[\mu, \theta] = \frac{\theta}{T} \sum_t \mu_t' W \mu_t \tag{6.21}$$

where $W = W^{\text{age},n}$ (the squared discretized derivatives of μ with respect to age) in the rest of this section. We assumed a zero mean for this prior, since the mean is irrelevant for the computation of the dimension of null space. This prior is defined over μ in $H[\mu, \theta]$, which represents the $A \times T$ dimensional space of the T age profiles. Let $\mathfrak{N}(W)$ denote the null space of W (which, as per Section 5.1, determines what patterns of μ the prior is indifferent to), and let $\dim(\mathfrak{N}) \equiv \text{nullity}(W)$ denote its dimensionality.

The null space of the non-parametric prior in Equation 6.21 is obtained by allowing the age profile of each year to vary *independently* from the other years, in $\mathfrak{N}$. This implies that the dimensionality of the null space of the prior 6.21 is $T \times \text{nullity}(W)$. When we project the prior on the space spanned by the covariates we obtain the following:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}_a' \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}. \tag{6.22}$$

This prior is defined on a much smaller space than Equation 6.21. Or, in other words, the prior in Equation 6.21 on μ, restricted to the space where $\mu = \mathbf{Z}\boldsymbol{\beta}$ holds, has a much smaller null space than before imposing the restriction.

In order to fix ideas, assume 17 age groups and 60 years of observations for a set of 7 covariates, with a non-parametric prior involving the second derivative only, so that $\text{nullity}(W) = 2$. The prior for μ is defined over $\mathbb{R}^{1020}$ ($1020 = 60 \times 17$), and its null-space has dimension 120 ($120 = 60 \times 2$). The prior on the coefficients in Equation 6.22 is defined over $\mathbb{R}^{119}$ ($119 = 7 \times 17$), which is less than the dimensionality of the whole null space of the non-parametric prior!

In order to study the dimensionality of the null space of the prior on the coefficients in Equation 6.22, it is convenient to start from the simple case in which the covariates do not vary across age groups (like GDP, for example). Therefore the number of covariates in age groups 1 to A is the same, and we denote it by k. In this case we have:

$$\mathbf{C}_{aa'} \equiv \mathbf{C} \quad \forall a, a'$$

where $\mathbf{C}$ is a symmetric $k \times k$ matrix and the prior has the form:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}_a' \mathbf{C} \boldsymbol{\beta}_{a'} \tag{6.23}$$

Since the dimensionality of the null space obviously does not depend on the particular coordinate system we use to compute it, we perform a convenient change of variables. We assume, without loss of generality, that the covariates are orthogonal. Therefore the substitution

$$\sqrt{\mathbf{C}} \boldsymbol{\beta}_a \to \boldsymbol{\beta}_a \tag{6.24}$$

is an invertible transformation (i.e., C^{-1} will exist because of the absence of collinearity among the covariates), and we can study the prior in Equation 6.23 in the new system of coordinates:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}_a' \boldsymbol{\beta}_{a'}.$$

Now denote by β_a^q the q-th component of the vector $\boldsymbol{\beta}_a$, so that $q = 1, \ldots, k$, and the $A \times 1$ vector $\boldsymbol{\beta}^q$ whose elements are β_a^q. Then we rewrite the expression above, which sums over age groups, as one which sums over covariates:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{q=1}^{k} (\boldsymbol{\beta}^q)' W \boldsymbol{\beta}^q. \tag{6.25}$$

where recall that $k = 7$ in our running numerical example. In order for $H[\boldsymbol{\beta}, \theta]$ to be zero each vector $\boldsymbol{\beta}^q$, with $q = 1, \ldots, k$, must be in the null space of W, which has dimension nullity(W). Therefore, *the null space of the prior in Equation 6.23 has dimension nullity*(W) $\times k$. Using the numbers in the example above we would have that the prior over the coefficients, which is defined over $\mathbb{R}^{119}$, has a null space of dimension 14 ($14 = 2 \times 7$).

6.4.2 Content of the Null Space

Expression 6.25 also allows us to identify the exact content of the null space. Let us choose all the $\boldsymbol{\beta}^q$ to be 0 except for $q = q^*$, and let us assume that the prior

over age groups is a standard smoothness prior with derivative of order $\mathtt{n}$ (for the mixed smoothness similar reasoning applies). Then the null space of W is the set of polynomials of degree $\mathtt{n}-1$. Therefore $\boldsymbol{\beta}^{q^*}$ is in the null space of W if it can be written as:

$$\beta_a^{q^*} = \sum_{j=0}^{\mathtt{n}-1} v_i^{q^*} a^j \ , \quad \text{for any } v_i \in \mathbb{R} \ , \quad i = 0, \dots \mathtt{n}-1$$

For this choice of coefficients, the patterns of log-mortality which belong to the null space of the prior are described as

$$\mu_{at} = z_t^{q^*} \beta_a^{q^*} = z_t^{q^*} \sum_{j=0}^{\mathtt{n}-1} v_j^{q^*} a^j \ , \quad \text{for any } v_j \in \mathbb{R} \ , \quad j = 0, \dots \mathtt{n}-1$$

These are patterns which at any point in time have an age profile which looks like a polynomial of degree $\mathtt{n}-1$, but whose coefficients evolve over time as the covariate $z_t^{q^*}$. Suppose for example that $\mathtt{n}=2$ and that q^* corresponds to the covariate GDP, so that $z_t^{q^*} = \mathrm{GDP}_t$. Then a pattern in the null space of the prior can be written as

$$\mu_{at} = \mathrm{GDP}_t(v_1 + v_2 a) \ \text{ for any } v_1, v_2 \in \mathbb{R}$$

This reasoning can be used for any given $q^* = 1, \dots k$, and taking a linear combination of the corresponding $\boldsymbol{\beta}^q$ we can obviously span the null space of the prior. Therefore the null space of the prior consists of patterns of log-mortality of the following general form:

$$\mu_{at} = \sum_{q=1}^{k} z_t^q \sum_{j=0}^{\mathtt{n}-1} v_j^q a^j, \ \text{ for any } v_j^q \in \mathbb{R},$$

where the coefficients v_j^q are arbitrary numbers (notice that there are exactly nullity$(W)\times$ k of them).

Covariates that Vary over Age Groups So far we have discussed the restrictive case in which the covariates are the same for all the age groups. The main observation necessary to understand the general case is that *the more the covariates differ across the age groups, the smaller the dimension of the null space of the prior.* The reason for this is that in order for a cross-sectional time series to be in the null space of the prior, the age profile for each year must be in the null space of W. That means that the coefficients have to satisfy, for each year, a complicated condition involving the covariates and the matrix W. In other words, *if the covariates have no regularity across the age groups it may be impossible for the prior to find a set of coefficients that satisfies a requirement of regularity for every year.*

We reinforce this intuition with the following example. Suppose the covariates z_{at}^r are zero mean, unit standard deviation, i.i.d. random variables. If T is large enough we will have

$$\mathbf{C}_{aa'}^{qr} = \frac{1}{T}\sum_{t=1}^{T} z_{at}^q z_{a't}^r \approx \delta_{aa'}\delta_{qr}$$

where δ is Kronecker's delta. In this case the prior 6.22 becomes

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T}\sum_a W_{aa}\|\boldsymbol{\beta}_a\|^2.$$

Since W is semi-positive definite its diagonal elements W_{aa} are positive, and therefore the null space of this functional is $\boldsymbol{\beta}_a = 0$ for all $a = 1, \ldots, A$, and the the prior is proper: The lack of correlation of the covariates across age groups has shrunk the null space of the prior to zero.

In order to quantify this intuition let us consider the case in which there are k covariates, but some of them may be missing in some age groups (by making the number of unique covariates large enough any case can be seen as a special case of this). Suppose for example we have 7 covariates, one of which is missing below a certain age. The prior in Equation 6.22 is now defined over a space of dimensionality smaller than $A \times k$. However, we can rewrite it as a prior defined over $\mathbb{R}^{A\times k}$, but with a constraint: we can "fill in" the missing covariates with arbitrary values, but constraining the corresponding coefficients to be zero. This constraint is formalized by saying that the coefficients $\boldsymbol{\beta}$ belong to a subspace $\mathbb{S}$ of $\mathbb{R}^{A\times k}$. Since the unconstrained prior has the same covariates in each age group, the dimensionality of its null space is $\dim(\mathfrak{N}) \times k$. It follows that the dimensionality of the null space of the constrained prior has to be lower than that, and therefore $\dim(\mathfrak{N}) \times k$ provides a convenient upper bound.

A lower bound is available as well. In fact, let us drop the covariate which is missing in some age groups altogether or, equivalently, let us set to zero the coefficients corresponding to this covariate for all age groups. The resulting prior also has the same covariate in all the age groups, and therefore the dimensionality of its null space is $\dim(\mathfrak{N}) \times (k-1)$. Since this is a projection of the prior in Equation 6.22 over a lower dimensional subspace this is a lower bound for the dimension of the null space of the prior 6.22. This result is comforting: it implies that *as long as there is at least one covariate (for example the constant) which is the same across all age groups the prior 6.22 is improper*. More precisely, if there are l covariates which are the same across the age groups, the dimension of the null space is at least $\dim(\mathfrak{N}) \times l$. We conjecture that this number is actually the correct dimensionality of the null space, but we have not proved it yet.

The discussion in this section applies to the a prior defined over age groups, but its logic also applies to any of the priors that we will describe in the next chapter,

where instead of smoothing the expected value of the dependent variable over age groups we smooth it, for example, over time or countries. A detailed example of how the null space depends on the covariates is presented in Section 7.1.

6.5 Choosing a Likelihood and Variance Function

Through most of this book, and in our software implementation, we specify the logarithm of the mortality rate to be normally distributed, a choice we refer to as "the normal specification". We also model the mean of the normal density as a linear function of the covariates, and we have let its variance be an unknown parameter σ_i^2, indexed by the cross-sectional index. Since our approach is Bayesian, we model the variances as random variables with a probability density $\mathcal{P}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. This density must satisfy two constraints: It must reflect knowledge we have about the variances, and it must lead to a computationally feasible model. The problem of finding a suitable model for $\mathcal{P}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ cannot be discussed without discussing the motivations behind our choice of the normal specification and the approximations and sources of errors associated with it. Therefore we start this section by reviewing the usual rationale for the normal specification.

6.5.1 Deriving The Normal Specification

The raw variable we observe is d_{it}, the number of people who die in year t in cross-section i. Because of the "count" nature of this variable, a reasonable starting point is to assume that d_{it} can be described by a Poisson process, with unknown mean λ_{it}. We summarize this information as follows:

$$d_{it} \sim \text{Poisson}(\lambda_{it}) \ , \quad \text{E}[d_{it}] = \lambda_{it}, \ , \quad \text{Var}[d_{it}] = \lambda_{it} \tag{6.26}$$

This model is highly constrained, since the mean and variance of this density are not independent. A more flexible model would be given by a Pólya process, rather than a Poisson process, where the Poisson density would be replaced by a negative binomial, in which the variance can be any number larger than the mean, or the generalized event count model which allows the variance to be greater than or less than the mean (King, 1989*a*; King and Signorino, 1996). We do not consider alternative processes here, since that would considerably lengthen our exposition without leading us in the end to different practical choices. From a conceptual point of view the various count models are appealing, and there is nothing in our model which prevents us from using any of them, since they simply correspond to different choices of the likelihood in the expression for the Bayesian estimator in Equation 4.3 (Page 68). However, since they would lead to fairly complicated implementations, we look for a computationally simpler alternative.

The key observation at this point is that if we think of the Poisson density as a function of a continuous random variable, then it can be well approximated, under

certain conditions and for appropriate choices of the parameters, by a lognormal density. The lognormal is a density with two free parameters, ν and ϱ, whose functional form is reported in Appendix B.3.3 (Page 254). In the following, if a random variable d has a lognormal density, we write $d \sim \log\mathcal{N}(\nu, \varrho^2)$. If we want to approximate a Poisson density with a lognormal density we must choose the parameters ν and ϱ in such a way that the mean and variance of the lognormal match the mean and variance of the Poisson density. Using the formulas in Appendix B.3.3 (Page 254) it is easy to see that the lognormal approximation to the Poisson density of Equation 6.26 is:

$$d_{it} \sim \log\mathcal{N}\left(\log\lambda_{it} + \frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right), \log\left(1+\frac{1}{\lambda_{it}}\right)\right). \tag{6.27}$$

The advantage of the approximation of the Poisson density by a lognormal is that

$$x \sim \log\mathcal{N}(\nu, \varrho^2) \Longleftrightarrow \log x \sim \mathcal{N}(\nu, \varrho^2).$$

Thus, if we could model the observed number of deaths by Equation 6.27 it would follow immediately that log-mortality would be modeled with a normal distribution. In fact, dividing d_{it} in Equation 6.27 by population p_{it} and using the property above we obtain:

$$m_{it} \sim \mathcal{N}\left(\log\frac{\lambda_{it}}{p_{it}} + \frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right), \log\left(1+\frac{1}{\lambda_{it}}\right)\right) \tag{6.28}$$

Since by definition $\lambda_{it}/p_{it} = \mathrm{E}[M_{it}]$, where $M_{it} = d_{it}/p_i t$, the expression above implies that

$$\mu_{it} \equiv \mathrm{E}[m_{it}] = \mathrm{E}[\log M_{it}] = \log\mathrm{E}[M_{it}] + \frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right) \geq \log\mathrm{E}[M_{it}]$$

As expected (from Jensen's inequality), the expected value of log mortality is larger than the log of the expected value of mortality, with the difference decreasing as the expected number of deaths increases.

Before discussing the implications of the expression above, we first study precisely when the approximation that led to it is appropriate.

6.5.2 Accuracy of the Lognormal Approximation to the Poisson

Here we compare the Poisson density, seen as a function of a continuous variable, with its lognormal approximation. This is done in Figure 6.8, where we show both densities for four different values of the mean.

As illustrated in the figure, the approximation improves as the expected value of the number of deaths, λ_{it}, increases, with large errors occurring when $\lambda_{it} < 5$. The crucial difference between the lognormal and the Poisson density is the behavior

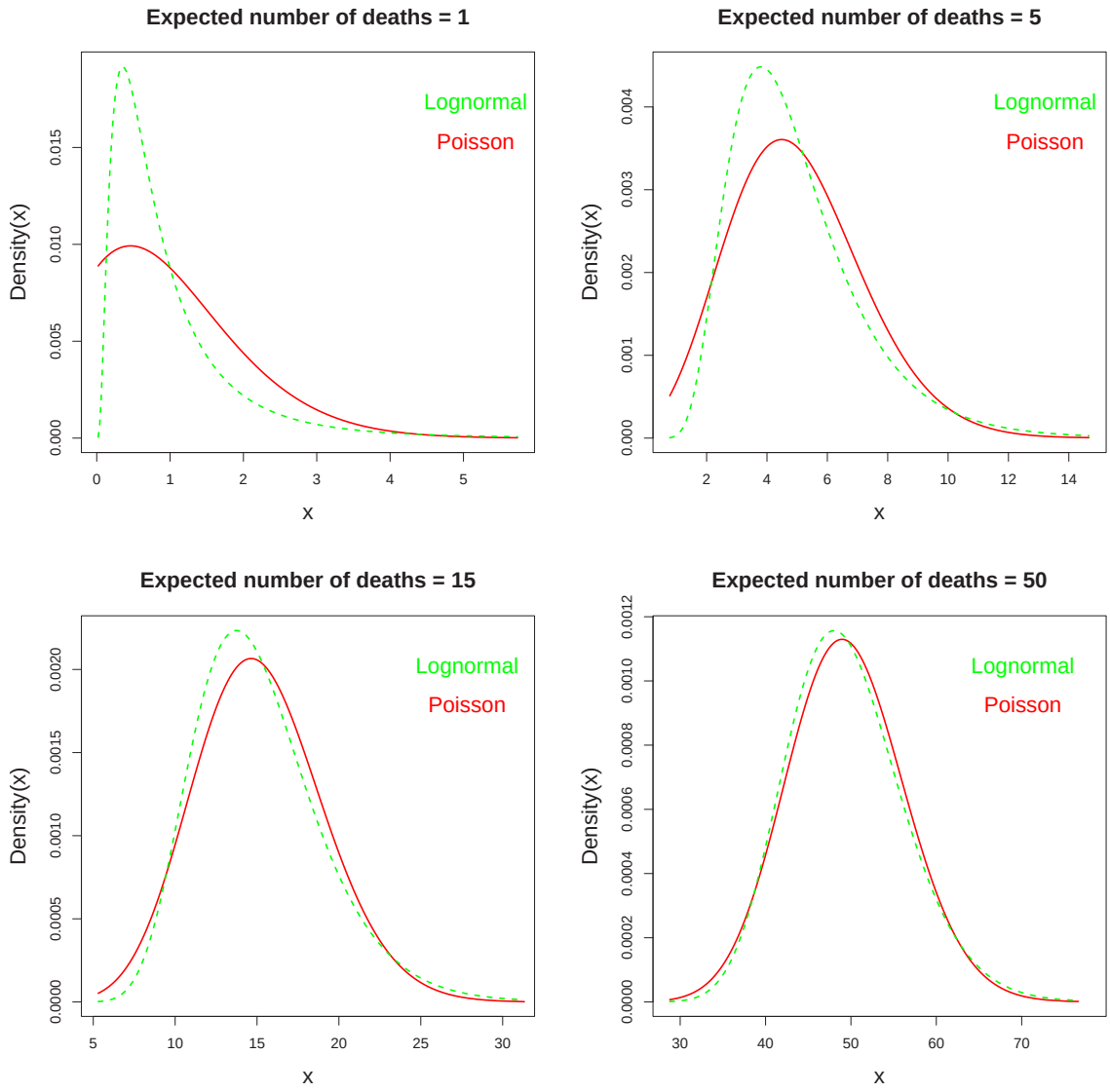

Figure 6.8: The Log-Normal Approximation (in green) to the Poisson density (in red), as a function of number of deaths, for different expected numbers of deaths.

at the origin: While the lognormal density is 0 at the origin (because it contains a negative exponential in $(\log d_{it})^2$), the Poisson is not. Therefore a sample from the lognormal may generate d_{it} close to zero, but never zero, while a sampling from the Poisson (as a discrete distribution) can certainly generate $d_{it} = 0$. This implies that any attempt to use the lognormal density in a likelihood where the data are generated by a Poisson distribution will result in an attempt to compute the logarithm of zero.

To give an idea of the numbers involved, with a Poisson density when $\lambda_{it} = 1$ the probability of observing a 0 is 37%, while when $\lambda_{it} = 5$ this probability drops to 0.7%, dropping to a negligible value of 4.5×10^{-5} when $\lambda_{it} = 10$.

These considerations suggest that there are three "regimes" in which we may need to operate:

1. Large value of λ_{it} This is the case in which the observed numbers of deaths contain no zeros. This is likely to happen when the expected number of deaths λ_{it} is always larger than 10 or 15. In this situation Equation 6.27 can be simplified:

$$d_{it} \sim \log\mathcal{N}\left(\log\lambda_{it} + \frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right), \log\left(1+\frac{1}{\lambda_{it}}\right)\right) \approx \log\mathcal{N}\left(\log\lambda_{it}, \frac{1}{\lambda_{it}}\right)$$

To see the amount of error involved in these approximations let us take $\lambda_{it} = 15$. In this case the term $\frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right)$ in the mean is -0.032, which is negligible when compared to $\log\lambda_{it} = 2.7$ (of a factor 100). For the term in the variance we have that $\log\left(1+\frac{1}{\lambda_{it}}\right) = 0.064$, which is well approximated by $\frac{1}{\lambda_{it}} = 0.066$.

The corresponding simplified specification for log-mortality is then:

$$m_{it} \sim \mathcal{N}\left(\log\frac{\lambda_{it}}{p_{it}}, \frac{1}{\lambda_{it}}\right). \tag{6.29}$$

In this regime, the expected value of log-mortality and the log of the expected value of mortality essentially coincide, and the variance of log-mortality is inversely proportional to the expected value of the number of deaths. This situation is very common when dealing with all-cause mortality or for the leading causes of death in countries which are not too small and, in most cases, for other than very young ages. The pattern is less common as we move to rarer causes of death, small countries or younger age groups.

To provide some specificity, consider a common cause of death, cardiovascular disease, in an hypothetical country, similar to the U.S., and of total population 280 million, for age group 70–74 among males. Reasonable values in this case are $\lambda_{it} = 60,000$ and $p_{it} = 4,000,000$, which correspond to $\mathrm{E}[M_{it}] = 1.5\%$. Under these conditions the problems of zeros is nonexistent, and the lognormal and Poisson density are virtually identical.

Now "scale" this hypothetical country down by a factor of 600, keeping the mortality rate constant. We would obtain an expected number of deaths of 100 in a

population of 6,666 people, in a country with a total population of about 470,000. For this country, we would still not expect any zeros in the observed number of deaths, and the lognormal approximation would still be appropriate. If we scaled our initial country down by a factor 36,000 then we could run into problems: The expected number of deaths would be only 1.6 in a population of 111 people, and the total population of the country would be only 7,777. Under these conditions we should expect a non-negligible number of zeros in the observed number of deaths, which would make the application of the model in Equation 6.29 practically impossible (since we would need to take the logarithm of zero) and imprecise (even if by luck we do not have zeros, the lognormal density does not approximate the Poisson density very well in this circumstance).

To make this discussion more empirical: In year 2000, two countries whose population in age group 70-74 were around 111: The Cook Islands and Palau. The observed number of deaths by cardiovascular disease for males in this age group for the two countries was 3 and 5, respectively, probably reflecting higher mortality rates than the U.S. In younger age groups, these countries exhibit non-negligible numbers of zeros in observed deaths.

2. Small value of λ_{it} This is the case in which the data will have some observed zeros, although most of the data will not contain zeros. We can expect this situation whenever λ_{it} is somewhere between 2 and 10 (for $\lambda_{it} = 2$ a Poisson density will generate data which are 0 about 13% of the time). In this case we face two problems: (1) in this range of λ_{it} the lognormal is not a very good approximation of the Poisson distribution, and (2) we do not know what to do when $d_{it} = 0$ since its logarithm is not defined. A common "fix" to the second problem consists of assigning to each cross-section an extra 0.5 deaths every year (Plackett, 1981).

Since there are great computational advantages in retaining a normal specification for log-mortality, like the one in Equation 6.29, we now study the size of errors associated with this procedure. To begin, suppose λ_{it} is small, but we add 0.5 deaths to each observation and proceed as if λ_{it} were large: How large is the error we make if we estimate λ assuming that Equation 6.29 still holds?

To study this question, we take 500,000 draws from a Poisson distribution with mean λ, for $0.5 \leq \lambda \leq 10$ (we omit the indexes it for simplicity, as if we were considering one specific cross-section in one specific year, and also take the total population to be 1). To these points we add a value of 0.5 and then we take their logarithm. We consider the result our sample of log-mortality, which we analyze as if it were normally distributed according to Equation 6.29. Let $\hat{\mu}$ be the empirical average of log-mortality in our sample. If Equation 6.29 holds then we can estimate λ as $\hat{\lambda} = e^{\hat{\mu}}$. We repeat this procedure for many different value of λ, and for each value we compute the percentage error $\frac{|\hat{\lambda}-\lambda|}{\lambda}$. We report our results in Figure 6.9.

The approximation error we obtain with this procedure is surprisingly small, dropping below 2% for λ greater than 2. Although this is reassuring, it does not mean,

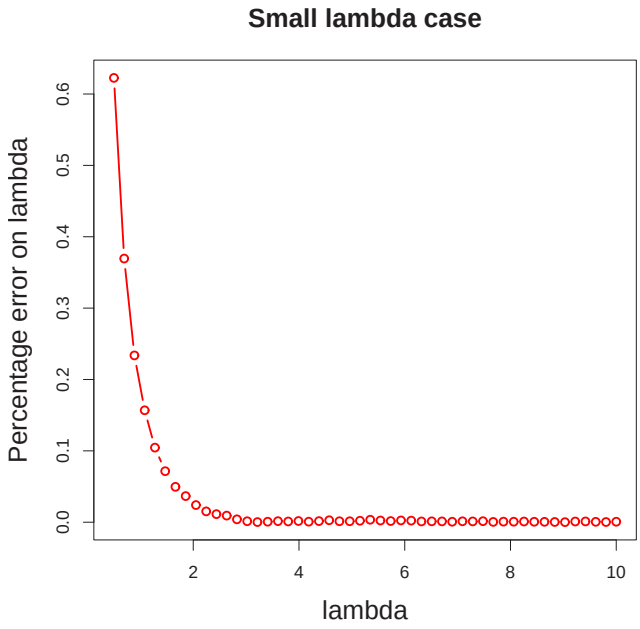

Figure 6.9: The error in estimating the expected number of deaths from log-mortality with zeros in observed deaths and 0.5 added to each observation. The expected number of deaths is estimated assuming Equation 6.29 is still valid.

however, that the density of log mortality with the "fix" is well represented by Equation 6.29: it merely says that its expected value is well approximated by the expected value of the density 6.29 (although a similar phenomenon holds for the variance too). In order to perform a more stringent test we perform a different simulation.

Thus, we generate a sample of 500,000 points from a Poisson distribution with mean λ, for $0.5 \leq \lambda \leq 10$, add 0.5 and take their logarithm as before. We consider the resulting sample our data for log-mortality, which we now analyze as if it were normally distributed according to $\mathcal{N}(\mu, \sigma^2)$, where estimates $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ are obtained in standard way. We do not make any assumption about how μ is related to λ. In order to estimate λ we sample from $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ to obtain a new sample for log-mortality, which we then convert to a sample for the number of deaths by simple exponentiation. This is the crucial step, since this sample will not look like the sample obtained from the Poisson density, especially when λ is small (it will have a lognormal distribution). As a final step we compute the empirical average of the mortality values from the new sample, which is an estimate of λ that we denote by $\hat{\lambda}$. We repeat this procedure for many different value of λ, and for each value we compute the percentage error $\frac{|\hat{\lambda}-\lambda|}{\lambda}$. We report our results in Figure 6.10, which now displays larger errors.

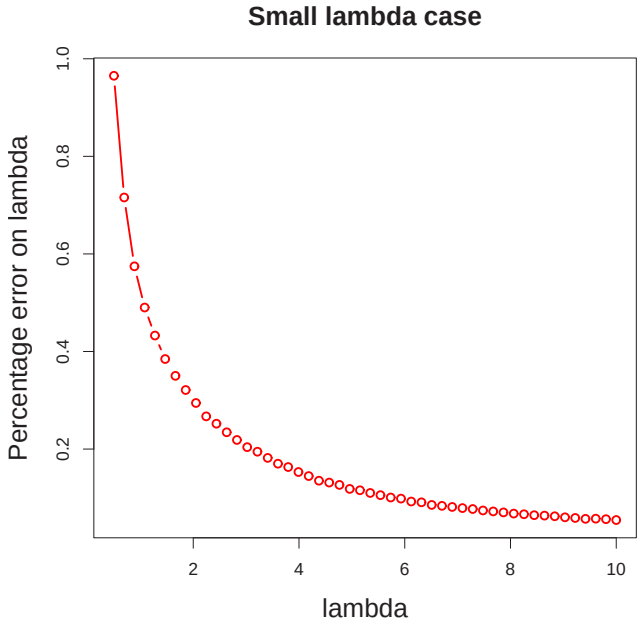

Figure 6.10: Error in estimating expected deaths from log-mortality with observed zeros in the number of deaths and with 0.5 deaths are added to each observation. We have not assumed that Equation 6.29 holds.

But how large are these errors? When λ is small, say three, the standard deviation

of the Poisson density is quite large (for $\lambda = 3$ it is $\sqrt{3} = 1.73$), and therefore there is a lot of variation built in the data. Thus, expecting great accuracy in estimating λ is unreasonable, even if we use the correct assumptions. Examining Figure 6.10 with this in mind, and noting that a percentage error on λ of even 20% is not large on this scale, we note that the "fix" of *adding 0.5 deaths to each observation could probably be used for values of λ as small as 3 without serious consequence.* This assessment takes explicitly into account the difference between the lognormal and the Poisson distribution, and therefore it is more informative than the one of Figure 6.9.

The figures shown so far suggest a range of values of λ_{it} such that adding 0.5 to the number of deaths does not noticeably destroy the information about the expected values of death, while still allowing one to use a normal specification for log-mortality. It remains to be shown that this procedure does not alter the structure for the variance. To this end it is instructive to perform a simple simulation, and compute the variance of $\log(d_{it} + 0.5)$ when $d_{it} \sim \text{Poisson}(\lambda_{it})$ for several values of λ_{it}. Given the results above we expect that, in the regime in which the Poisson and the lognormal density are not too far apart, we would expect this variance to be equal to $\log(1 + \frac{1}{\lambda_{it}})$ (see Equation 6.27). We test this hypothesis in Figure 6.11, where on the horizontal axis we have λ_{it}, and the curve in green is $\log(1 + \frac{1}{\lambda_{it}})$ as a function of λ_{it}, while the one in red is $\text{V}[\log(d_{it} + 0.5)]$ for $d_{it} \sim \text{Poisson}(\lambda_{it})$. Again, *serious deviations between these two curves occur only for $\lambda_{it} < 3$.*

Figure 6.11 underlines an important point about the variance of log-mortality (a topic also raised in Girosi and King 2005). In general it is not possible to define the random variable log-mortality $m_{it} = \log d_{it}$ (assuming population $p_{it} = 1$), unless we know that d_{it} never assume zero values, which is certainly not the case when λ_{it} is small and d_{it} is a Poisson process. Therefore it does not make sense to talk of the variance of log-mortality when λ_{it} is very small. It does makes sense to define the random variable $\log(d_{it} + \alpha)$, where α is any number larger than 0. It is tempting therefore to interpret the variance of log-mortality as the variance of $\log(d_{it} + \alpha)$ and then let α go to 0. Unfortunately Figure 6.11 suggests that this cannot be done. The figure, which corresponds to $\alpha = 0.5$, is representative of the behavior of the variance of $\log(d_{it} + \alpha)$: for $\lambda_{it} = 0$ the variance is 0 (since the density is concentrated at the origin) and in a neighbor of $\lambda_{it} = 0$ the variance increases with λ_{it}, before starting to decrease. For values of α smaller than 0.5 the location of the maximum will shift to the left and the curve will become steeper, but the shape of the curve remains the same. As a result the limit of this curve for α going to 0 does not exist at the origin (the curve becomes discontinuous: it is 0 at the origin and then becomes a finite, large number as soon as we leave the origin). Therefore writing $\text{V}[\log m_{it}] \approx \frac{1}{\lambda_{it}}$ for λ_{it} going to 0 cannot be correct.

In order to understand in what situations correspond to "small" expected values of deaths, we perform an exercise similar to the one we have done for λ large. We start with the same large country as above, (total population 280,000,000), and consider a cause of death not as common as cardiovascular disease, for example homicide, in

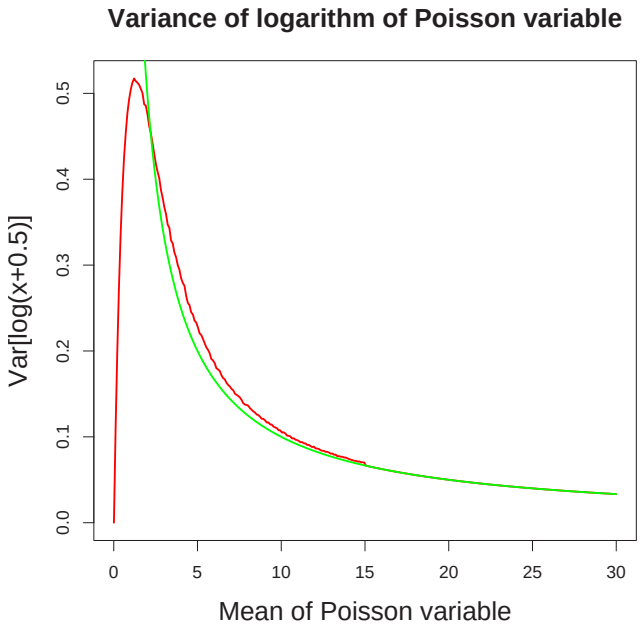

Figure 6.11: Approximating the Variance of the logarithm of a Poisson Variable: In red we report $\mathrm{V}[\log(d_{it} + 0.5)]$ for $d_{it} \sim \text{Poisson}(\lambda_{it})$, as a function of λ_{it}. In green we report the function $\log(1 + \frac{1}{\lambda_{it}})$, the variance of the logarithm of the number of deaths, as a function of λ_{it}, when the Poisson and the lognormal density are close to each other and there is no 0.5 additional term.

the same age groups as before, that is 70-74. If we consider the male population, a reasonable value for λ_{it} is 135, and for p_{it} is 4,000,000, which corresponds to $\mathrm{E}[M_{it}] = 3.4 \times 10^{-5}$. Let us now scale this country down by a factor of 45, keeping mortality the same: this makes the total population approximately 6,200,000, with $\lambda_{it} = 3$ and $p_{it} \approx 88,000$. For such a country we would expect to see a number of zero observed deaths, and we would have to add 0.5 deaths to each observation if we wanted to retain the normal specification for log-mortality.

Again, we compare these calculations to real data: In the year 2000, two countries with population in the age group 70-74 were close to 88,000: Denmark and Finland, and both report some zeros for the number of deaths. The total population of both countries is around 5.1 million. For Finland the average number of deaths over the last 10 years has been 2.1, while for Denmark it has been 0.7, and in both cases the number of zeros observed in the time series is consistent with expected deaths of that size (ignoring the downward time trend).

3. Very Small Value of λ_{it} This corresponds to situations with many number of observations where $d_{it} = 0$. This is likely to happen when λ_{it} is smaller than 2 or 3. In these cases, the data contain very little information and assigning 0.5 deaths to each observation could be highly distortive. For these cases the problem is not so much the correct specification of the density, but the paucity of data: Absent prior information, it is not clear that any sort of meaningful statistical inference can be performed on the data. Since we do have prior information, we use it to deal with these cases with an appropriate pre-processing imputation stage. Whenever zeros are found in the data, we fill them in with values borrowed from nearby cross-sections and nearby points in time. An alternative to this pre-processing is simply to consider the zero values as missing and let the prior take over, although this risks selection bias. We use the pre-processing approach mostly because of implementation issue: Mortality data have the convenient feature that if the number of deaths is observed in one age groups it is observed in all age groups, and we use this feature in our implementation.

This regime is common when studying very small countries, such as islands in the Pacific. In these cases it can easily happen that an entire age profile is 0, even for causes of death which are not very rare. Obviously adding 0.5 deaths would be wrong in these cases and the distinction between an entirely zero age profile and entirely missing mortality rate is very small. Less dramatic cases occur in small countries such as Honduras or Nicaragua for causes of death such as breast cancer. For example in Nicaragua, in the period before 1980, in every year the age profiles only had 3 or 4 non-zero observations, with the typical value hovering around $d_{it} = 3$. Since the shape of the age profile for breast cancer is fairly well known, it is not difficult to use the non-zero observations to fit a reasonable age profile, and therefore impute the observations corresponding to the zero values.

In our applications, We have not found reasons to give up the computational

advantages of the normal specification to use a Poisson or negative binomial specification. This is especially true since our primary interest is in the forecast point estimate. We have seen in this section, and confirmed in our experiments, that when the expected number of deaths λ_{it} is large, the Poisson specification does not help, and when λ_{it} is small but not very small adding 0.5 deaths to each observation does not cause enough error to justify changing the specification. In short, once 0.5 deaths are added to each observation, Equation 6.29 is a reasonable choice.

The results of this section leave open the possibility that instead of a Poisson density a different density should be used as starting point for evaluating our approximation. In particular, a different density could allow the variance of the number of deaths to be less tightly tied to the expected value λ_{it}. We address this issue by using a specification for the variance slightly more general than the one suggested by Equation 6.29, a topic which we discuss in the next section.

6.5.3 Variance Specification

If the normal specification in Equation 6.29 is correct we would expect to see the variance of log-mortality to be inversely proportional to the expected number of deaths. Since for every year and cross-section we only have one observation, we cannot test this hypothesis directly. The value of λ_{it} could be approximated with the observed value d_{it}, but we cannot do the same to get the variance of m_{it}, for which we need at least two observations. The problem obviously is that the random variables involved are not stationary. A quick way around that is to assume that they are "temporarily" stationary, so that we can assume m_{it} and $m_{i,t+1}$ are drawn from the same distribution. In this case we take their average absolute differences as an estimate of the standard deviation of m_{it}, which according to the model should be $\frac{1}{\sqrt{\lambda_{it}}} \approx \frac{1}{\sqrt{d_{it}}}$. Therefore to check how well the following relationship holds, we make this comparison:

$$\frac{1}{T}\sum_{t} |m_{i,t+1} - m_{it}| \approx \frac{1}{T}\sum_{t} \frac{1}{\sqrt{d_{it}}} \tag{6.30}$$

where we are averaging over time in order to reduce the estimation variance in computing the standard deviation.

We offer some illustrative examples for cardiovascular disease in men, in Figure 6.12, and breast cancer in women, in Figure 6.13. Each of the four graphs in each Figure is specific to a country, cause of death, and gender. Each graph plots the left side of Equation 6.30, on the vertical axis, against its right side, where the cross-sectional index i varies over 17 age groups. A red line is drawn at 45 degrees, where the points should be if the relationship in Equation 6.30 held exactly (and without measurement error). Note that the vertical axis, which determines the meaning of specific deviations from the 45 degree line, is different in each graph.

In some of these graphs, such as for cardiovascular disease in Australian males and for breast cancer in females in Italy and Malta, the relationship holds quite well,

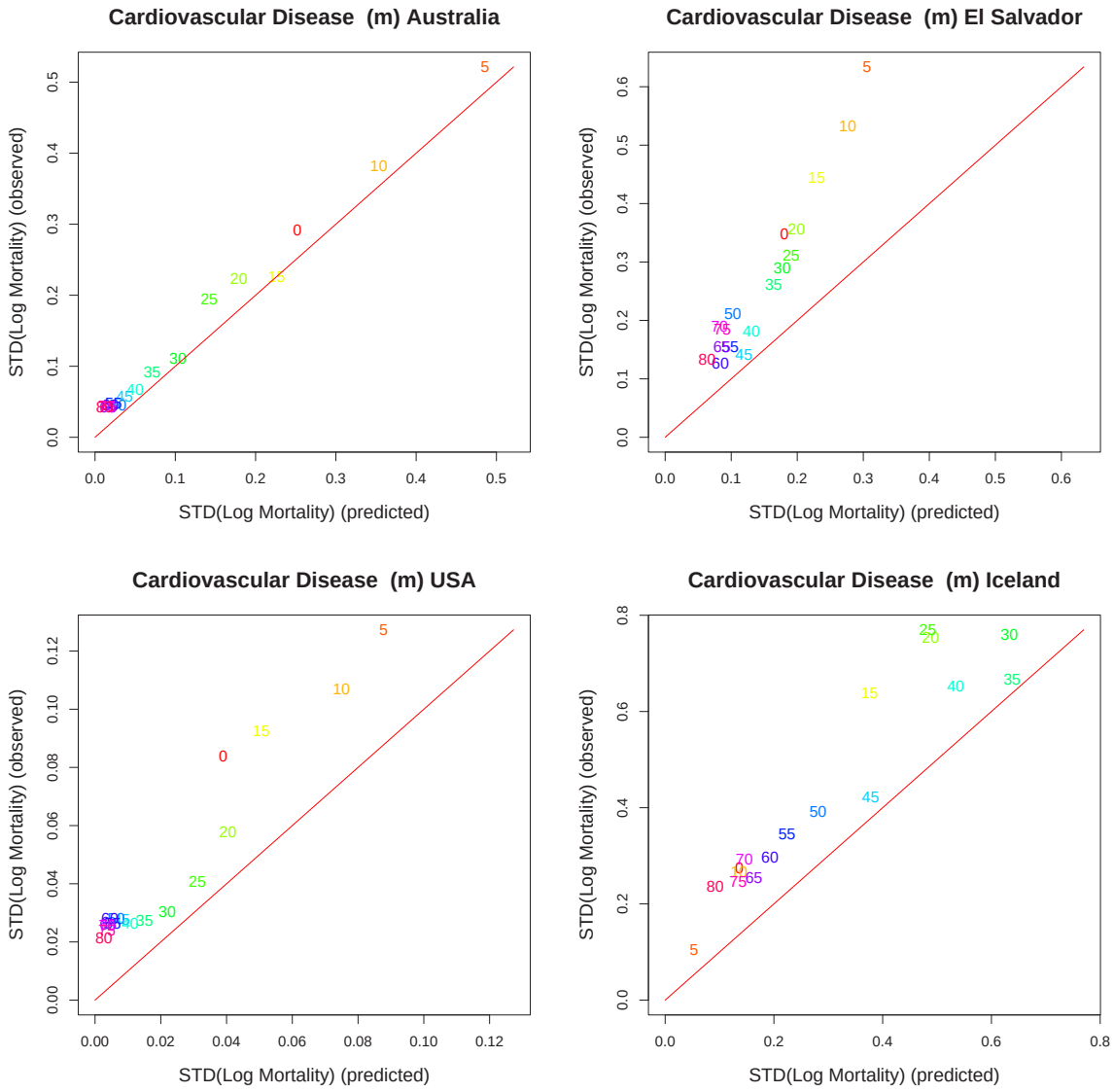

Figure 6.12: The Log-Normal Variance Approximation for Cardiovascular Disease in Men. The left and right sides of Equation 6.30 are plotted against each other for 17 age groups. Deviation from the 45 degree line indicate countries and diseases where the approximation holds less well.

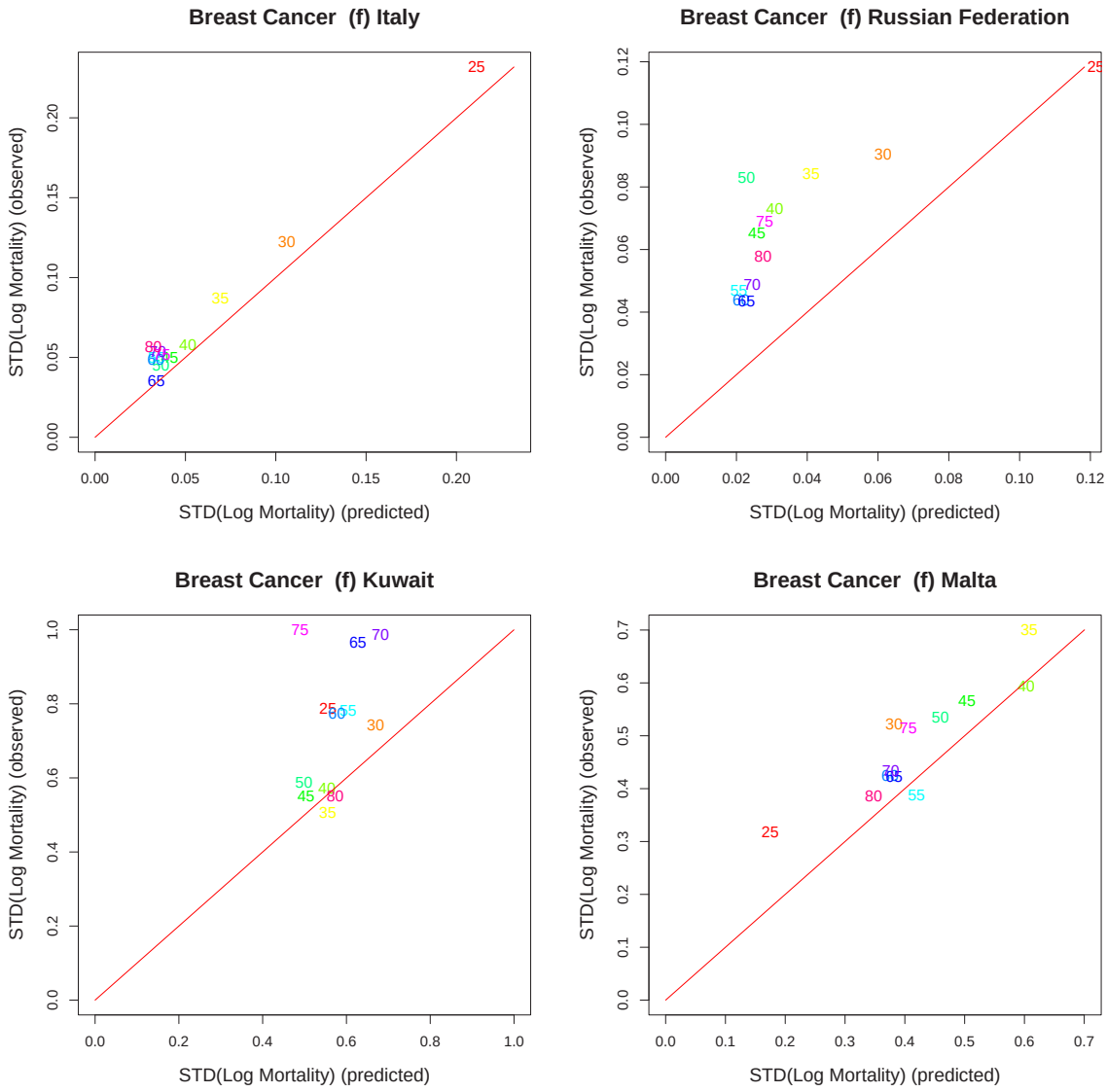

Figure 6.13: The Log-Normal Variance Approximation for Breast Cancer in Women. The left and right sides of Equation 6.30 are plotted against each other for 17 age groups. Deviation from the 45 degree line indicate countries and diseases where the approximation holds less well.

especially considering the approximation involved. In other cases, like in the U.S. and Iceland for cardiovascular disease in males, the relationship holds qualitatively, in the sense that the pattern is correct. In cases like breast cancer in Kuwait it is hard to say, since the points are so dispersed relative to the narrow range in which mortality varies over ages. Examples like El Salvador are clearly violations, but in practice, we let our variance approximation be a scalar multiple of the true variance, which means that deviations from the 45 degree line that fall around a line through the origin, such as in El Salvador, fit well. We could also introduce an additive scalar correction, to allow for linear deviations that do not pass through the origin, but we have not found this necessary.

After having looked at many different countries and many different causes of death the tentative conclusion we have drawn is that when the number of deaths is not too small it is qualitatively true that the variance of log-mortality decreases with the expected number of deaths, although the exact functional form may vary. Going back to the question posed at the beginning of this section: How do we translate this knowledge into a prior for the standard deviations?

Begin by assuming the usual linear specification for the expected value of log-mortality $\mu_{it} = \mathbf{Z}_{it}\boldsymbol{\beta}_i$ and that λ_{it} is large enough so that Equation 6.29 is valid. Then we can identify the expected value of log-mortality with the logarithm of the expected value of mortality, and therefore set $\lambda_{it} = p_{it}\exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)$. If we truly believed in this model we would then have a different standard deviation σ_{it} for each observation. A way to capture the inverse proportionality between σ_{it} and λ_{it} would be to make the σ_{it} correlated with the regression coefficients $\boldsymbol{\beta}_i$, and write $\mathcal{P}(\sigma^2 \mid \boldsymbol{\beta}) = \prod_{it}\mathcal{P}(\sigma_{it}^2 \mid \boldsymbol{\beta}_i)$, and set $\mathcal{P}(\sigma_{it}^2 \mid \boldsymbol{\beta}_i)$ to some density whose mean value is $\frac{1}{\lambda_{it}} = \frac{1}{p_{it}}\exp(-\mathbf{Z}_{it}\boldsymbol{\beta}_i)$. An obvious choice consists in setting $\sigma_{it} = \frac{\sigma_{it}^*}{\sqrt{\lambda_{it}}}$ where σ_{it}^* is a random variable whose expected value is 1. An extreme case of this model consists in setting $\sigma_{it}^2 = \frac{1}{p_{it}}\exp(-\mathbf{Z}_{it}\boldsymbol{\beta}_i)$.

There are at least two problems with this approach. First, having σ and $\boldsymbol{\beta}$ correlated is computationally complicated, and would lead to a fairly slow implementation in term of Monte Carlo Markov Chains. Second, having one standard deviation for each observation leads to an unreasonable number of parameters.

Thus, we first modify this approach by removing the dependency of the standard deviations on time, and hence writing σ_i instead of σ_{it}. This step is reasonable because the variation over time is small relative to the variation over cross-sections (age groups in particular). Second, we model σ_i as inversely proportional to some average historical level of the number of deaths for cross-section i, which we assume to be known a priori and denote by $\bar{\lambda}_i$. If we think of $\bar{\lambda}_i$ as a number which sets the order of magnitude of σ_i, then we model the uncertainty around σ_i by writing $\sigma_i = \frac{\sigma_i^*}{\sqrt{\bar{\lambda}_i}}$ and taking σ_i^* to be a random variable with mean value 1. This model still leads us to introduce $C \times A$ random variables, which is large in some applications. We have also found it useful to reduce further the number of parameters by allowing

σ_i^* to vary only by age, although this latter choice is grounded in our experience and not in any theory that guarantees that it will hold in other applications.

Summarizing, our variance specification has the form:

$$\sigma_{ca} = \frac{\sigma_a^*}{\sqrt{\bar{\lambda}_{ca}}} \ , \quad \mathrm{E}[\sigma_a^*] = 1 \tag{6.31}$$

The presence of $\bar{\lambda}_{ca}$ in the variance has a flavor similar to empirical Bayes. In fact, it is not reasonable to expect that we can elicit estimates of these quantities directly from experts, and some method which uses data needs to be used. A straightforward implementation would use the average historical value of number of deaths as an estimate of $\bar{\lambda}_{ca}$, or the average of the predicted values of an LS regression. This we find too data-dependent. In order to remain close in spirit to the rest of the book, we choose to borrow the value of $\bar{\lambda}_{ca}$ from neighboring cross-sections, using the same weights we use in the prior for the regression coefficients. Although not entirely satisfactory, this approach seems to be a good compromise between practicality and theory.

Chapter 7

Adding Priors Over Time and Space

We now extend our results for generating priors to priors defined over sets of cross-sections defined over indices other than discretized continuous variables like age groups. We model prior knowledge of the expected value of the dependent variable and the extent to which it varies smoothly over time. We consider situations where we have prior knowledge about how the time trend of the expected value of the dependent variable, rather than the value itself, varies smoothly across cross-sections. We also allow more general interactions, such as if the age profile of mortality varies smoothly over time and this pattern varies smoothly across neighboring countries.

Mathematically, this Chapter explicitly extends the model for cross-sections labeled by indices which vary over sets that are continuous in nature but discretized (like a set of age groups or income brackets), to point continuous without discretizing (like time or distance from the population center), to variables composed of discrete sets with no metric structure (such as a list of countries, diseases, or ethnic groups).

7.1 Smoothing over Time

Another form of prior knowledge we are likely to have, and indeed do have in our running example, is that the expected value of the dependent variable μ_{it} varies smoothly over time.[1] Since time is a continuous variable we can use the same reason-

[1]We might also have more specific knowledge, for example that μ_{it} decreases monotonically over time, but this is more difficult to deal with a priori because we would need to specify the degree of drop in μ, which is less clearly known a priori. Two easier, if less fully satisfactory, ways to incorporate this type of information could be used. One would be to include time as a covariate (as is often done in mortality studies as a rough proxy for technology) and to put a prior directly on its coefficient. Another possibility is to use the prior in this section and to make forecasts but to truncate the posterior via rejection sampling to ensure the desired pattern. However, we find in

ing we developed for smoothing over age, in Section 5.2. Hence, denoting by i as a generic cross-sectional index, we use an analogous smoothness functional to smooth over time:

$$H[\mu, \theta] \equiv \frac{\theta}{N} \sum_i \int_0^T dw^{\text{time}}(t) \left(\frac{d^{\mathfrak{m}} \mu(i,t)}{dt^{\mathfrak{m}}} \right)^2, \tag{7.1}$$

where N is the total number of cross-sectional units and the measure $dw^{\text{time}}(t)$ allows us to weight some time periods more than others (for example we could use it to exclude a time period in which we know that our smoothness assumptions do not hold, like the time of an epidemic or a war; see Section 8.4). The discretization of Equation 7.1 works exactly as the discretization of the smoothness functional over age groups, so we do not report it here. The resulting implied prior for $\boldsymbol{\beta}$ is

$$H^{\mu}[\boldsymbol{\beta}, \theta] = \frac{\theta}{N} \sum_i \boldsymbol{\beta}_i' \mathbf{C}_{ii}^{\text{time},n} \boldsymbol{\beta}_i \tag{7.2}$$

where we have defined the matrix:

$$\mathbf{C}_{ii}^{\text{time},n} \equiv \frac{1}{T} \left(\frac{d^{\mathfrak{m}} \mathbf{Z}_i}{dt^{\mathfrak{m}}} \right)' \left(\frac{d^{\mathfrak{m}} \mathbf{Z}_i}{dt^{\mathfrak{m}}} \right) \tag{7.3}$$

While Equation 7.2 is mathematically similar to Equations 4.18 (Page 81) and 7.6, it differs in a substantively important way. One way to see this is that Equation 7.2 contains no interactions among any cross-sections, so that for example a random permutation of the cross-sectional index will leave this expression unchanged. From a probabilistic point of view this means that the coefficients $\boldsymbol{\beta}_i$ are *independent* (not i.d.) random variables, while the whole point of smoothing over age groups in Equation 4.18 (Page 81) and countries in Equation 7.6 is precisely that the $\boldsymbol{\beta}_i$ are *dependent* in specific interesting ways.

The obverse is that Equations 4.18 (Page 81) and 7.6 are insensitive to any temporal behavior of the covariates, since time enters into those equations only through the product $\mathbf{Z}_i' \mathbf{Z}_j$: a random permutation of the time index will leave this quantity unchanged. In contrast, the whole point of Equation 7.2 is to take into account the temporal behavior of the covariates, since it explicitly incorporates the time derivatives of the covariates.

7.1.1 Prior Indifference and the Null Space

The smoothness functional in Equation 7.1 is a standard smoothness functional, of the type discussed in Chapter 5. Therefore, in terms of μ the null space contains profiles of log-mortality which, in each cross-section, evolve over time as polynomials

practice that these steps are unnecessary since the likelihood contains plenty of information about the downward trend.

in t of degree $\mathtt{n}-1$. What happens when we project on the subspace spanned by the covariates? Since the smoothness functional in Equation 7.1 simply sums over the cross-sections, the null space of the prior can be studied independently for each cross-section, and so for simplicity in the following we assume that there is only one cross-section, that we denote with the index i.

Restricted to the subspace defined by our covariates and linear functional form, the null space is simply the null space of the matrix $\mathbf{C}_{ii}^{\text{time},n}$. Since for any matrix V we know that V and $V'V$ share the same null space, the null space of $\mathbf{C}_{ii}^{\text{time},n}$ in Equation 7.3 is simply the null space of $\frac{d^{\mathtt{n}}\mathbf{Z}_i}{dt^{\mathtt{n}}}$. Generic covariates, such as GDP or tobacco consumption, are not perfectly correlated, and it is reasonable to expect that their time derivatives are also linearly independent. Therefore if the data matrix only had covariates of this type the matrix $\frac{d^{\mathtt{n}}\mathbf{Z}_i}{dt^{\mathtt{n}}}$ would have full rank, the null space would be trivial, and the prior would not be indifferent to any pattern. Therefore the structure of the null space would be lost going from the space of μ to the space of the coefficients, which would be unfortunate. Fortunately, the covariates will usually include the constant, and probably time: this allows the matrix $\frac{d^{\mathtt{n}}\mathbf{Z}_i}{dt^{\mathtt{n}}}$ not to be full rank. The best way to see what the null space would be is through successive examples, which we order in terms of increasing complexity.

1. Suppose we have only one age group and one country but multiple time periods. If $\mathtt{n}=1$ then only constant levels are in the null space on the scale of μ. If only a constant term is included in the covariate matrix $\mathbf{Z}$, then after the prior is restricted to the subspace defined by the covariates and our functional form $\mathbb{S}_{\mathbf{Z}}$, all patterns are in the null space. That is, since the prior can only affect the constant term, and the constant term can have no effect on the smoothness of μ_t over time, the prior has no parameters to adjust to achieve smoothness and it will do nothing. Thus, in this situation, the prior will have no effect on the empirical results, which is equivalent to a likelihood-only analysis, or a Bayesian analysis with an improper uniform prior.

2. If we continue with the first example, but change $\mathtt{n}=2$, then the null space for μ includes constant shifts as well as changes in the slope of μ_t over time. However, since the covariates still only include the constant term, the prior will have no effect on the constant term or the empirical estimates. So nothing changes from the first example.

3. If $\mathtt{n}=1$, and $\mathbf{Z}$ includes a constant term and GDP, then the null space for μ, and after restriction to the subspace $\mathbb{S}_{\mathbf{Z}}$, includes only constant shifts. This means that the prior will have no effect on the constant term in the regression. The prior smooths expected log-mortality in this example by requiring the squared first derivative with respect to time to be small. However, the only way the prior can have an effect such as this is by affecting the coefficient on GDP. If GDP varies a lot over time, then this prior can only impose smoothness by reducing the size of its coefficient.

4. Continuing with the previous example, but changing the degree of smoothness to $\mathbb{m} = 2$, the null space in μ becomes larger: The prior would now be indifferent to both changes in levels and slopes of μ_t over time. However, the null space restricted to $\mathbb{S}_{\mathbf{Z}}$ is the same as the previous example because patterns linear in time are not in the span of the covariates (unless GDP happened to be exactly linear in time). The nonnull space has changed from the previous example since the prior now penalizes the second derivative of GDP. In other words, the prior is now sensitive to, and tries to smooth, μ only as affected by the nonlinear portions of GDP. It will do this by reducing the size of the coefficient on GDP. (The fact that GDP may be nearly linear is immaterial, since any nonlinearities are enough to let the prior use its coefficient to achieve the desired degree of smoothness.)

5. Continuing with the previous example, suppose we add a time trend to the constant and GDP in $\mathbf{Z}$. Since $\mathbb{m} = 2$, the null space on the scale of μ includes shifts in both the level and slope, as before. Since the covariates are sufficiently rich to represent these patterns, the null space restricted to $\mathbb{S}_{\mathbf{Z}}$ also includes level and slope shifts. In this example, the prior would then only have an effect only on the coefficient of GDP. This coefficient is adjusted by the prior to keep μ smooth, and so it would be reduced if the second derivatives of this variable were large. The constant and slope on the linear trend are unaffected by the prior.

7.2 Smoothing over Countries

When Coale and Demeny developed their now widely used model life tables, they began with 326 male and 326 female mortality age profiles, and reduced them to 192 tables by discarding those with apparent data errors (judged from large male-female deviations). They then classified these age profiles inductively into following four distinct patterns. When they examined which countries fell in each category, they found that the countries in each of the four categories were geographically clustered (Coale and Demeny, 1966). As is widely recognized in building life tables by hand, such as when filling in mortality age patterns in countries with missing data, "inferences are often drawn from the mortality experienced by neighboring countries with better data. This borrowing is made on the assumption that neighboring countries would have similar epidemiological environments, which would be reflected in their cause of death distributions and hence their age patterns of mortality" (Preston, Heuveline and Guillot, 2001, p.196). We now use this generalization about mortality patterns to develop priors that smooth over countries or other geographic areas, borrowing strength from neighbors to improve the estimation in each area.

In this section, we consider the case where the cross-sectional index i is a label that does not come with a continuous structure naturally associated with it. In this

case an appropriate mathematical framework to describe smoothness is graph theory. To keep things simple, we proceed here intuitively, leaving the formal connection to graph theory and our precise definitions to Appendix E.

To fix ideas we focus on the case in which i is a country index (and we have only one age group), so that $i = c$, $c = 1, \ldots, C$ and the expected value of the dependent variable is a matrix with elements μ_{ct} (where time is treated as a discrete variable). We assume the following prior knowledge: At any point in time the expected value of the dependent variable varies smoothly across countries. That is, it has the tendency to change less across neighboring countries than between countries that are far apart.

The only ingredient we need to build a smoothness functional in this case is the notion of a "neighbor", which can be based on contiguity, proximity, similarity, or the degree to which the people in any two countries interact. This is easily formalized by introducing the symmetric matrix s^{cntry}, whose positive elements $s^{\text{cntry}}_{cc'}$ are "large" only if c and c' are countries which are "neighbors", that is countries for which we have a priori reasons to assume that the expected value of the dependent variable takes similar values. In full analogy with Section 5.2 we write a smoothness functional of the form:

$$H[\mu, \theta] = \frac{\theta}{2T} \sum_{cc't} s^{\text{cntry}}_{cc'} (\mu_{ct} - \mu_{c't})^2 \tag{7.4}$$

Smoothness functionals of this type are common in applications of Markov Random Fields in different disciplines, from agricultural field experiments (Besag and Higdon, 1999) to computer vision (Geman and Geman, 1984; Besag, 1986; Li, 1995). Defining, as usual the matrix $W^{\text{cntry}} = (s^{\text{cntry}})^+ - s^{\text{cntry}}$, we can rewrite the functional in Equation 7.4 as

$$H[\mu, \theta] = \frac{\theta}{T} \sum_{cc't} W^{\text{cntry}}_{cc'} \mu_{ct} \mu_{c't}. \tag{7.5}$$

Now that the prior is in a form similar to Equation 5.12 (Page 100), we repeat the steps of Section 5.2 to derive the prior for the coefficients $\boldsymbol{\beta}$. Plugging the specification $\mu_{ct} = \mathbf{Z}_{ct}\boldsymbol{\beta}_c$ in Equation 7.5 we obtain, predictably:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{cc'} W^{\text{cntry}}_{cc'} \boldsymbol{\beta}'_c \mathbf{C}_{cc'} \boldsymbol{\beta}_{c'} \right), \tag{7.6}$$

where we have defined the matrix

$$\mathbf{C}_{cc'} \equiv \frac{1}{T} \mathbf{Z}'_c \mathbf{Z}_{c'} \,,$$

and $\mathbf{Z}_c$ is the usual data matrix for country c, that is the matrix whose rows are the vectors $\mathbf{Z}_{ct}$. The key here is the perfect correspondence between Equation 7.6 and Equation 5.16 (Page 108): The only difference is that while for the prior over age

groups the matrix $W^{\text{age},\mathbb{m}}$ was determined easily by the choice of the scalar $\mathbb{m}$, for the matrix W^{cntry} here we have to do more work and build the adjacency matrix s^{cntry} by hand, using experts' opinions to figure out which countries should be considered neighbors. Mathematically, then the two forms are the same. The only difference is due to the substantive differences between the two problems.

7.2.1 Null Space and Prior Indifference

The smoothness functional of Equation 7.5 defines a prior density for μ through the relationship:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_t \mu_t' W^{\text{cntry}} \mu_t\right) \tag{7.7}$$

where μ_t is the $C \times 1$ vector with elements μ_{ct}. By definition, the rows and columns of W^{cntry} sum up to 0, and therefore W^{cntry} is singular. If the adjacency matrix s^{cntry} has been built in such a way that it is possible to go from one country to any other country traveling from neighbor to neighbor (that is there is only one continent and no "islands"), then one can show that W^{cntry} has only one zero eigenvalue (Biggs, 1993). Therefore we have:

$$\text{rank}(W^{\text{cntry}}) = C - 1 \ , \quad \text{nullity}(W^{\text{cntry}}) = 1$$

The null space of W^{cntry} is simply the one-dimensional space of constant vectors, and the prior 7.7 is indifferent with respect to the transformation:

$$\mu_{ct} \rightsquigarrow \mu_{ct} + f_t \ , \quad \forall f_t \in \mathbb{R}.$$

Therefore, while we know something about how the dependent variable μ varies from one country to the next, we are totally ignorant about the absolute levels it may take.

Suppose that the adjacency matrix s^{cntry} is built with "islands," so within each group it is possible to go from one country in one group (or "island") to every other country in that island by traveling from neighbor to neighbor; however, it is not possible to go from any country in one island to any country in another island. In this situation, each island adds an extra zero eigenvalue to W^{cntry}, and thus increases its nullity by one. The null space of W^{cntry}, and hence the prior in Equation 7.7, is indifferent to a different constant shift for all countries c included in *each* island $j(c)$:

$$\mu_{ct} \rightsquigarrow \mu_{ct} + f_{j(c),t} \ , \quad \forall f_{j(c),t} \in \mathbb{R}.$$

Although using islands to add flexibility to the prior and expand the null space in this way can be very useful in practice, such datasets can be analyzed separately for the group of countries on each island. As such, we only analyze the case with no

islands (i.e., one world island) in the rest of this section. Obviously, the same result apply separately and independently within each island.

Since there are T time periods and therefore T independent choices of the values f_t the null space is a T-dimensional subspace, consisting of a log-mortality profile which is constant across countries and that evolves arbitrarily over time. When we add to the prior 7.7 the information coming from the specification $\mu_{ct} = \mathbf{Z}_{ct}\boldsymbol{\beta}_c$, however, the structure of this subspace will be altered: the time evolution of log-mortality is now determined by the covariates, and we will not be able to produce patterns of log-mortality with arbitrary behaviors over time (and constant across countries).

More precisely, the null space of the prior as determined by the coefficients $\boldsymbol{\beta}$ will be the intersection of the null space of the prior 7.7 with the subspace $\mathbb{S}_{\mathbf{Z}}$ defined by the specification $\mu_{ct} = \mathbf{Z}_{ct}\boldsymbol{\beta}_c$. Excluding pathological combinations of the covariates, this implies that all the covariates which are country-specific must have zero coefficients in the null space. In other words, to get the μ's to be similar, the prior will reduce the value of the coefficients with $\mathbf{Z}$'s that vary over countries.

Suppose now there exist k covariates $z_t^{(1)}, \dots z_t^{(k)}$ which are the same across all the countries, such as the constant and time: Then, for each, we can set the corresponding coefficient equal to an arbitrary country independent constant, obtaining a log-mortality profile which is constant across countries and that evolves over time as $z_t^{(k)}$. Therefore the null space of the prior on the scale of $\boldsymbol{\beta}$ in Equation 7.6 is k-dimensional and can be described as follows:

$$\mu_{ct} = b_1 z_t^{(1)} + b_2 z_t^{(2)} + \cdots + b_k z_t^{(k)} \qquad\qquad b_1, \dots b_k \in \mathbb{R}$$

In practice it is likely that the only covariates which are common to all countries are time and the constant. Therefore the null space will consists of patterns of the form $\mu_{ct} = b_1 + b_2 t$, for any b_1 and b_2. In terms of the regression coefficient $\boldsymbol{\beta}$ this implies that if we add arbitrary numbers b_1 and b_2 to the coefficients of the constant and the time covariates the prior remains constant. Therefore, when it comes to these coefficients, the prior carries information only about their relative levels.

7.2.2 Interpretation

In Chapter 5, we considered the case of smoothness functionals for functions of variables which are continuous in principle, although discrete in practice. In this case the key ingredient was the possibility of using the derivative of order $\mathtt{n}$ as a measure of local variation. We also saw that priors written in terms of derivatives, as in Equation 5.10, could be written, once discretized, as in Equation 5.14 (Page 105), a form which we used as a starting point for the prior in Equation 7.4. We noticed that when the derivative in Equation 5.10 is of order 1 the weights $s_{aa'}^{\mathrm{age},\mathtt{n}}$ which connect one age group to another should be positive. Since the weights in Equation 7.4 $s_{cc'}^{\mathrm{cntry}}$ are, by construction, positive, it is natural to ask whether the expression 7.4 can be related to some notion of first derivative with respect to the country label. We now show

that this is indeed the case, by giving an intuitive description and leaving the details to Appendix E

The derivative is a measure of local variation, and therefore if we want to define the derivative of μ_{ct} with respect to the country variable, at the point c, we start by simply collecting in one vector $\nabla^c \mu_{ct}$ the differences $\mu_{ct} - \mu_{c't}$ for all countries c' which are neighbors of c (the superscript c stands for country and is not an index):

$$\nabla^c \mu_{ct} \equiv (\mu_{ct} - \mu_{c_1 t}, \dots \mu_{ct} - \mu_{c_n t}) \quad c_1 \dots c_n \text{ neighbors of } c$$

The sign of these differences is irrelevant at this point, since we will square them at the end. As the notation suggests, we think of $\nabla^c \mu_{ct}$ as the gradient of μ_{ct} with respect to the country label, although this quantity, unlike the usual gradient, is a vector of possibly different lengths at c and c', depending on the local neighborhood structure. We now verify that this notion of gradient is useful. In the case of continuous variables, like age (a), we obtain a smoothness functional by taking the derivative of a function at a point a, squaring it and integrating over a. Let us do the same with the discrete variable c. Thus, we "square the derivative at a point" by simply taking the squared Euclidean norm of $\nabla^c \mu_{ct}$ at the point c, which we denote by $\|\nabla^c \mu_{ct}\|^2$, and integrate by summing this quantity over all the countries. The resulting candidate for the smoothness functional is

$$H[\mu, \theta] \equiv \frac{\theta}{2T} \sum_{ct} \|\nabla^c \mu_{ct}\|^2, \tag{7.8}$$

where the factor $\frac{1}{2}$ is included to avoid double counting (the difference $\mu_{ct} - \mu c' t$ appears both in the gradient at c and in the gradient at c'). It is now easy to verify that the expression above is indeed a smoothness functional, and in fact it is the same smoothness functional of Equation 7.4:

$$H[\mu, \theta] = \frac{\theta}{2T} \sum_{ct} \|\nabla^c \mu_{ct}\|^2 = \frac{\theta}{2T} \sum_{cc't} s^{\text{cntry}}_{cc'} (\mu_{ct} - \mu_{c't})^2. \tag{7.9}$$

This derivation of the smoothness functional does not add anything from a technical point of view. However, it allows to write a smoothness functional for a discrete variable using the same formalism we use for continuous variables, hence unifying two apparently different frameworks. This is useful for example for when we consider more complicated forms of smoothness functionals, combining derivatives with respect to ages and countries in the same expression. A limit of this formulation is that it does not provide an easy generalization of the concept of derivative of order higher than 1.[2]

[2]It is a trivial observation, that nevertheless will be useful later on, that the derivative of order 0 is always well defined, since it corresponds to the identity operator. This implies that the correct generalization of a smoothness prior with derivative of order 0 is obtained simply by setting $W^{\text{cntry}} = I$.

7.3 Smoothing Simultaneously over Age, Country and Time

With the tools developed in this chapter thus far, we can now mix, match, and combine smoothness functionals as we like. Here we report for completeness the result of using all three simultaneously, since this is what we often use in applications, and since the results will always have the same unified and simple form as that for all the other priors specified in this book. Although each component will have in practice its own smoothness parameter θ, and its own order of derivative $\mathtt{m}$, for simplicity of notation we assume that they share the same θ and $\mathtt{m}$. We adopt the continuous variable notation for age and time, so that $\mu(c,a,t)$ is the expected value of the dependent variable for country c, age a and at time t. Assuming the Lebesgue measure for the age and time, the smoothness functional is:

$$H[\mu,\theta] \equiv \frac{\theta}{CAT}\sum_c \int_0^A da \int_0^T dt \theta^{\text{age}} \left[\left(\frac{d^{\mathtt{m}}\mu(c,a,t)}{da^{\mathtt{m}}} \right)^2 + \theta^{\text{ctr}} \|\nabla^c \mu(c,a,t)\|^2 + \theta^{\text{time}} \left(\frac{d^{\mathtt{m}}\mu(c,a,t)}{dt^{\mathtt{m}}} \right)^2 \right] \tag{7.10}$$

Noitce that we introduced a redundant parametrization, in which the smoothness parameters associated to each smoothness functional (θ^{age},θ^{ctr} and θ^{time}) are multiplied by a common "scaling factor". This helps to maintain some of the notation consistent with other parts of the book, and can be useful in practice (for example one may determine a priori the relative weight of the 3 smoothness parameters and carry on in the Gibbs sampling only the global parameter θ).

Now it is just a matter of going through the exercise of the previous section while appropriately accounting for the indices c, a and t. The final result of our usual second step is the same prior expressed on the scale of $\boldsymbol{\beta}$, which is

$$H^{\mu}[\boldsymbol{\beta},\theta] = \theta \left[\sum_{caa'} \frac{\theta^{\text{age}}}{C} W^{\text{age},\mathtt{m}}_{aa'} \boldsymbol{\beta}'_{ca} \mathbf{C}_{ca,ca'} \boldsymbol{\beta}_{ca'} + \sum_{cc'a} \frac{\theta^{\text{ctr}}}{A} W^{\text{cntry}}_{cc'} \boldsymbol{\beta}'_{ca} \mathbf{C}_{ca,c'a} \boldsymbol{\beta}_{c'a} + \frac{\theta^{\text{time}}}{CA} \sum_{ca} \boldsymbol{\beta}'_{ca} \mathbf{C}^{\text{time},\mathtt{m}}_{ca,ca} \boldsymbol{\beta}_{ca} \right].$$

Fortunately this expression can be rewritten in a much simpler way using the following definitions and the multi-indices $i \equiv ca$ and $j = c'a'$:**FG check the following**

$$\begin{aligned} W_{ij} \equiv & W_{ca,c'a'} \equiv \frac{\theta^{\text{age}}}{C} W^{\text{age},\mathtt{m}}_{aa'} \delta_{cc'} + \frac{\theta^{\text{ctr}}}{A} W^{\text{cntry}}_{cc'} \delta_{aa'} \\ \mathbf{C}_{ij} \equiv & \frac{1}{T} \mathbf{Z}'_i \mathbf{Z}_j + \frac{\theta^{\text{time}} \delta_{ij}}{CATW_{ii}} \left(\frac{d^{\mathtt{m}}\mathbf{Z}_i}{dt^{\mathtt{m}}} \right)' \left(\frac{d^{\mathtt{m}}\mathbf{Z}_i}{dt^{\mathtt{m}}} \right) \end{aligned}$$

The final expression for the prior in terms of $\boldsymbol{\beta}$ is therefore:

$$H^{\mu}[\boldsymbol{\beta},\theta] = \theta \sum_{ij} W_{ij} \boldsymbol{\beta}'_i \mathbf{C}_{ij} \boldsymbol{\beta}_j \tag{7.11}$$

and as a result our prior for $\boldsymbol{\beta}$ is obtained by substituting Equation 7.11 in Equation 4.13:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) = K(\theta) \exp\left(-\frac{1}{2}\theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j\right), \tag{7.12}$$

where $K(\theta)$ is a normalization constant.

The remarkable feature of this expression is its simplicity: it is a normal prior, no more mathematically complex than the prior implied by the original prior on coefficients in Equation 4.5, and yet it embeds information about smoothness over ages, countries and time and on the scale of the dependent variable. In addition, while the weight matrix W^{cntry} has to be constructed by hand, the matrix $W^{\text{age},\mathtt{n}}$ is automatically determined once the integer $\mathtt{n}$ has been chosen. It follows from the construction in this section that the general structure of the prior in Equation 7.12 does not depend on the particular choice that we have made in Equation 7.10: We could easily add terms with mixed derivatives with respect to ages and times, or ages and countries, or triple term interactions that combine ages, countries and time, and still the final result would be of the type of Equation 7.12. We turn to this topic in the next section.

7.4 Smoothing Time Trend Interactions

Sections 5.2 and 7.2 allow researchers to specify smoothness priors on the expected value of the dependent variable across (possibly discretized) continuous variables like age and unordered nominal variables like country, respectively. In both cases, the priors operated directly on the *levels* of μ. Similarly, Section 7.1 enables researchers to specify smoothness priors on the time trend of the expected value of the dependent variable. In this section, we generalize these results to allow the priors to operate on the *time trends* in these variables, which is often very useful in applications.

7.4.1 Smoothing Trends over Age Groups

In addition, or as an alternative, to the smoothness functional for age in Equation 5.8 (Page 95) we now show how to allow the time *trend* of the expected value of the dependent variable to vary smoothly across age groups. For example, we often expect log-mortality to decrease at similar rates for all age groups (except possibly in infants). For example, most demographers would normally be highly skeptical of mortality forecasts for a country, sex, and cause, that trended upward for 25-year-olds but downward for 30-year-olds. In this case an appropriate smoothness functional can be obtained by replacing the level $\mu(a,t)$ in Equation 5.8 (Page 95) with the time derivative $\partial\mu(a,t)/\partial t$, and averaging over time as well:

$$H[\mu,\theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{\partial^{\mathtt{n}+1}\mu(a,t)}{\partial a^{\mathtt{n}} \partial t}\right)^2. \tag{7.13}$$

This is one of those cases in which having a measure $dw^{\text{age}}(a)$ could be very important, and so we have written it in explicitly. In particular, if μ is a log-mortality rate the smoothness assumption above does not hold well at very young ages, where mortality frequently drops at a faster rate — because of technological developments and political necessity — than in other age groups. In this situation, the measure $dw^{\text{age}}(a)$ should be defined so that younger ages are not penalized much, if at all, for having the rate of decrease of log-mortality differ from neighboring age groups. An extreme choice would be to set $dw^{\text{age}}(a) = 0$ for, say, $a < 5$ and a constant otherwise, although a smoother choice would probably be preferable.

With a smoothness functional like that in Equation 7.13, the prior for the coefficients $\boldsymbol{\beta}$ has exactly the same form as the one in Equation 4.19 (Page 82), the only difference being that the covariates should be replaced by their time derivatives, and therefore the matrices $\mathbf{C}_{aa'}$ should be replaced by

$$\mathbf{C}_{aa'}^{\text{time},1} \equiv \frac{1}{T}\left(\frac{d\mathbf{Z}_a}{dt}\right)'\left(\frac{d\mathbf{Z}_{a'}}{dt}\right). \tag{7.14}$$

Obviously time derivatives of order $\mathtt{m}_t > 1$ could be considered too, if desired, by simply replacing the matrix $\mathbf{C}_{aa'}^{\text{time},1}$ with a similarly defined matrix $\mathbf{C}_{aa'}^{\text{time},\mathtt{m}_t}$, in which the first derivative is replaced with the derivative of order $\mathtt{m}_t$.

7.4.2 Smoothing Trends over Countries

Just as we sometimes may want to smooth the trend of the expected value of the dependent variable across *age groups*, we may also want to do the same across *countries*. This is often a less restrictive but useful form of prior knowledge, which avoids us having to make statements about the levels of the expected value of the dependent variable. This is especially useful in situations where two countries, with different base levels of the dependent variable, pursue similar policies over time, or benefit from the same relevant technological advances.

By simply repeating the argument of Section 7.4.1, we can see that a prior corresponding to this kind of knowledge has the same form as the one implied by Equation 7.3, in which the covariates have been replaced by their time derivative, and therefore the matrices $\mathbf{C}_{cc'}$ have been replaced by:

$$\mathbf{C}_{cc'}^{\text{time},1} \equiv \frac{1}{T}\left(\frac{d\mathbf{Z}_c}{dt}\right)'\left(\frac{d\mathbf{Z}_{c'}}{dt}\right). \tag{7.15}$$

7.5 Smoothing with General Interactions

In this final section where we build priors, we give detailed calculations for a generic term involving a triple interaction of age, country and time. By setting appropriate

matrices equal to the identity, researchers will be then able to derive formulas for all the other pairwise interactions.

We begin with a smoothness functional of the form

$$H^{\mathtt{n}_a,\mathtt{n}_t}[\mu,\theta] \equiv \frac{\theta}{C}\sum_c \int_0^T dw^{\text{time}}(t)\int_0^A dw^{\text{age}}(a)\left\|\nabla^c \frac{\partial^{\mathtt{n}_a+\mathtt{n}_t}\mu(c,a,t)}{\partial a^{\mathtt{n}_a}\partial t^{\mathtt{n}_t}}\right\|^2. \tag{7.16}$$

where $dw^{\text{time}}(t)$ and $dw^{\text{age}}(a)$ are probability measures allowing one to impose different degrees of smoothness in different parts of the integration domain, and $\mathtt{n}_a$ and $\mathtt{n}_t$ are integers denoting the order of derivatives with respect to age and time respectively.

We first discretize the derivatives with respect to age and time:

$$\frac{\partial^{\mathtt{n}_a+\mathtt{n}_t}\mu(c,a,t)}{\partial a^{\mathtt{n}_a}\partial t^{\mathtt{n}_t}} \Rightarrow \mu'_{cat} \equiv \sum_{a't'} \mathrm{D}^{\mathtt{n}_a}_{aa'}\mathrm{D}^{\mathtt{n}_t}_{tt'}\mu_{ca't'}$$

where we use the notation μ' to remind us that this quantity is a derivative. Then we compute the gradient with respect to c of μ'_{cat} and square it:

$$\|\nabla^c \mu'_{cat}\|^2 = \sum_{c'} s^{\text{cntry}}_{cc'}(\mu'_{cat}-\mu'_{c'at})^2.$$

Now we sum this expression over c, a and t, weighting the sums over a and t with the weights w^{age}_a and w^{time}_t, which are the discrete versions of the probability measures $dw^{\text{age}}(a)$ and $dw^{\text{time}}(t)$:

$$H^{\mathtt{n}_a,\mathtt{n}_t}[\mu,\theta] = \frac{\theta}{C}\sum_{cat} w^{\text{age}}_a w^{\text{time}}_t \sum_{c'} s^{\text{cntry}}_{cc'}(\mu'_{cat}-\mu'_{c'at})^2 = \theta\sum_{cc'at} W^{\text{cntry}}_{cc'} w^{\text{age}}_a w^{\text{time}}_t \mu'_{cat}\mu'_{c'at}$$

where we define $W^{\text{cntry}} = C^{-1}[(s^{\text{cntry}})^+ - s^{\text{cntry}}]$ as in Section 7.2. Now we substitute the expression for μ'_{cat} and obtain:

$$H^{\mathtt{n}_a,\mathtt{n}_t}[\mu,\theta] = \theta\sum_{cc'at} W^{\text{cntry}}_{cc'} w^{\text{age}}_a w^{\text{time}}_t \sum_{a't'} \mathrm{D}^{\mathtt{n}_a}_{aa'}\mathrm{D}^{\mathtt{n}_t}_{tt'}\mu_{ca't'} \sum_{a''t''} \mathrm{D}^{\mathtt{n}_a}_{aa''}\mathrm{D}^{\mathtt{n}_t}_{tt''}\mu_{c'a''t''}.$$

Reshuffling the order of the sums we obtain:

$$H^{\mathtt{n}_a,\mathtt{n}_t}[\mu,\theta] = \theta\sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'}\left(\sum_a \mathrm{D}^{\mathtt{n}_a}_{aa'} w^{\text{age}}_a \mathrm{D}^{\mathtt{n}_a}_{aa'}\right)\left(\sum_t \mathrm{D}^{\mathtt{n}_t}_{tt'} w_t \mathrm{D}^{\mathtt{n}_t}_{tt'}\right)\mu_{cat}\mu_{c'a't'}$$

Defining the following matrices:

$$W^{\text{age},\mathtt{n}_a} \equiv (\mathrm{D}^{\mathtt{n}_a})'\text{diag}[w^{\text{age}}_a]\mathrm{D}^{\mathtt{n}_a} \qquad W^{\text{time},\mathtt{n}_t} \equiv (\mathrm{D}^{\mathtt{n}_t})'\text{diag}[w^{\text{time}}_a]\mathrm{D}^{\mathtt{n}_t} \tag{7.17}$$

we obtain:

$$H^{\mathtt{n}_a,\mathtt{n}_t}[\mu,\theta] = \theta\sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathtt{n}_a}_{aa'} W^{\text{time},\mathtt{n}_t}_{tt'}\mu_{cat}\mu_{c'a't'}$$

Now the prior for $\boldsymbol{\beta}$ can be obtained by simply substituting the specification $\mu_{cat} = \mathbf{Z}_{cat}\boldsymbol{\beta}_{ca}$ in the expression above, obtaining:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{n}_a}_{aa'} W^{\text{time},\mathbb{n}_t}_{tt'} \mathbf{Z}_{cat}\boldsymbol{\beta}_{ca}\mathbf{Z}_{c'a't'}\boldsymbol{\beta}_{c'a'}$$

By rewriting $\mathbf{Z}_{cat}\boldsymbol{\beta}_{ca}$ as $\boldsymbol{\beta}'_{ca}\mathbf{Z}_{cat}$ and changing the order of the sums we write:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'} W^{\text{cntry}}_{c'c''} W^{\text{age},\mathbb{n}_a}_{aa'} \boldsymbol{\beta}'_{ca} \left(\sum_{tt'} \mathbf{Z}'_{cat} W^{\text{time},\mathbb{n}_t}_{tt'} \mathbf{Z}_{c'a't'} \right) \boldsymbol{\beta}_{c'a'}$$

Now we define the matrix:

$$\mathbf{C}^{\mathbb{n}_t}_{ca,c'a'} \equiv \mathbf{Z}'_{ca} W^{\text{time},\mathbb{n}_t} \mathbf{Z}_{c'a'} \tag{7.18}$$

where $\mathbf{Z}_{ca}$ is the usual data matrix for cross-section ca, which has for each row vector $\mathbf{Z}_{cat}$. Using the $\mathbf{C}$ matrices defined above the smoothness functional for $\boldsymbol{\beta}$ simplifies to:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{n}_a}_{aa'} \boldsymbol{\beta}'_{ca} \mathbf{C}^{\mathbb{n}_t}_{ca,c'a'} \boldsymbol{\beta}_{c'a'} \tag{7.19}$$

Defining the multi-indices $i = ca$ and $j = c'a'$, and letting $W^{\mathbb{n}_a,\text{cntry}}_{ij} \equiv W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{n}_a}_{aa'}$, we simplify the expression above further and write it in the usual form:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{ij} W^{\mathbb{n}_a,\text{cntry}}_{ij} \boldsymbol{\beta}'_i \mathbf{C}^{\mathbb{n}_t}_{ij} \boldsymbol{\beta}_j \tag{7.20}$$

However, expression 7.19 probably the most useful form when deriving new smoothness functionals, since it allows researchers to plug into it the desired values and derive, as special cases, all the priors discussed in this book. Equation 7.20 will often be the form most useful for estimation.

Example Consider the problem of smoothing the time trend over age groups, as described in Section 7.4.1. The corresponding smoothness functional is a particular special case of Equation 7.16 in which, instead of the gradient with respect to the country variable we have the derivative of order 0, in which $\mathbb{n}_t = 1$ and $\mathbb{n}_a$ is arbitrary. As pointed out in Section 7.2.2, the derivative of order 0 with respect to countries corresponds to the choice $W^{\text{cntry}} = I$. Plugging these choices in Equation 7.19 we obtain the smoothness functional:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{caa'} W^{\text{age},\mathbb{n}_a}_{aa'} \boldsymbol{\beta}'_{ca} \mathbf{C}^{1}_{ca,ca'} \boldsymbol{\beta}_{ca'}$$

where, from Equation 7.18, we have defined:

$$\mathbf{C}^1_{ca,ca'} = \mathbf{Z}'_{ca} W^{\text{time},1} \mathbf{Z}_{ca'}$$

In order to compare with Equation 7.14 we need to consider the special case, considered in that section, of a uniform measure over time: $dw^{\text{time}}(t) = T^{-1}dt$. Substituting this choice for $dw^{\text{time}}(t)$ in Equation 7.17 we obtain:

$$W^{\text{time},1} = \frac{1}{T}(\mathrm{D}^1)'\mathrm{D}^1$$

Substituting this expression in the definition of $\mathbf{C}^1_{ca,ca'}$ above we obtain, as expected, the same expression of Equation 7.14:

$$\mathbf{C}^1_{ca,ca'} = \frac{1}{T}\mathbf{Z}'_{ca}(\mathrm{D}^1)'\mathrm{D}^1\mathbf{Z}_{ca'} = \frac{1}{T}(\mathrm{D}^1\mathbf{Z}_{ca})'(\mathrm{D}^1\mathbf{Z}_{ca'}) = \frac{1}{T}\left(\frac{d\mathbf{Z}_a}{dt}\right)'\left(\frac{d\mathbf{Z}_{a'}}{dt}\right).$$

⊠

7.6 Choosing a Prior for Multiple Smoothing Parameters

The smoothing parameter θ determines how much weight to put on the prior as compared to the data in the estimation and thus how smooth the forecasts will be. In Chapter 6, we showed that, when only one prior is being used, the only information needed to set θ is the average standard deviation of the prior. We also showed in Section 6.2.2 that setting the average standard deviation of the prior simultaneously set all the properties of the samples from the prior. When more than one prior is used, as should be the case in many applications, the reasoning of Chapter 6 still applies, although the implementation is more involved. We describe these procedures here.

Suppose we are using K priors, each with a smoothness parameter θ_k ($k = 1, \ldots, K$). If we only used a single prior k then we can use the result in Chapter 6 that θ_k is uniquely determined by the average standard deviation of the prior, which we denote σ_k (see Equation 6.14). Since the parameter σ_k is interpretable and uniquely determines θ_k, we use σ_k to parametrize our single priors. This implies the more general result that the expected value of any summary measure $F(\mu)$ is a function of the K parameters σ_k. Therefore, in order to estimate what the parameters σ_k should be, all we have to do is to find K summary measures F_k ($k = 1, \ldots K$), for which we have information about in terms of their expected values, and which we denote by $\bar{F}_k$. Then the values of the smoothness parameters are determined by solving the following system of equations:

$$\mathrm{E}_\perp[F_k(\mu) \mid \sigma_1, \ldots, \sigma_K] = \bar{F}_k \ , \quad k = 1, \ldots, K \tag{7.21}$$

With more than one prior it is not possible to solve these equations analytically even for summary measures which are quadratic in μ, and so numerical procedures must be employed. In addition, if more than K summary measures are available, it is advisable to use all of them. The system of equations (7.21) then become over-determined, and an approximate solution is required, but the advantage is that one gains a better insight into the properties of the prior.

We have found that the following summary measures are well suited for our purposes:

$$\begin{aligned}
\mathrm{SD}(\mu) &\equiv \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T}(\mu_{at}-\bar{\mu}_a)^2 \\
F_{\mathrm{age}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|\mu_{at}-\mu_{a-1,t}| \\
F_{\mathrm{time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|\mu_{at}-\mu_{a,t-1}| \\
F_{\mathrm{age/time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|(\mu_{at}-\mu_{a,t-1})-(\mu_{a-1,t}-\mu_{a-1,t-1})|
\end{aligned} \tag{7.22}$$

The summary measure SD is the average standard deviation of the prior, , which measures how much samples from the prior differ from the average age profile $\bar{\mu}$. Summary measure F_{age} measures how much log-mortality changes going from one age group to the next and F_{time} summarizes the changes in log-mortality from one time period to the next. Finally, $F_{\mathrm{age/time}}$ is a measure of how much the time trend changes from one age group to the next. Each of these quantities are easily interpretable, and with some clarification about what they mean we find that demographers and other experts often have have a reasonable estimate of their expected values. If such expert knowledge is not available, one can still get an idea of the expected values of these quantities using a procedure which has the flavor of empirical Bayes, and which we describe in the following section.

An important point is that for certain choices of $\bar{F}_i$, Equations 7.21 may have no solution. But instead of this posing a methodological problem, it indicates that the prior is unable to produce samples with the desired characteristics, or in other words that some of the expert's (or analyst's) choices were logically inconsistent. Learning about such logical inconsistencies can be helpful to an expert in making these choices. Furthermore, it is easy to imagine how this can happen: Suppose for example that one prefers a prior with a tiny overall standard deviation, but also one that allows wide variations in log-mortality from one year to the next, or from one age group to the next. These choices are clearly not compatible, and no set of prior parameters will produce such a result. Problems of this type are more likely to arise when relatively few covariates are being used, because samples from the prior are more constrained

in those cases. With more covariates, the prior has more coefficients to adjust to produce patterns consistent with a wider range of patterns.

For these reasons, we recommend, rather than trying to solve Equations 7.21 numerically, that analysts study the behavior of the expected values of the summary measures as a function of the parameters σ_k, and in particular the range of values that they can assume. This can be done, for example, using multiple scatterplots, as we illustrate below.

In general we recommend the following strategy for choosing appropriate values of σ_k:

1. Define at least N_s ($N_s \geq K$) summary measures for which expected value are approximately known;

2. Assign a reasonably wide range of variation to each σ_k (for example 0.01 to 2) and use it to define a grid in the space of the σ_k (for example each interval $[0.01, 2]$ could be divided in 5 sub-intervals).

3. For each combination of σs corresponding to a point in the grid, which we label γ, draw a large number of samples from the prior. Use these samples to compute numerically the implied expected value of the summary measures, which we denote by $\tilde{F}_i$. Store the results in a table whose rows have the structure:

$$(\sigma_1^\gamma, \ldots, \sigma_K^\gamma,\ \tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma)$$

4. Produce all pairs of scatterplots of the summary measures as a function of the σs and of each other. Qualitatively assess where the target values $\bar{F}_i$ of the summary measures fall in the scatterplots.

5. Define a distance $D(\cdot\ ;\ \cdot)$ in the space of the N_s summary measures and use it to find the combination of prior parameters $\hat{\sigma}_k$ which produces empirical values of the summary measures that are closest to the target. Formally, define

$$\hat{\gamma} \equiv \arg\min_\gamma D(\tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma; \bar{F}_1, \ldots, \bar{F}_{N_s})$$

   and then set $\hat{\sigma}_k = \sigma_{\hat{\gamma}}$.

This procedure depends on the choice of the distance measure $D(\cdot;\ \cdot)$. The usual Euclidean distance in general will not work well for two reasons: (1) the target values $\bar{F}_1, \ldots, \bar{F}_{N_s}$ may have quite different scales, in which case the Euclidean distance will tend to ignore target values with small scale; and (2) the Euclidean distance allows the optimization to "trade" one target value for another, while for our purposes it would be preferable that the selected values are uniformly close to the target values.

The first problem is solved by simply scaling the values $\tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma$ so that they have unit standard deviation (in the data or some relevant reference data set). The second problem is solved by using the L_∞ distance instead of the Euclidean distance:

$$D_{L_\infty}(\tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma; \bar{F}_1, \ldots, \bar{F}_{N_s}) \equiv \max(|\tilde{F}_1^\gamma - \bar{F}_1|, \ldots, |\tilde{F}_{N_s} - \bar{F}_{N_s}|)$$

A detailed example of the implementation of this procedure is described in section 11.3. In order to fix ideas, here we simply offer a preview of how the scatterplots mentioned above may look like and what kind of information they convey.

7.6.1 Example

Here we consider the special case of "deterministic forecasts", that is where the covariates are known to have a well defined analytical form as a function of time alone. This is useful when realistic covariates are missing, but a linear time trend is known not to be sufficient and a non-linear trend may be needed. We consider a specification of the form:

$$\mu_{at} = \beta_a^{(0)} + \beta_a^{(1)} t + \beta_a^{(2)} \log(t - \alpha)$$

where α is a given number[3], and assume that we are interested in forecasting mortality by lung cancer in males. The details of this example is provided in Section 11.3. For this choice, the target values of the summary measures 7.22 have been derived with the procedure described in the next section, and are as follows:

$$\bar{\text{SD}} \approx 0.33 \ , \quad \bar{F}_{\text{age}} \approx 0.56 \ , \quad \bar{F}_{\text{time}} \approx 0.029 \ , \quad \bar{F}_{\text{age/time}} \approx 0.006$$

We use a smoothness functional consisting of 3 terms:

$$\begin{aligned} H[\mu, \theta_{\text{age}}, \theta_{\text{time}}, \theta_{\text{age/time}}] \equiv & \frac{\theta_{\text{age}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{da^2} (\mu(a,t) - \bar{\mu}(a)) \right)^2 \\ & + \frac{\theta_{\text{time}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{dt^2} \mu(a,t) \right)^2 \\ & + \frac{\theta_{\text{age/time}}}{TA} \int_0^T dt \int_0^A da \left(\frac{\partial^3 \mu(a,t)}{\partial a \partial t^2} \right)^2 \end{aligned}$$

and therefore need to estimate 3 smoothness parameters, θ_{age}, θ_{time} and $\theta_{\text{age/time}}$. As explained at the beginning of this section, it is convenient to re-parametrize the smoothness functional using the standard deviations of the prior, σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$. We remind the reader that σ_{age} is simply the average standard deviation of the prior over age groups, if it were used in isolation, and it is linked to θ_{age} by

[3]In this example we set $\alpha = 1876$. The justification of this choice is presented in Section 11.3.

Equation 6.14 (Page 118), which we rewrite below (see Equations 6.7, Page 117 and 6.9, Page 117 for the definition of other quantities in this formula):

$$\theta_{\text{age}} = \frac{\text{Tr}\left(\mathbf{Z}D^{+}_{\text{age}}\mathbf{Z}'\right)}{AT\sigma^2_{\text{age}}}$$

The same formula applies, with the obvious modifications, for θ_{time} and $\theta_{\text{age/time}}$.

Once the smoothness functional and the specification have been chosen, the prior is defined, and all we have to do is to draw samples from it and compute empirically the expected value of the summary measures for many different values of the prior parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$. Scatterplots of the prior parameters against the empirical values of the summary measures are shown in Figure 7.1.

The main message emerging from the analysis of the scatterplots is that the target value of the summary measure $\bar{\text{SD}} = 0.33$ is not compatible with the value of the summary measure $\bar{F}_{\text{time}} = 0.029$. This is easily seen in the scatterplot of SD against F_{time} (see the fourth row and sixth column in Figure 7.1). Fixing the value of SD around 0.3, we see that only very small values of F_{time} can be realized (around 0.010). Therefore if we want a prior with a summary measure F_{time} closer to its target value of 0.029 we will need to settle for a higher value of the average standard deviation SD. The reason underlying this behavior of the prior is that we have very few covariates, each of which is a fixed function of time. Since the prior can only operate by influencing the coefficients on these variables, there just isn't much room to maneuver. (The same issue would occur if we had a prior that suggested that log-mortality move according to a quadratic but only a linear term for time was included among the covariates; no amount of adjusting the coefficients would produce the desired effect.) Therefore samples from the prior are very constrained to begin with, and if we want to achieve the level of variation over time corresponding to a value $\bar{F}_{\text{time}} = 0.029$ we need the prior to have a large standard deviation.

These considerations imply that we will need to settle for an approximate solution of Equation 7.21, that can be found using the procedure described above in this section. In Table 7.1, we display the combination of parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ that lead to the empirical value of the summary measures closest to the target value. The table shows the 25 closest values, sorted according to their distance to the target. Therefore the first row gives us the optimal values $\hat{\sigma}_k$, and the following rows represent combinations of σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ most likely to produce similar results.

7.6.2 Estimating the Expected Value of the Summary Measures

In order to use the procedure outlined above we need to begin with some substantively reasonable ranges for the expected values of the summary measures. While it would be possible to elicit some of these measures from subject matter experts, here we pretend that expert opinion is unavailable, and get our estimates using a procedure

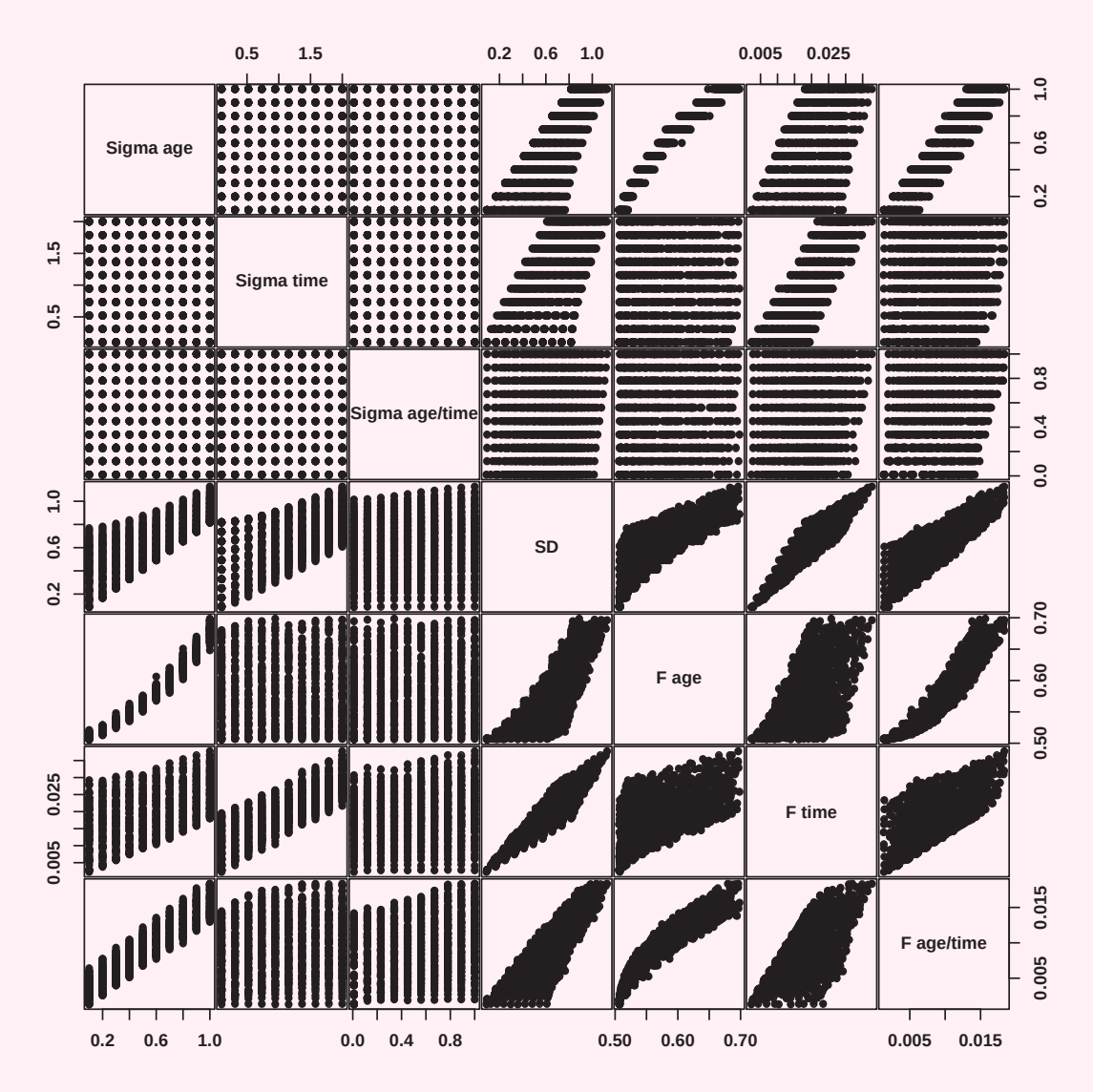

Figure 7.1: Scatterplots of summary measures by prior parameters: The plot shows the relationship between the prior parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ and summary measures SD, F_{age}, F_{time} and $F_{\text{age/time}}$.

σ_{age}	σ_{time}	$\sigma_{\text{age/time}}$	SD	F_{age}	F_{time}	$F_{\text{age/time}}$
0.3	1.578	0.12	0.543	0.528	0.022	0.005
0.2	1.367	0.78	0.559	0.523	0.022	0.006
0.2	1.367	1.00	0.575	0.525	0.022	0.007
0.4	1.156	0.89	0.575	0.549	0.022	0.009
0.2	1.578	0.45	0.576	0.523	0.023	0.006
0.4	1.156	1.00	0.579	0.553	0.021	0.010
0.2	1.367	0.89	0.568	0.524	0.021	0.007
0.3	1.367	0.67	0.580	0.537	0.022	0.008
0.4	1.156	0.67	0.562	0.545	0.021	0.009
0.3	1.578	0.34	0.584	0.533	0.022	0.006
0.4	1.156	0.78	0.570	0.549	0.021	0.009
0.3	1.367	0.78	0.591	0.538	0.022	0.008
0.4	1.367	0.45	0.592	0.551	0.021	0.009
0.4	1.578	0.12	0.585	0.537	0.021	0.006
0.2	1.578	0.56	0.596	0.522	0.023	0.006
0.5	1.156	0.45	0.594	0.563	0.021	0.009
0.3	1.367	0.89	0.599	0.538	0.021	0.008
0.3	1.789	0.12	0.600	0.531	0.023	0.005
0.3	1.367	0.56	0.567	0.533	0.020	0.007
0.2	1.789	0.23	0.589	0.518	0.020	0.004
0.4	1.578	0.23	0.602	0.537	0.021	0.007
0.2	1.789	0.12	0.571	0.518	0.022	0.003
0.3	1.367	0.45	0.550	0.533	0.020	0.007
0.5	1.367	0.23	0.604	0.562	0.021	0.008
0.3	1.367	1.00	0.606	0.534	0.022	0.008

Table 7.1: Summary Measures and Parameter Values: Combinations of different values of the parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ together with the corresponding value of the summary measures SD, F_{age}, F_{time} and $F_{\text{age/time}}$. The rows are sorted according to their distance to the target values for the summary measures.

which has the flavor of empirical Bayesian analysis. In practice, we recommend that this procedure be used in conjunction with expert involvement, perhaps as a starting point to orient the experts.

Instead of looking at the data to determine the parameters of the prior, we look at smoothed versions of the data, obtained by making forecasts of the time series and then selecting the in sample portion of the predictions. But to examine a set of forecasts, we need to start with some baseline model, although the particular model we choose should not matter much. Our recommendation is to use a simple version of our model, although one could also use LS when it happens to produce reasonable in sample fits in a particular application.

Once a smooth and realistic version of the data is available, we can then use these time series to compute an estimate of the expected value of the summary measures in Equation 7.22. For example, denoting by $\hat{\mu}_{cat}$ the in sample prediction of the model, the expected value of the summary measure F_{age} can be estimated as:

$$\bar{F}_{\text{age}} \approx \frac{1}{CAT} \sum_{c=1}^{C} \sum_{t=1}^{T} \sum_{a=1}^{A} |\hat{\mu}_{cat} - \hat{\mu}_{c,a-1,t}|$$

In addition to the mean, it may be useful to look at the entire distribution of the terms in the sum above, in order to get an idea of its spread and possible skewness.

In order to provide an example of such distributions we consider the case of death by lung cancer in males. We use a simple specification with a linear trend and a logarithmic trend in order to get a basic set of forecasts. The basic forecasts are initially obtained using our Bayesian method with a prior that smooths over age groups only (using a second derivative and constant weights) for all the countries with more than 15 observations, using a standard deviation of the prior equal to 0.3. Since not all the forecasts look reasonable, we eliminate those who do not, and re-run our method, trying a few alternative values for the standard deviation of the prior. The whole point of this procedure is to create a fairly large number of smoothed versions of the in sample data which look realistic. In these data, we find that a value of the standard deviation of the prior equal to 0.2 produces a reasonable in sample prediction, which we use as baseline starting point.

The distributions of the quantities involved in the computation of the summary measures 7.22 are shown in Figure 7.2, together with their mean values.

While the procedure outlined above is not rigorous, it is an example of the kind of qualitative analysis one can perform to set reasonable starting values for the standard deviation of the prior to help orient experts. The main point here is that it is possible to link the standard deviation of the prior to other quantities that are, at least in principle, observable, and about which we might have real prior knowledge. In our example this link is provided by Figure 7.1.

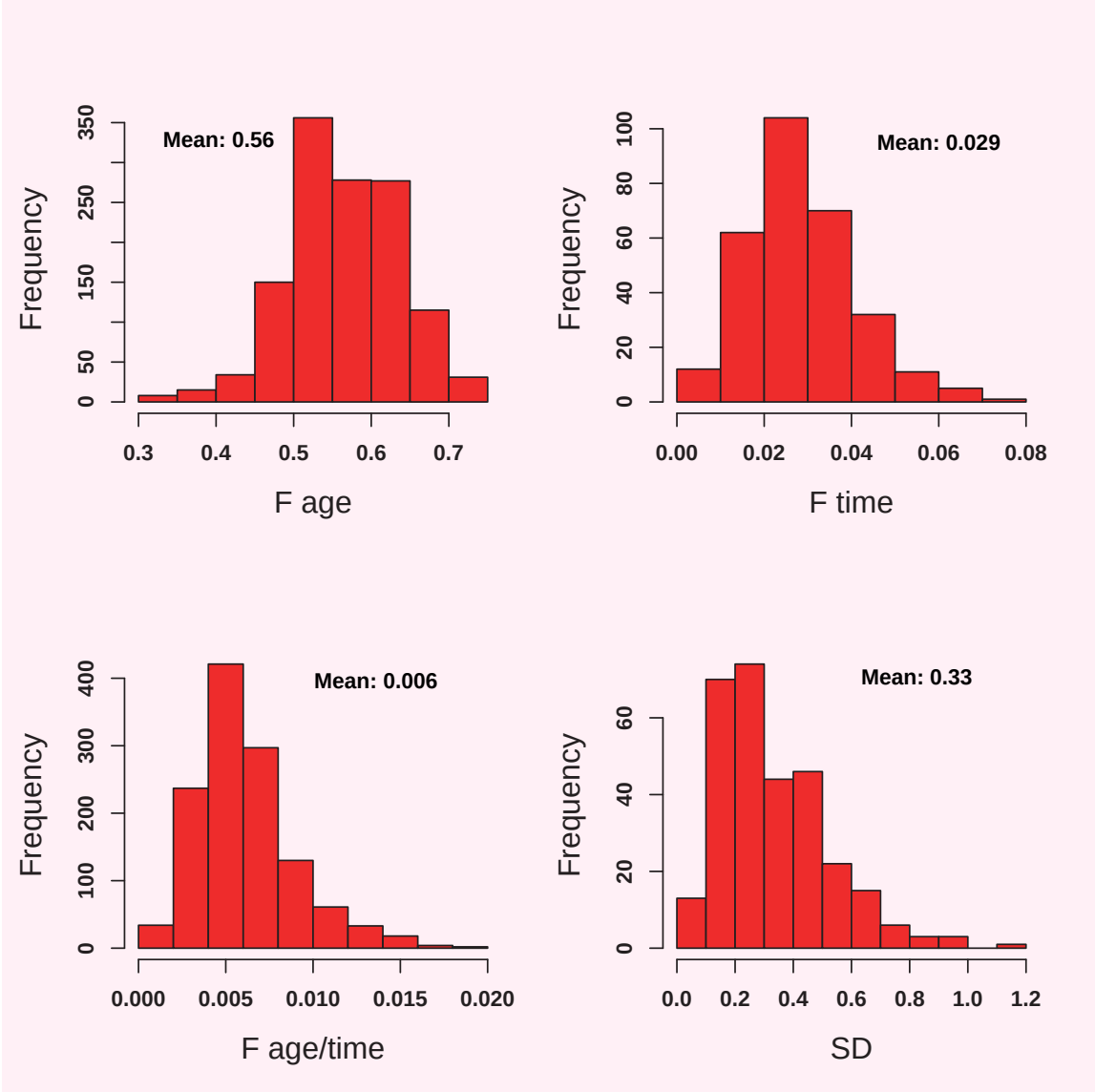

Figure 7.2: Result of the Empirical Bayes-like Procedure for Setting Summary Measure Target Values: Empirical distribution of the quantities involved in the computation of the summary measures in Equation 7.22. The dependent variable is log-mortality for lung cancer in males, for 25 countries.

7.7 Concluding Remark

This chapter offers a rich set of priors for analyzing mortality rates. But it also offers a set of tools researchers can use to adapt new priors in new substantive problems. The key features of our approach in this chapter involve the application of the two step method (introduced in Chapter 4) for specifying priors on the expected value of the dependent variable, rather than on the coefficients directly, new methods for the analysis of prior indifference via null spaces, and ways we developed to set priors with genuine prior knowledge.

Chapter 8

Comparisons and Extensions

In this Chapter, we provide some general procedures for understanding the priors built in Chapters 5–7. We begin in Section 8.1 with a systematic comparison of the priors on coefficients with our priors built with knowledge about the expected value of the dependent variable. We then prove, in Section 8.2, that priors specified in the large literature on hierarchical Bayesian models that feature exchangable clusters of cross-sections are special cases of our models. The results in this section demonstrate that all our results about the inappropriateness of putting priors on coefficients apply to the hierarchical literature as well. It also demonstrates how our approach has the same attractive features of empirical Bayes but without having to leave (as empirical Bayes does) the standard Bayesian approach to inference.

8.1 Priors on Coefficients vs. Dependent Variables

In this section, we compare the prior on coefficients from Section 4.2 with that on the expected value of the dependent variable, in Section 4.4 and Chapters 5 and 7. We provide intuition by comparing the joint densities in Section 8.1.1 and the conditional densities in Section 8.1.2. Section 8.1.3 describes connections between the results in our first two sections with theoretical results from the pattern recognition literature.

8.1.1 Joint Densities

In order to facilitate comparison, we write each prior in two equivalent forms with the aid of the quadratic form identity (Appendix B.2.6, Page 253). Then, we put the prior on $\boldsymbol{\beta}$ in Equation 4.5 (Page 69) side-by-side with the prior on μ from Equation 7.11 (Page 153) as follows:

$$H^{\mu}[\boldsymbol{\beta},\theta] = \theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \qquad \Leftrightarrow \qquad H^{\beta}[\boldsymbol{\beta},\Phi] = \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \Phi \boldsymbol{\beta}_j \tag{8.1}$$

$$= \frac{1}{2}\theta \sum_{ij} s_{ij} \|\mu_i - \mu_j\|^2 \qquad \Leftrightarrow \qquad = \frac{1}{2} \sum_{ij} s_{ij} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}^2 \tag{8.2}$$

where $\mu_i = \mathbf{Z}_i \boldsymbol{\beta}_i$ is a $T \times 1$ vector, $\|\cdot\|$ is the Euclidean norm and $\|\mathbf{b}\|_{\Phi}^2 = \mathbf{b}'\Phi\mathbf{b}$ is the Mahalanobis norm of the vector $\mathbf{b}$ (the left-hand side of Equation 8.2 can be proved to be equal to the left hand side of Equation 8.1 by direct substitution of μ_i and μ_j into the expression). Notice that the matrix s does not have to be the same in the two priors, but since it has similar meaning and its explicit form is irrelevant here we just take it to be the same to ease notation.

The difference between imposing smoothness on μ and $\boldsymbol{\beta}$ is now starting to become clear: when imposing smoothness on $\boldsymbol{\beta}$, researchers use s_{ij} as a distance in the space of cross-sectional units, but, for fixed i and j, no natural definition of distance between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ exists. Therefore, the usual procedure is to parametrize the distance between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ as the Mahalanobis distance $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}$. This approach obviously cannot be used when $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ have different dimensions, or correspond to different covariates, since the set of coefficients would have no obvious metric structure and so would not be comparable.

In contrast, in our case, rather than comparing the coefficients $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, we compare their predicted patterns for the expected value of the dependent variable, μ_i and μ_j, taking advantage of the fact that μ_i and μ_j are interpretable and there exists a natural distance between them, no matter what covariates are included. This distance is Euclidean (rather than Mahalanobis; see Appendix B.1.3, Page 233) and so the normalization matrix Φ is not required. In other words, we project $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ into *the same higher-dimensional metric space* through the covariate matrices $\mathbf{Z}_i$ and $\mathbf{Z}_j$, and then compare them. The covariates play here the role of "translators", allowing one to compare vectors of disparate quantities. This they do through the matrices $\mathbf{C}_{ij}$, that allow us to project a vector of "type i" onto a vector of "type j".

This result can be seen more clearly in Equation 8.1 where the prior on coefficients contains a sum of scalar products $\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j$, which does not have meaning unless $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are of the same type. However, in the right hand side of Equation 8.1 we see that following our approach the scalar products $\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j$ are well defined: vector $\boldsymbol{\beta}_j$ of type j is converted to a vector of type i by the matrix $\mathbf{C}_{ij}$, and then the usual Euclidean scalar product is computed (since $\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j = \boldsymbol{\beta}_j'\mathbf{C}_{ji}'\boldsymbol{\beta}_i$ we can also say that vector $\boldsymbol{\beta}_i$ of type i is converted to a vector of type j by the matrix $\mathbf{C}_{ji}'$). The matrices $\mathbf{C}_{ij}$, despite their simplicity, allow us to impose a metric structure on a set which does not have any existing structure: notice that while the set of coefficients $\boldsymbol{\beta}$ does not even have the structure of a vector space, since the sum of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ is not defined, a notion of distance is defined between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, when translated to the scale of the expected value, by the expression:

$$d^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j | \mathbf{Z}_i, \mathbf{Z}_j) \equiv \|\mathbf{Z}_i\boldsymbol{\beta}_i - \mathbf{Z}_j\boldsymbol{\beta}_j\|^2 = \boldsymbol{\beta}_i'\mathbf{C}_{ii}\boldsymbol{\beta}_i + \boldsymbol{\beta}_j'\mathbf{C}_{jj}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j \tag{8.3}$$

This expression should be compared with the Mahalanobis distance which, written in terms of scalar products, is as follows:

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_\Phi^2 = \boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_i + \boldsymbol{\beta}_j'\Phi\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j$$

This comparison reinforces the point that the expression $\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j$ is the "natural" scalar product between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, and, indeed, it has all the properties of a scalar product (except of course the fact that a true scalar product is always defined between elements of the same set). Similarly, the distance in Equation 8.3 satisfies all the axioms of a distance, or, to be precise, of a semi-distance, since $d^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j | \mathbf{Z}_i, \mathbf{Z}_j) = 0$ does *not* imply that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ (for a formal definition see Appendix B, Page 233).

One way to summarize this discussion is to say that the intuition of putting a prior on coefficients is reasonable except that the notion of similarity should be defined using prior knowledge about the expected value of the dependent variable.

8.1.2 Conditional Densities

Another useful way to compare the two approaches is to examine the implied conditional priors. Thus, the prior density on the $\boldsymbol{\beta}$'s implies the following conditional prior distribution of the coefficient $\boldsymbol{\beta}_i$ given the values of all the other coefficients $\boldsymbol{\beta}_{-i}$:

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}, \Phi \sim \mathcal{N}\left(\sum_j \frac{s_{ij}}{s_i^+}\boldsymbol{\beta}_j \ , \ \frac{\Phi^{-1}}{s_i^+} \right) \tag{8.4}$$

The expression above confirms the intuition that *smoothing is achieved by letting $\boldsymbol{\beta}_i$ be a weighted average of the regression coefficients of the neighboring cross-sections*, and obviously loses any meaning when the $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are not comparable. Performing a similar calculation for our prior in Equation 7.12 we obtain the following conditional prior:

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}, \theta \sim \mathcal{N}\left(\sum_j \frac{s_{ij}}{s_i^+}\mathbf{C}_{ii}^{-1}\mathbf{C}_{ij}\boldsymbol{\beta}_j, \frac{1}{\theta s_i^+}\mathbf{C}_{ii}^{-1} \right) \tag{8.5}$$

The key to this expression is the presence of two sets of matrices, with different roles: in the conditional mean, the matrix $\mathbf{C}_{ij}$ converts vectors of "type j" into vectors of "type i", but also produces a vector with different measurement units, and so the matrix $\mathbf{C}_{ii}^{-1}$ converts this result to the correct measurement units, to ensure that the coefficients $\boldsymbol{\beta}$ have measurement units which are the inverse of the measurement units of the covariates.

The presence of $\mathbf{C}_{ii}^{-1}$ in the conditional variance ensures that the Bayes estimator 4.3 (Page 68) based on prior in Equation 7.12 (Page 154) produces forecasts which are invariant for scaling of the covariates *in each cross-sectional unit*. In other words, we can decide to use pounds instead of dollars in some cross-sectional units and still obtain the same forecast (and obviously a different set of appropriately scaled coefficients). If we used the prior on coefficients in Equation 4.4 (Page 69) not only would we have to make sure that the covariates in the different cross-sections are measured in the same units, but if we changed units we would also have to change the scale of the covariance parameter Φ.

8.1.3 Connections to "Virtual Examples" in Pattern Recognition

Expression 8.5 has an interesting interpretation in terms of what in the pattern recognition literature are called "virtual examples". The connection to virtual examples is useful here to clarify the meaning of the prior and, in Chapter 10, as a starting point for a fast estimation procedure that does not require Markov Chain Monte Carlo algorithms.

In order to simplify the exposition, we do not smooth over time, so that $\mathbf{C}_{ij} = \frac{1}{T}\mathbf{Z}_i'\mathbf{Z}_j$, and let us interpret Equation 8.5 as saying that, conditional on the values of all the other coefficients $\boldsymbol{\beta}_{-i}$, we expect $\boldsymbol{\beta}_i$ to be in a neighborhood of the conditional mean, a fact that we write informally as

$$\boldsymbol{\beta}_i \approx \sum_j \frac{s_{ij}}{s_i^+}(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\mathbf{Z}_j\boldsymbol{\beta}_j$$

Then, noting that the quantities $\mathbf{Z}_j\boldsymbol{\beta}_j = \mu_j$ are the predicted values for the dependent variable in cross-section j, we write

$$\boldsymbol{\beta}_i \approx (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\sum_j \frac{s_{ij}}{s_i^+}\mu_j.$$

The sum in the expression above is simply the average of the predicted values for the dependent variable in the cross-sections that are neighbors of cross-section i (excluding i itself since $s_{ii} = 0$), and we call this quantity $\bar{\mu}_i$, rewriting:

$$\boldsymbol{\beta}_i \approx (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\bar{\mu}_i \tag{8.6}$$

This expression is a standard LS estimator, and has a simple interpretation. Given the values of all the other coefficients $\boldsymbol{\beta}_{-i}$, we could get an a priori likely estimate of $\boldsymbol{\beta}_i$ in two steps:

1. Obtain an estimate for the dependent variable in cross-section i by averaging the predicted values of the cross-sections that are neighbors of i (the vector $\bar{\mu}_i$)

2. Then, to obtain the coefficients in cross-section i, run a LS regression of $\bar{\mu}_i$ on $\mathbf{Z}_i$.

Since the vector $\bar{\mu}_i$ is not a vector of observed values, or "examples", but rather is inferred using prior knowledge, we say that it is a vector of "virtual examples", and in this sense we could say that *the role of the prior knowledge we have on the problem is to create suitable sets of virtual examples.* For more discussion of the connection between prior information and virtual examples see Abu-Mostafa (1992), Bishop (1995), and Niyogi, Girosi and Poggio (1998).

8.2 Extensions to Hierarchical Models and Empirical Bayes

In this Section, we demonstrate two fundamental results. First, our methods incorporate the key attractive features seen in empirical Bayes approaches, but without having to resort to the problematic empirical Bayes theory of inference. In empirical Bayes, hyperparameters from the last level of a hierarchical model are estimated rather than chosen a priori. Although this procedure might seem better because it brings the data to bear on the problem of making difficult choices about obscure hyperparameters, many scholars question the inferential validity of this approach: It uses the data twice and inferences must be corrected in various ad hoc ways to avoid underestimating the width of confidence intervals. Despite the inferential problems, however, this procedure is still frequently used, one important reason for which is because using the data in this way turns out to be equivalent to making the prior indifferent to certain chosen parameters (Carlin and Louis, 2000). For example, with empirical Bayes it is possible to achieve shrinkage among a set of parameters without having to specify the mean of the parameters. Of course, our formal approach to prior indifference accomplishes exactly the same task, but entirely within the standard Bayesian framework. As such, at least some of the reason for using empirical Bayesian approaches would seem to vanish. We demonstrate this equivalence here.

Second, we prove here that Bayesian hierarchical models, with clusters of exchangable units, are a special case of the Bayesian spatial models we are analyzing in this book. As such, our results about the inappropriateness of putting priors directly on coefficients in spatial models (see Section 4.3) also extends without modification to the Bayesian hierarchical modeling literature. Taken together, it would seem that the vast majority of all Bayesian models that use covariates are using prior densities that inappropriately reflect their prior knowledge. All our techniques for putting priors on the expected value of the dependent variable, and developing priors indifferent to chosen features of the parameters, apply to hierarchical models as well.

8.2.1 The Advantages of Empirical Bayes without Empirical Bayes

We begin by considering a hierarchical linear model, with N cross-sections and N vectors of coefficients $\boldsymbol{\beta}_i$. A common assumption is the following "shrinkage" prior:

$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\gamma,\ \tau^2\right)$$

The "direct" effect of this prior is to shrink the coefficients $\boldsymbol{\beta}_i$ toward the same mean γ. The "indirect" effect is that the coefficients are shrunk toward each other. It is often the case that the indirect effect is more desirable than the direct one: one can be confident that the coefficients $\boldsymbol{\beta}_i$ should be similar to each other without necessarily knowing what value they should assume. In other words, the researcher may be agnostic (indifferent) about the absolute level of the coefficients but may be knowledgeable about their relative size. Let us apply the idea of using subspaces to represent indifference. It is sufficient to work with one dimensional coefficients, so we assume $\boldsymbol{\beta}_i, \gamma \in \mathbb{R}$ in the following. Taking $\tau = 1$ for simplicity, the prior above can be rewritten as:

$$\mathcal{P}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\beta}_i - \gamma)^2\right) \tag{8.7}$$

Defining the $N \times 1$ vectors $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$ and $\boldsymbol{\gamma} \equiv (\gamma, \ldots \gamma)$, we rewrite the expression above in vector form:

$$\mathcal{P}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|^2\right). \tag{8.8}$$

While this prior is defined over $\mathbb{R}^N$, there is a whole subspace of $\mathbb{R}^N$ we are indifferent to: This is the set $V \subset \mathbb{R}^N \equiv \{x \mid x = (k, \ldots, k),\ \forall k \in \mathbb{R}\}$, which coincides with the diagonal of the positive orthant[1] in $\mathbb{R}^N$. In other words, we are indifferent between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i + k$, for any $k \in \mathbb{R}$. Denoting by $P_\perp$ the projector onto $V_\perp$, the orthogonal complement of V, the prior 8.8 can be made indifferent to V by simply projecting its argument onto $V_\perp$. Therefore we define a new prior:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|P_\perp(\boldsymbol{\beta} - \boldsymbol{\gamma})\|^2\right)$$

Since $\boldsymbol{\gamma} \in V$ by construction, then $P_\perp \boldsymbol{\gamma} = 0$ and the prior above becomes simply:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|P_\perp \boldsymbol{\beta}\|^2\right) \tag{8.9}$$

This expression makes clear that the only part of $\boldsymbol{\beta}$ we have prior knowledge about is $P_\perp \boldsymbol{\beta}$, which can be interpreted as the portion of $\boldsymbol{\beta}$ which contains only "relative"

[1] An "orthant" is a quadrant in three or more dimensions.

information. Let us find an explicit expression for $P_\perp$. By the properties of projection operators in Appendix B.1.13 (Page 240) we have $P_\perp = I - P_\circ$, where $P_\circ$ is the projector onto V, which we now recognize as the null space of the prior. $P_\circ$ is easily built in terms of an orthonormal basis for the subspace V, which is given by the constant row vector $\mathbf{v} = \frac{1}{\sqrt{N}}(1, \ldots, 1)$, where the factor $\sqrt{N}$ ensure normalization. Then the projector $P_\circ$ is given by $P_\circ \equiv \mathbf{v}'\mathbf{v}$ (see Page 240). The form of both $P_\perp$ and $P_\circ$ is given as

$$P_\circ \equiv \frac{1}{N}\begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix} \qquad P_\perp \equiv \begin{pmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \ldots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \ldots & -\frac{1}{N} \\ \vdots & \vdots & \ldots & \vdots \\ -\frac{1}{N} & -\frac{1}{N} & \ldots & 1-\frac{1}{N} \end{pmatrix}.$$

Therefore the projector $P_\perp$ operates on a vector $\boldsymbol{\beta}$ as

$$P_\perp\boldsymbol{\beta} = \boldsymbol{\beta} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i(1, \ldots, 1) \equiv \boldsymbol{\beta} - \bar{\boldsymbol{\beta}}(1, \ldots, 1)$$

where $\bar{\boldsymbol{\beta}} = \sum_{i=1}^{N}\boldsymbol{\beta}_i/N$ is the average of the elements of $\boldsymbol{\beta}$. Using this notation we can rewrite the prior 8.9 as follows:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\beta}_i - \widetilde{\boldsymbol{\beta}})^2\right) \tag{8.10}$$

This expression should be compared to the prior in Equation 8.7: the crucial difference between these two is that while the prior in Equation 8.7 shrinks $\boldsymbol{\beta}_i$ to a common, *predetermined* value γ, the prior in Equation 8.10 simply shrinks them to some common value, which is not known a priori, but is determined by the data. The prior in Equation 8.10 is similar to the empirical Bayes prior, with the difference that in empirical Bayes the value $\widetilde{\boldsymbol{\beta}}$ is replaced by an average of *empirical estimates* of the $\boldsymbol{\beta}_i$. It shares with the empirical Bayes prior the property of being independent of the absolute scale of $\boldsymbol{\beta}$, but it obviously does not require the empirical Bayesian theory of inference.

8.2.2 Hierarchical Models as Special Cases of Spatial Models

It is instructive to rewrite the prior in Equation 8.10 in a way that makes it more similar to the conditionally autoregressive priors described earlier in this chapter. First we notice that since $P_\perp$ is symmetric and is a projection operator, then $\|P_\perp\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}'P_\perp'P_\perp\boldsymbol{\beta} = \boldsymbol{\beta}'P_\perp\boldsymbol{\beta}$. Since the rows of $P_\perp$ sum to 0, we can use the quadratic form identity of Appendix B.2.6 (Page 253) to rewrite the prior in Equation 8.10 as:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2N}\sum_{i,j=1}^{N}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)^2\right). \tag{8.11}$$

This prior has the same form of the priors described by the left column of Equation 8.2, in which we have set $s_{ij} = 1$ for all $i, j = 1, \dots N$, and therefore it is the simplest form of conditionally autoregressive prior. *This proves that hierarchical models are special cases of spatial models in which all elements of a cluster are defined to be "neighbors" of all other elements.* All results in this book described in the context of spatial models thus also apply to hierarchical models.

8.3 Smoothing vs. Forecasting

In several instances researchers may be interested in smoothing mortality patterns, rather than forecasting them. More precisely, they might have noisy and possibly incomplete mortality data and are interested in removing the noise and possibly filling in the missing values.

This problem is easily handled in our framework and does not require the development of any new technique. Let us consider the case in which there is only one country. We have already defined a suitable set of priors for the expected value of the dependent variable:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}H[\mu, \theta]\right) \tag{8.12}$$

where the smoothness functional $H[\mu, \theta]$ will have, in general, a component for smoothing over age groups and a component for smoothing over time. Using smoothness functional over age groups and time as those in equation 5.8 and 7.1 respectively, the discretized version of the smoothness functional is[2]:

$$H[\mu, \theta] = \frac{\theta^{\text{age}}}{T} \sum_{aa't} W^{\text{age},\mathbb{n}}_{aa'} \mu_{at}\mu_{a't} + \frac{\theta^{\text{time}}}{A} \sum_{att'} W^{\text{time},\mathbb{k}}_{tt'} \mu_{at}\mu_{at'}$$

In the expression above $\mathbb{n}$ and $\mathbb{k}$ are the order of smoothness of the smoothness functional over age and time respectively.

Unlike in the regression case, the quantity we are interested in is μ itself, and we do not need to link μ to a set of covariates here. Therefore the smoothing problem consists simply in estimating μ given the prior 8.12 and the likelihood associated to the specification:

$$m_{at} \sim \mathcal{N}\left(\mu_{at}, \frac{\sigma_a^2}{b_{at}}\right) \qquad a = 1, \dots, A\ , \quad t = 1, \dots T$$

It is worth writing explicitely the negative log-posterior distribution for μ:

[2]We are considering a zero mean prior here. If a non-zero mean prior is needed, the rest of the analysis remains the same, but μ must be interpreted as the mean-centered age profile.

$$\log \mathcal{P}(\mu|m,\theta,\sigma) \propto \sum_{at} \frac{b_{at}}{\sigma_a^2} \left(m_{at} - \mu_{at}\right)^2 + \left[\frac{\theta^{\text{age}}}{T} \sum_{aa't} W_{aa'}^{\text{age},\mathbb{n}} \mu_{at}\mu_{a't} + \frac{\theta^{\text{time}}}{A} \sum_{att'} W_{tt'}^{\text{time},\mathbb{k}} \mu_{at}\mu_{at'} \right] \tag{8.13}$$

This expression fits squarely in the standard framework of non-parametric smoothing, and it can be seen as a simple application of standard Bayesian smoothing theory with[3]. Usually the estimate for μ is obtained my maximizing the posterior distribution, that is by minimizing the expression in equation 8.13 over μ, using a variety of methods, including cross-validation, to determine the parameters θ and σ. This approach takes advantage of the fact that the log-posterior is quadratic in μ, and therefore can use linear methods to solve part of the problem. Alternatively, one can develop a full Gibbs sampling strategy for the computation of the mean of the posterior.

In our case we do not need to develop new methods, or even write new code. We observe that any estimation strategy used to solve the regression problem can be immediately applied to solve the smoothing problem. In fact, we can always construct an artificial set of covariates such that the regression coefficients can be interpreted as estimates of the expected value of the dependent variable.

In order to see this, let us consider the regression problem with the usual specification $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$. Now let us choose as covariates a set of T dummy variables, with one dummy variables associated to each year from 1 to T[4]. This is equivalent to take a covariate matrix $\mathbf{Z}_a$ equal to the T-dimensional identity matrix. The specification $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$ can now be rewritten as

$$\mu_{at} = \beta_a^{(t)}$$

Therefore estimates of the coefficients are effortlessy translated into estimates for μ: A user who wishes to smooth the data and is not interested in forecasting has simply to decide the priors to use, create a set of dummy variables and run any estimation algorithm.

In the next section we present an application of this idea to a case which is not often considered in standard smoothing theory, which is smoothing in presence of discontinuities.

[3]We note that if there are missing values m_{at}, they are handled by simply filling them with a random number and settiung the weight b_{at} to zero.

[4]This implies that we drop the costant term for the specificiation

8.4 Priors when the Dependent Variable Changes Meaning

A common problem in the analysis of cause-specific mortality rates that the International Classification of Diseases and related health problem (ICD) used to label the cause of death changes at points in time (ICD are revised usually every 10 years). If a change is large enough, it could lead to visible discontinuities in the log-mortality time series, violating the assumptions that log-mortality is smooth over time. In Figure 8.1, we report the time series of log-mortality for "other infectious diseases" in males aged 0 to 4, for 4 different countries.

The jumps in years 1968 and 1979 do not correspond to the sudden beginning and end of some worldwide epidemic, but rather a change in the way some infectious diseases have been coded. In particular they appear to reflect the adoption of ICD-8 codes in 1968 and of ICD-9 in 1979.

Several ways exist for dealing with data with one or more such jumps (other than ignoring the problem). One consists of fixing the problem by pre-processing, that is, modifying the time series in order to make it comparable across the whole period of analysis. This can sometimes be done using "comparability ratios", which basically allow one to translate one meaning (or ICD code change) into another. However, comparability ratios are often unavailable (after all, if such a simple translation were possible, the international public health establishment would probably not have gone to such lengths to change the ICD code in the first place), and we are stuck with discontinuous time series. In addition, a discontinuity may exist for other reasons than ICD revision: For example a country which was previously unable to report deaths from certain isolated regions might be suddenly find the resources to increase coverage.

In principle, the dependent variable changes meaning after every jump, and since we are interested in forecasting only the last meaning, the obvious thing to do is to discard all the data before the last jump. This however is extreme, since it assumes no correlation between meanings. An alternative consists of making some assumptions about how log-mortality before and after the jump are related. Here we consider the simplest assumption: The dynamics of log-mortality remain unchanged, except for a shift and a change in slope at the time of the change (which we assume known). This can be incorporated into the model by including two new variables among the covariates: One is an indicator variable that is 1 before the change and 0 after, and the other is linear before the change and 0 (or constant) after (i.e., an indicator variable for the change and an interaction between a time trend and the indicator variable). Once these variables have been introduced, however, we also have to change our prior, since we no longer expect log-mortality to vary smoothly over time, and our smoothness assumption must be replaced by something weaker (less constraining).

Thus, we denote by t^* the year in which the discontinuity occurs (the extension to data with more than one jump will be obvious). The new prior knowledge can be

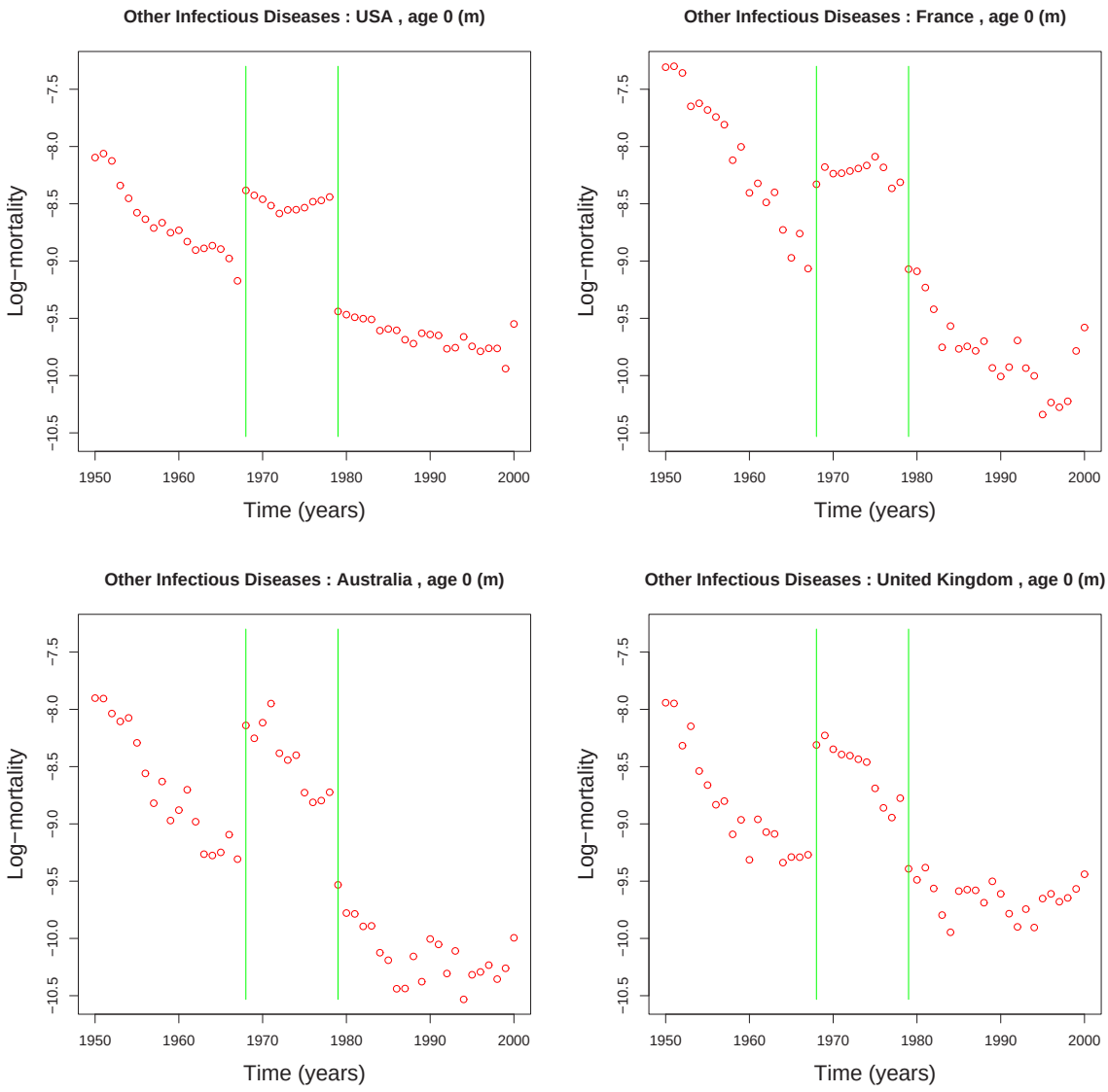

Figure 8.1: The Effects of Changes in ICD Codes: Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for four countries. The discontinuous behavior is likely due to changes in ICD codes. The green lines mark the years of the change.

formulated as follows: Log-mortality varies smoothly over time before t^* and after t^*, and we now write a smoothness functional which encodes this knowledge. It suffices to realize that the generic smoothness functional over time of Equation 7.1 (Page 146) can be rewritten as follows:

$$H[\mu, \theta] \equiv \frac{\theta}{N} \sum_i \left[\int_0^{t^*} dw^{\text{time}}(t) \left(\frac{d^{\mathbb{n}} \mu(i,t)}{dt^{\mathbb{n}}} \right)^2 + \int_{t^*}^{T} dw^{\text{time}}(t) \left(\frac{d^{\mathbb{n}} \mu(i,t)}{dt^{\mathbb{n}}} \right)^2 \right] \quad (8.14)$$

This smoothness functional has the desidered property, since it enforces smoothness independently before and after the jump, but it does not penalize functions which have a jump at time t^*. The null space for this functional is the set of *piecewise* polynomials of degree $\mathbb{n} - 1$, where the two "pieces" correspond to the period before and after t^*. Take for example the standard choice $\mathbb{n} = 2$: This implies that we are indifferent to patterns of mortality that are linear in time, but with different slopes and intercepts before and after the change. In other words, we make no assumptions about the coefficients of the two new variables.

Equation 8.14 underscores a key point that should always be kept in mind when choosing a prior: We must be clear about its domain of definition, that is the set of functions we can plug into it. When we write the prior of Equation 7.1 (Page 146) we implicitly assume that log-mortality is at least continuous, since if not the functional assumes an infinite value. However, the prior in Equation 8.14 is defined also for patterns of log-mortality that are discontinuous at t^*. In other words, the domain of definition of the functional in Equation 8.14 is larger then the one of the functional in Equation 7.1, although the two functionals coincide at the domain of Equation 7.1. Put differently, if we did not add the two variables there would be nothing gained by using functional 8.14 rather than functional 7.1 (in practice there would be something lost, due to the discretization of the derivative operator, which is always poorer at the extrema of the domain of the integral).

In order to understand the difference between using the smoothness functionals in Equations 8.14 and 7.1, we use both functionals, with $\mathbb{n} = 2$, to smooth (rather than forecast) the log-mortality patterns of Figure 8.1. Since the standard smoothness functional 7.1 "does not know" about the jumps, and it assumes that the underlying function is continous, we expect it to make large errors around the jumps, resulting in over-smoothing in those areas. The functional in Equation 8.14, which has been modified to include two jumps rather than one, one in year 1968 and one in year 1978, should smooth in the three regions independently. We report the results in Figure 8.2. The red dots are the data, the green dashed line is the results of smoothing with functional 7.1, which ignores the discontinuity, and the blue continuous line is the results of the functional 8.14. The smoothness parameter θ has been chosen large enough that the differences between the two smoothed curves are clear. These results are very pleasing, since the modified smoothness functional does exactly what it is supposed to do: It smooths the data while preserving the discontinuities.

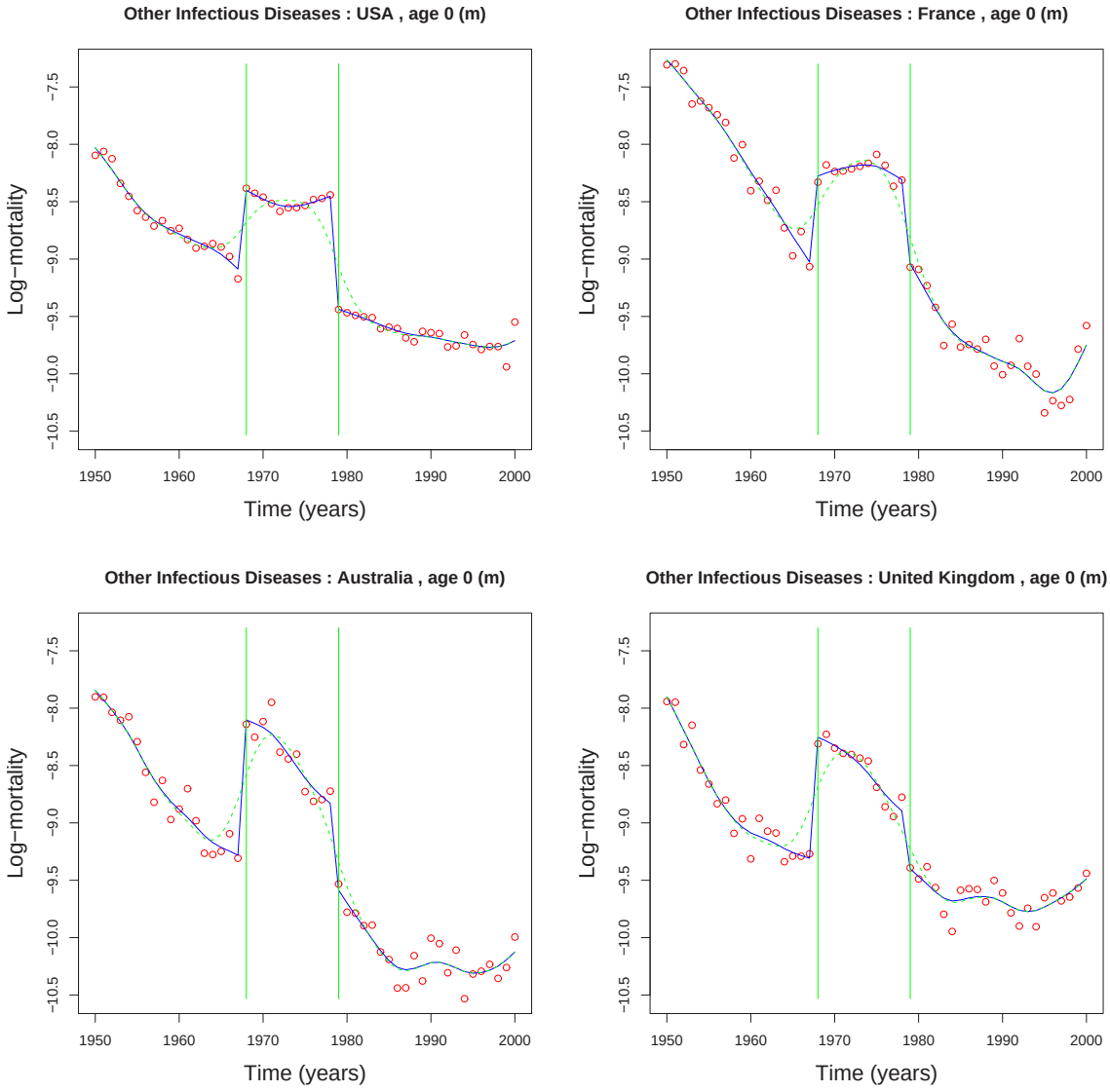

Figure 8.2: Modeling The Effects of Changes in ICD Codes: Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for four countries. The green curve smooths the data with the standard smoothness functional, while the blue curve smooths with the modified smoothness functional, allowing for discontinuities. The smoothness parameter θ has been set to 10, which is probably close to optimal.

In order to check that the modified smoothness functional also has the right null space, we smooth the data with the same smoothness functionals, but with a near-infinity value of the smoothness parameter. In so doing we force the smoothed curve to lie in the null space of the functional, providing the best approximation to the data from the null space. We report these results in Figure 8.3, using the same color coding as before. Note that the green curve is a straight line, because the null space of the smoothness functional 7.1 with $\mathtt{m} = 2$ is the set of polynomials of degree 1. For the modified smoothness functional in Equation 8.14, the smoothed curve is a *piecewise* polynomial of degree 1, as predicted by the theory.

8.5 Concluding Remark

Taken together with the other results in this book, the practical issues raised in this chapter are intended to eliminate the necessity of making uncomfortable or arbitary choices in applications. We have systematically related all parameters and hyperprior values directly to features of prior knowledge, in the form with which experts in the field seem most familiar.

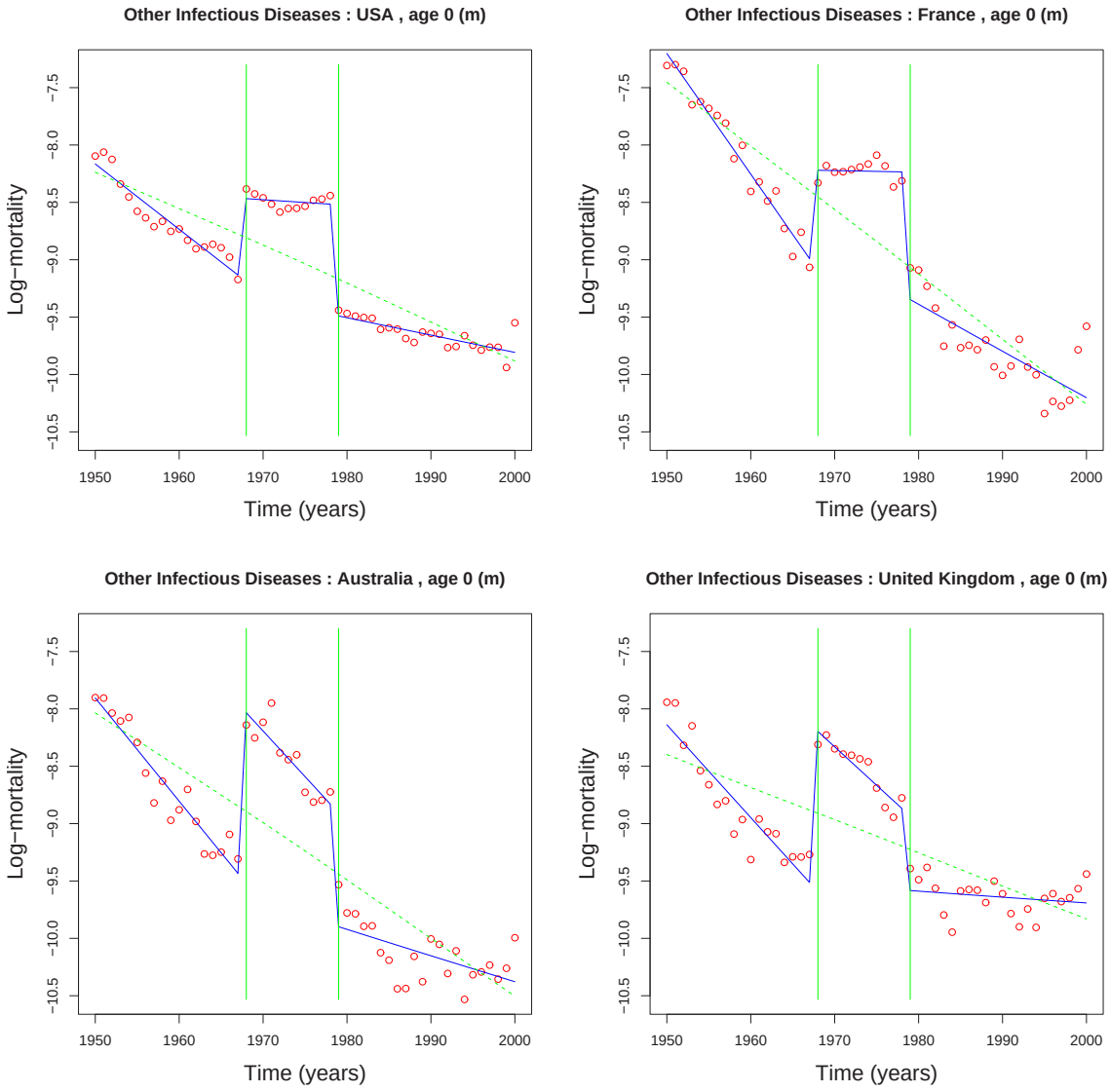

Figure 8.3: The Null Space for Models of Changes in ICD Codes: Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for 4 countries. The green curve smooths the data with the standard smoothness functional, while the blue curve smooths with the modified smoothness functional, allowing for discontinuities. The smoothness parameter θ has been set to 100,000, forcing the smoothed curve into the null space of the smoothness functional.

Part III
Estimation

In this Part, we show how to estimate and implement the models introduced in Part II. Chapter 9 implements the full Bayesian version of our model via Markov Chain Monte Carlo algorithms. Chapter 10 shows how to implement a faster regularization theory-based estimation procedure without Markov Chains.

Chapter 9

Markov Chain Monte Carlo Estimation

In practical applications, researchers can build a prior using any combination of the results developed in Part II. Once this step has been performed we are left with a prior of the mathematical form given in Equation 7.12 (Page 154), except for the fact that the exponent is likely to contain a sum of l terms of that form, with different parameters $\theta_1, \ldots, \theta_l$ and different matrices $\mathbf{C}_{ij}$. Then the only thing left to do is to assume some reasonable prior densities for σ and θ, plug them in Equation 4.3 and estimate the mean of the posterior distribution of $\boldsymbol{\beta}$. In this chapter, we summarize the complete model, filling in these remaining details, and then describe a method of estimation based on the Gibbs sampler to calculate quantities of interest (Tanner, 1996). We report our calculations for the case in which there is only 1 prior (for example the one for smoothness over age groups). The full details of the more general case, which follows the same lines but has longer and more cumbersome notation, appears in the manual accompanying our software (see Appendix F).

9.1 Complete Model Summary

We now review the model, and identify the main quantities involved in the estimation. We begin with the full posterior density (reproduced from Equation 4.2, Page 68):

$$\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m) \propto \mathcal{P}(m \mid \boldsymbol{\beta}, \sigma)\mathcal{P}(\boldsymbol{\beta} \mid \theta)\mathcal{P}(\theta)\mathcal{P}(\sigma) \tag{9.1}$$

In this section, we now formally define each of densities on the right side of Equation 9.1, so that we can then compute the mean posterior of the coefficients (from Equation

4.3):

$$\boldsymbol{\beta}^{\text{Bayes}} \equiv \int \boldsymbol{\beta}\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m) d\boldsymbol{\beta} d\theta d\sigma. \tag{9.2}$$

and our forecasts.

9.1.1 Likelihood

Each cross-section i includes T_i observations and has its own standard deviation σ_i. As explained in Section 3.1.2, we allow the observed values of the dependent variable m_{it} to be weighted. Therefore instead of m and $\mathbf{Z}$, and the likelihood $\mathcal{P}(m \mid \boldsymbol{\beta}, \sigma)$, we use their weighted counterparts y and $\mathbf{X}$ (see Equations 3.8 and 3.10 at Page 54). The weighted version of the likelihood is then

$$\mathcal{P}(y \mid \boldsymbol{\beta}, \sigma) \propto \left(\prod_i (\sigma_i^{-2})^{\frac{T}{2}} \right) \times \exp\left(-\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} \sum_t (y_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_i)^2 \right). \tag{9.3}$$

9.1.2 Prior for $\boldsymbol{\beta}$

We consider a prior for $\boldsymbol{\beta}$ of the form described in Equation 7.12 (Page 154). This prior has only one hyperparameter, θ, and can be expressed as

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) = K\theta^{\frac{r}{2}} \exp\left(-\frac{1}{2}\theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \right), \tag{9.4}$$

where r is the rank of the matrix defining the quadratic form in the exponent in 9.4 (see Equation C.8, page 262).

9.1.3 Prior for σ_i

The functional form of the prior for σ_i is chosen according to convenience. Therefore we follow standard practice, and choose an inverse Gamma prior for σ_i^2:

$$\sigma_i^{-2} \sim \mathcal{G}(\mathbb{d}/2, \mathbb{e}/2) \tag{9.5}$$

Here $\mathbb{d}$ and $\mathbb{e}$ are user specified parameters which determine the mean and the variance of σ_i^{-2} as follows:

$$\mathbb{E}[\sigma_i^{-2}] = \frac{\mathbb{d}}{\mathbb{e}}, \quad \mathbb{V}[\sigma_i^{-2}] = \frac{\mathbb{d}}{\mathbb{e}^2}. \tag{9.6}$$

In order to specify these parameters the user must specify the mean and variance of σ_i^{-2} and then solve equation 9.6 for $\mathbb{d}$ and $\mathbb{e}$. The user may not have prior knowledge

on σ_i^{-2}, and she is more likely to have some knowledge on σ_i (see section 6.5.3). Therefore one should relate the parameters $\mathbb{d}$ and $\mathbb{e}$ to moments of σ_i, rather than σ_i^{-2}. This is not totally straightforward, since the resulting formulas do not allow for a closed form solutions, and a numerical solution will need to be computed. We derive here all the necessary formulas.

Since the object about which we have prior knowledge is σ_i it is important to understand how the prior density of equation 9.5 looks like in terms of σ_i. Defining the auxiliary variable $s_i \equiv \sigma_i^{-2}$ From the definition of Gamma density, we rewrite the density of equation 9.5 as:

$$\mathcal{P}(s_i; \mathbb{d}, \mathbb{e}) = \frac{1}{\Gamma(\frac{\mathbb{d}}{2})} \left(\frac{\mathbb{e}}{2}\right)^{\frac{\mathbb{d}}{2}} {s_i}^{\frac{\mathbb{d}}{2}-1} e^{-\frac{\mathbb{e}}{2}s_i}$$

The density for σ_i is derived with a simple change of variable and it has the following form:

$$\mathcal{P}(\sigma_i; \mathbb{d}, \mathbb{e}) = \frac{2}{\Gamma(\frac{\mathbb{d}}{2})} \left(\frac{\mathbb{e}}{2}\right)^{\frac{\mathbb{d}}{2}} \sigma_i^{-\mathbb{d}-1} e^{-\frac{\mathbb{e}}{2}\frac{1}{\sigma_i^2}}$$

A lengthy but straightforward calculation shows that the moments of the density above are as follows:

$$\mathbb{E}[\sigma_i^n] = \frac{\Gamma(\frac{\mathbb{d}-n}{2})}{\Gamma(\frac{\mathbb{d}}{2})} \left(\frac{\mathbb{e}}{2}\right)^{\frac{n}{2}} \tag{9.7}$$

The mean and variance of the prior density for σ_i are therefore:

$$\begin{aligned} \mathbb{E}[\sigma_i] =& \frac{\Gamma(\frac{\mathbb{d}-1}{2})}{\Gamma(\frac{\mathbb{d}}{2})} \left(\frac{\mathbb{e}}{2}\right)^{\frac{1}{2}} \\ \mathbb{V}[\sigma_i] =& \frac{\mathbb{e}}{\mathbb{d}-2} - \mathbb{E}[\sigma_i]^2 \end{aligned} \tag{9.8}$$

Notice that the variance of σ_i goes to infinity as $\mathbb{2}$ tends to 2, and therefore we impose the constraint $\mathbb{d} > 2$. The relationship 9.8 cannot be inverted to express $\mathbb{d}$ and $\mathbb{e}$ as a function of $\mathbb{E}[\sigma_i]$ and $\mathbb{V}[\sigma_i]$, and a numerical procedure has to be used. To this end it is convenient to rewrite the equations in 9.8 as:

$$\begin{aligned} \frac{\mathbb{E}[\sigma_i]}{\sqrt{\mathbb{V}[\sigma_i] + \mathbb{E}[\sigma_i]^2}} =& \frac{\Gamma(\frac{\mathbb{d}-1}{2})}{\Gamma(\frac{\mathbb{d}}{2})} \sqrt{\frac{\mathbb{d}}{2} - 1} \\ \mathbb{e} =& (\mathbb{d}-2)\left(\mathbb{V}[\sigma_i] + \mathbb{E}[\sigma_i]^2\right) \end{aligned} \tag{9.9}$$

The top part of equation 9.9 is easily solved numerically (for $d > 2$), since it has only one solution, and once a value for $\mathbb{d}$ is obtained the value for $\mathbb{e}$ follows from simple substitution.

9.1.4 Prior for θ

Here again we follow standard practice and choose a Gamma prior for θ:

$$\theta \sim \mathcal{G}(\mathbb{f}/2, \mathbb{g}/2) \tag{9.10}$$

where $\mathbb{f}$ and $\mathbb{g}$ are user specified parameters, which control the mean and the variance of θ through formulas like those in Equation 9.6. We have shown in Section 6.2 (Page 113) how to make reasonable choices for the mean and the variance of θ, and therefore we can use those results to set $\mathbb{f}$ and $\mathbb{g}$.

9.1.5 The Posterior Density

Now we can bring together all the quantities defined above and write the full posterior density as

$$\begin{aligned}\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid y) \propto & \left(\prod_i (\sigma_i^{-2})^{\frac{\mathbb{d}+T_i}{2}-1} e^{-\frac{1}{2}\mathbb{e}\sigma_i^{-2}} \right) \left(\theta^{\frac{\mathbb{f}+r}{2}-1} e^{-\frac{\mathbb{g}}{2}\theta} \right) \times \\ & \times \exp\left(-\frac{1}{2} \left[\sum_i \frac{1}{\sigma_i^2} \sum_t (y_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_i)^2 + \theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \right] \right).\end{aligned} \tag{9.11}$$

In the next section we briefly describe the Gibbs algorithm used to sample from this density and to compute the posterior mean in Equation 9.1.

9.2 The Gibbs Sampling Algorithm

We evaluate of the conditional mean in Equation 9.2 using a Monte Carlo Markov Chain (MCMC) approach. In this Section, we give the expressions needed to implement the Gibbs sampler (Geman and Geman, 1984; Gelfand, A.E. and Smith, A.F.M., 1990), one of the most commonly used MCMC techniques. We refer the reader to standard textbooks on MCMC for a description of the Gibbs sampler (Gelman et al., 1995; Gilks, Richardson and Spiegelhalter, 1996; Tanner, 1996). We describe this algorithm with reference to the prior in Equation 7.12, with only one hyper-parameter θ and only one type of cross-sectional index, generically denoted by i.

To draw random samples from the posterior density in Equation 9.11, we use the Gibbs sampling algorithm. The essence of the Gibbs sampler lies in breaking a complicated joint probability density into a set of full conditional densities, and sampling one variable (or a group of variables) at a time, conditional on the values of the others.

In our case we have three sets of variables, $\boldsymbol{\beta}$, σ and θ, so that one iteration of the algorithm consists of sampling each of these sets. To simplify our notation, we denote the density of a variable x conditional on all the others by $\mathcal{P}(x \mid \text{else})$. Then, we write an iteration of the Gibbs sampler containing the following steps:

1. $\sigma_i^{-2} \sim \mathcal{P}(\sigma_i^{-2} \mid \text{else}) \quad i = 1, \ldots, n$
2. $\theta \sim \mathcal{P}(\theta \mid \text{else})$
3. $\boldsymbol{\beta}_i \sim \mathcal{P}(\boldsymbol{\beta}_i \mid \text{else}) \quad i = 1, \ldots, n$

Once we know how to sample from the conditional densities above, we compute the posterior mean of $\boldsymbol{\beta}$ in Equation 9.1 by averaging the values of $\boldsymbol{\beta}$ obtained by repeating these steps a large number of times (after having discarded a suitable number of "burn-in" iterations to ensure that the algorithm has converged, possibly also with separate chains; we do not worry about autocorrelation in the series unless we are computing standard errors). Since the conditional densities need to be known only up to a normalization factor, we only need terms in the posterior that include the variables of interest.

We now derive each of these conditional densities. We also show that each can be simply understood as a weighted average of a maximum likelihood estimate and the prior mean.

9.2.1 Sampling σ

The Conditional Density

When we choose the prior for σ_k we implicitly assume that the relevant variable (the one with the gamma density) was σ_k^{-2}, rather than σ_k. Consistently with that choice we use σ_k^{-2} as the sampled variable, and pick from Equation 9.11 all the terms that contain σ_k^{-2}, grouping the others in a generic normalization constant. We thus obtain:

$$\mathcal{P}(\sigma_k^{-2} \mid \text{else}) \propto (\sigma_k^{-2})^{\frac{\mathbb{d}+T_k}{2}-1} e^{-\frac{1}{2}\mathbb{e}\sigma_k^{-2}} \exp\left(-\frac{1}{2}\sigma_k^{-2}\sum_t (y_{kt} - \mathbf{X}'_{kt}\boldsymbol{\beta}_k)^2\right)$$

which is a Gamma distribution for σ_k^{-2}. Thus, by defining

$$\text{SSE}_k \equiv \sum_t (y_{kt} - \mathbf{X}'_{kt}\boldsymbol{\beta}_k)^2 , \tag{9.12}$$

we conclude that sampling for σ_k^{-2} should be as follows:

$$\sigma_k^{-2} \mid \text{else} \sim \mathcal{G}\left(\frac{\mathbb{d}+T_k}{2}, \frac{\mathbb{e}+\text{SSE}_k}{2}\right). \tag{9.13}$$

Interpretation

In order to clarify this expression further, we write the conditional expected value of σ_k^{-2}:

$$\mathbb{E}[\sigma_k^{-2} \mid \text{else}] = \frac{\mathbb{d} + T_k}{\mathbb{e} + \text{SSE}_k}$$

and define, respectively, $\sigma^2_{\text{P},k}$, which is related to the expected value of σ_k^2 under the prior in Equation 9.5, and $\sigma^2_{k,\text{ML}}$, the usual maximum likelihood estimator of σ_k^2:

$$\sigma^2_{\text{P},k} \equiv \frac{\mathbb{e}}{\mathbb{d}} = \frac{1}{\mathbb{E}[\sigma_k^{-2}]} \ , \quad \sigma^2_{k,\text{ML}} = \frac{\text{SSE}_k}{T_k}.$$

Now rewrite the equation for the conditional mean of σ_k^{-2}:

$$\mathbb{E}[\sigma_k^{-2} \mid \text{else}] = \left(\frac{\mathbb{d}\sigma^2_{\text{P},k} + T_k \sigma^2_{k,\text{ML}}}{T_k + \mathbb{d}} \right)^{-1}.$$

This expression helps clarify that when $\mathbb{d}$ is large — when the prior density is highly concentrated around its mean — the conditional mean of σ_k^{-2} is very close to the prior mean. On the other side, when the number of observations T_k is large then the likelihood dominates, and the conditional mean becomes determined by the likelihood. Although this conclusion is not surprising, it is useful, since it makes clear the "dual" role of the quantities $\mathbb{d}$ and T_k, which control the trade-off between the prior and the likelihood, and that can be thought as measures of concentration around the mean. It is also possible to show a similar kind of duality between $\mathbb{e}$ and SSE_k.

9.2.2 Sampling θ

The Conditional Density

Proceeding in the same way as above we write the conditional distribution for θ as

$$\mathcal{P}(\theta \mid \text{else}) \propto \theta^{\frac{\mathbb{f}+r}{2}-1} e^{-\frac{1}{2}\theta \mathbb{g}} \exp\left(-\frac{1}{2}\theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \right),$$

which is again a Gamma distribution. Thus, sampling for θ may be done according to the following, expressed in standard form:

$$\theta \sim \mathcal{G}\left(\frac{\mathbb{f}+r}{2}, \frac{\mathbb{g}}{2} + \frac{1}{2}\sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \right). \tag{9.14}$$

Interpretation

Again, we examine the conditional mean of θ, which is:

$$\mathbb{E}[\theta \mid \text{else}] = \frac{\mathbb{f} + r}{\mathbb{g} + \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j}.$$

In order to interpret this expression we define, respectively, θ_{P}, the expected value of θ under its prior (see Equation 9.10), and θ_{ML}, the maximum likelihood estimator of θ, that is the value of θ which maximizes $\mathcal{P}(\boldsymbol{\beta} \mid \theta)$ (see Equation 9.4):

$$\theta_{\mathrm{P}} \equiv \mathbb{E}[\theta] = \frac{\mathbb{f}}{\mathbb{g}}\ , \quad \theta_{\mathrm{ML}} \equiv \frac{r}{\sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j}$$

Rewriting the equation above for the conditional mean as an equation for its reciprocal is easier. Although the mean of the reciprocal is not the reciprocal of the mean these quantities are related, which is enough for the purpose of explanation:

$$\frac{1}{\mathbb{E}[\theta \mid \text{else}]} = \frac{\mathbb{f}\frac{1}{\theta_{\mathrm{P}}} + r\frac{1}{\theta_{\mathrm{ML}}}}{\mathbb{f} + r}.$$

As is the case for σ, this expression depends on the trade-off between two terms: One which relates to the prior of θ and one which relates to its likelihood. The parameters which control this tradeoff are $\mathbb{f}$ and r. The parameter $\mathbb{f}$ controls how concentrated is the prior distribution of θ around its mean. By the same token we would expect that r, the rank of the matrix in the exponent of $\mathcal{P}(\boldsymbol{\beta} \mid \theta)$, describes the concentration of $\mathcal{P}(\boldsymbol{\beta} \mid \theta)$ around its mean. That is the case can be seen by writing the density for the random variable $H = \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j$, which is what is relevant from the point of view of θ. Using the techniques described in Appendix C we can show that:

$$H \sim \mathcal{G}(r, \theta)$$

From here we can see immediately that r plays for the likelihood the same role played by $\mathbb{f}$ in the prior, and is indeed a measure of concentration.

9.2.3 Sampling $\boldsymbol{\beta}$

The Conditional Density

In order to find the distribution of $\boldsymbol{\beta}_k$ with all the other variables held constant, we need to isolate from the posterior all terms that depend on $\boldsymbol{\beta}_k$. As a first pass, we eliminate unnecessary multiplicative terms in Equation 9.4 and write:

$$\mathcal{P}(\boldsymbol{\beta}_k \mid \text{else}) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_k^2}\sum_t (y_{kt} - \mathbf{X}_{kt}'\boldsymbol{\beta}_k)^2 + \theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j\right]\right). \qquad (9.15)$$

We collect in a generic term K all the terms which do not depend on $\boldsymbol{\beta}_k$ (we reuse the symbol K to refer to possibly different constants for each equation below). For the first term in Equation 9.15 we have:

$$\sum_t (y_{kt} - \mathbf{X}'_{kt}\boldsymbol{\beta}_k)^2 = \boldsymbol{\beta}'_k\mathbf{X}'_k\mathbf{X}_k\boldsymbol{\beta}_k - 2\boldsymbol{\beta}'_k\mathbf{X}'_k y_k + K$$

For the second term in Equation 9.15 we have:

$$\sum_{ij} W_{ij}\boldsymbol{\beta}'_i\mathbf{C}_{ij}\boldsymbol{\beta}_j = W_{kk}\boldsymbol{\beta}'_k\mathbf{C}_{kk}\boldsymbol{\beta}_k + 2\sum_{j\neq k} W_{jk}\boldsymbol{\beta}'_k\mathbf{C}_{kj}\boldsymbol{\beta}_j + K$$

Using the quadratic form identity in Appendix B.2.6 (Page 253), $W = s^+ - s$, and so we rewrite the expression above as:

$$\sum_{ij} W_{ij}\boldsymbol{\beta}'_i\mathbf{C}_{ij}\boldsymbol{\beta}_j = s^+_k\boldsymbol{\beta}'_k\mathbf{C}_{kk}\boldsymbol{\beta}_k - 2\sum_j s_{jk}\boldsymbol{\beta}'_k\mathbf{C}_{kj}\boldsymbol{\beta}_j + K.$$

Putting everything together,

$$\mathcal{P}(\boldsymbol{\beta}_k \mid \text{else}) \propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}'_k\left(\frac{\mathbf{X}'_k\mathbf{X}_k}{\sigma_k^2} + \theta s^+_k\mathbf{C}_{kk}\right)\boldsymbol{\beta}_k - 2\boldsymbol{\beta}'_k\left(\frac{\mathbf{X}'_k y_k}{\sigma_k^2} + \theta\sum_j s_{jk}\mathbf{C}_{kj}\boldsymbol{\beta}_j\right)\right]\right),$$

we now define

$$\Lambda_k^{-1} \equiv \frac{\mathbf{X}'_k\mathbf{X}_k}{\sigma_k^2} + \theta s^+_k\mathbf{C}_{kk}$$

and

$$\overline{\boldsymbol{\beta}_k} \equiv \frac{\mathbf{X}'_k y_k}{\sigma_k^2} + \theta\sum_j s_{jk}\mathbf{C}_{kj}\boldsymbol{\beta}_j.$$

With these definitions we have:

$$\begin{aligned}\mathcal{P}(\boldsymbol{\beta}_k \mid \text{else}) &\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}'_k\Lambda_k^{-1}\boldsymbol{\beta}_k - 2\boldsymbol{\beta}'_k\overline{\boldsymbol{\beta}_k}\right]\right)\\ &= \exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta}_k - \Lambda_k\overline{\boldsymbol{\beta}_k})'\Lambda_k^{-1}(\boldsymbol{\beta}_k - \Lambda_k\overline{\boldsymbol{\beta}_k})\right]\right)\end{aligned}$$

Therefore, we need to sample $\boldsymbol{\beta}_k$ as follows:

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\Lambda_k\overline{\boldsymbol{\beta}_k}, \Lambda_k) \tag{9.16}$$

which is easily done by setting:

$$\boldsymbol{\beta}_k = \Lambda_k\overline{\boldsymbol{\beta}_k} + \sqrt{\Lambda_k}\mathbf{b} \quad , \quad \mathbf{b} \sim \mathcal{N}(0, I).$$

Interpretation

Despite the apparent complexity, Equation 9.16 has a clear interpretation, similar to the interpretation of the formulas for the conditional means of σ_k^{-2} and θ. In those cases the conditional mean was a weighted average of two terms: one was interpreted as a maximum likelihood estimate, and the other was the prior mean. Since we expect to see the same phenomenon, we define the two quantities:

$$\boldsymbol{\beta}_k^{\mathrm{ML}} \equiv (\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k' y_k \ , \quad \boldsymbol{\beta}_k^{\mathrm{P}} \equiv \sum_j \frac{s_{kj}}{s_k^+}\mathbf{C}_{kk}^{-1}\mathbf{C}_{kj}\boldsymbol{\beta}_j$$

The quantity $\boldsymbol{\beta}_k^{\mathrm{ML}}$ is simply the maximum likelihood estimator of $\boldsymbol{\beta}_k$. The quantity $\boldsymbol{\beta}_k^{\mathrm{p}}$ is the conditional mean of the prior, in Equation 8.5 (Page 171).

In order to see the meaning of Equation 9.16, we consider a particular, but informative, case. Remember that by definition we have $\mathbf{C}_{kk} = T^{-1}\mathbf{Z}_k'\mathbf{Z}_k$. Here $\mathbf{Z}_k$ is a vector of covariates extending over T time periods — the time over which we think prior knowledge is relevant. In general, the data matrix $\mathbf{Z}_k$ differs from the data matrix $\mathbf{X}_k$: $\mathbf{X}_k$ might include population weights, and it reflects the same pattern of missing values of m_{it}. Therefore, even without population weighting, the rows of $\mathbf{X}_{it}$ are a subset of the rows of $\mathbf{Z}_k$. Here, for the purpose of explanation, we assume that $\mathbf{X}_k$ and $\mathbf{Z}_k$ are identical, so that $\mathbf{C}_{kk} = T_k^{-1}\mathbf{X}_k'\mathbf{X}_k$. Using this assumption and a bit of algebra, we rewrite the conditional mean for $\boldsymbol{\beta}_k$ as follows:

$$\mathbb{E}[\boldsymbol{\beta}_k \mid \text{else}] = \frac{\frac{T_k}{\sigma_k^2}\boldsymbol{\beta}_k^{\mathrm{ML}} + \theta s_k^+ \boldsymbol{\beta}_k^{\mathrm{P}}}{\frac{T_k}{\sigma_k^2} + \theta s_k^+}.$$

As expected, the conditional mean of $\boldsymbol{\beta}_k$ is a weighted average of $\boldsymbol{\beta}_k^{\mathrm{ML}}$ and $\boldsymbol{\beta}_k^{\mathrm{P}}$. The weight of $\boldsymbol{\beta}_k^{\mathrm{ML}}$ is large when the number of observations T_k is large, or when the noise affecting the observation (σ_k), which also measures the variance of $\boldsymbol{\beta}_k$ (in the likelihood), is small. In order to interpret the weight on $\boldsymbol{\beta}_k^{\mathrm{P}}$ we need to inspect Equation 8.5 (page 171): From this equation we see that the term θs_k^+ is inversely proportional to the (conditional) variance of $\boldsymbol{\beta}_k$ under the prior. Therefore the weight on $\boldsymbol{\beta}_k^{\mathrm{P}}$ is large when $\boldsymbol{\beta}_k$ has large prior variance; it is the counterpart of $\frac{1}{\sigma_k^2}$ in the weight on $\boldsymbol{\beta}_k^{\mathrm{ML}}$.

9.2.4 Uncertainty Estimates

Once the Gibbs sampler has been implemented, no additional effort is needed to estimate model-based standard errors or confidence intervals for the forecast. This is done by producing, at every iteration of the Gibbs sampler (after the "burn-in" period), a forecast for each cross-section based on the current sample from $\boldsymbol{\beta}$ and adding to it a random disturbance, sampled from a normal distribution with standard deviation given by the current sample from σ. The standard deviation for this random

set of forecasts will give us an estimate of the standard errors. Of course, model-based uncertainty estimates do not take into account the most important source of error, which is the specification error, for which other techniques must be used.

9.3 Concluding Remark

This chapter offers a method of computing forecasts from our model in Chapter 4 given the choice of any of the priors in Chapters 5 or 7. The following chapter offers speedier versions that do not rely on the Gibbs algorithm.

Chapter 10

Fast Estimation Without Markov Chains

In Chapter 9, we focused on the Gibbs sampling algorithm, since our goal was to compute the mean of the posterior for $\boldsymbol{\beta}$ (Equation 9.2, Page 188). However, given the posterior distribution $\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid m)$ in Equation 9.11, we could alternatively use its maximum as a point estimate, the so-called Maximum A Posteriori (MAP) estimator.

Thus, from the (marginal) posterior distribution of the coefficients, $\mathcal{P}(\boldsymbol{\beta}|y)$, we could compute

$$\beta_{\text{MAP}} \equiv \arg\max_{\boldsymbol{\beta}} \mathcal{P}(\boldsymbol{\beta}|y) = \arg\max_{\boldsymbol{\beta}} \int \mathcal{P}(\boldsymbol{\beta}, \sigma, \theta \mid y) d\sigma d\theta. \tag{10.1}$$

We show here that we can compute this quantity or approximations to it without Gibbs sampling, as the solution to a maximization problem. Whether this MAP estimator is better than the mean of the posterior, and whether it is easier to implement it may depend on the problem and on the kind of software and expertise available to the user.

In the three main sections of this Chapter, we offer three alternative estimators — a maximum a posteriori (MAP) estimator, a marginal MAP estimator, and a conditional MAP estimator. In applications, we have used only the conditional MAP estimator to any large extent, but we present all three so that a user interested in implementing any of them will have many of the necessary calculations already done. All three estimators also provide additional insights about the model.

10.1 Maximum A Posteriori Estimator

Instead of maximizing the marginal posterior for $\boldsymbol{\beta}$ we could alternatively maximize the whole posterior, obtaining estimates for σ and θ as well. This has the advantage

of enabling one to compute forecast standard errors easily via simulation. In order to compute this estimator we have to solve the following maximization problem:

$$\max_{\boldsymbol{\beta},\sigma,\theta} \mathcal{P}(\boldsymbol{\beta},\sigma,\theta \mid y) \tag{10.2}$$

It is now straightforward to write the first order conditions for this problem:

$$\begin{aligned}
\boldsymbol{\beta}_k &= \left(\frac{\mathbf{X}_k'\mathbf{X}_k}{\sigma_k^2} + \theta s_k^+ \mathbf{C}_{kk}\right)^{-1} \left(\frac{\mathbf{X}_k' y_k}{\sigma_k^2} + \theta \sum_j s_{jk}\mathbf{C}_{kj}\boldsymbol{\beta}_j\right), \\
\sigma_k^{-2} &= \frac{\mathbb{d} + T_k - 2}{\mathbb{e} + \mathrm{SSE}_k}, \\
\theta &= \frac{\mathbb{f} + r - 2}{\mathbb{g} + \sum_{ij} W_{ij}\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j}.
\end{aligned}$$

Thus, we have written the equation for the coefficient $\boldsymbol{\beta}_k$ in such a way that $\boldsymbol{\beta}_k$ is not on the right side (since $s_{kk} = 0$). This suggests a simple iteration algorithm in which one starts, for example, with $\boldsymbol{\beta}$ and σ obtained via equation-by-equation least squares and then updates the estimates for all three parameters according to the expressions above until some suitable convergence condition is satisfied.

10.2 Marginal Maximum A Posteriori Estimator

We begin by factoring the marginal posterior into the product of two terms:

$$\mathcal{P}(\boldsymbol{\beta}|y) \propto \left[\int d\sigma \mathcal{P}(m \mid \boldsymbol{\beta},\sigma)\mathcal{P}(\sigma)\right]\left[\int d\theta \mathcal{P}(\boldsymbol{\beta} \mid \theta)\mathcal{P}(\theta)\right].$$

The first term is often called an "effective" likelihood, that is a likelihood in which we already have incorporated the effect of the uncertainty about σ by integrating it out. Similarly, the second term can be thought of as an "effective" or marginal prior, which also takes in account the fact that the parameters of the prior are known with uncertainty. The integrals in the expressions above can be readily computed:

$$\int d\sigma \mathcal{P}(m \mid \boldsymbol{\beta},\sigma)\mathcal{P}(\sigma) \propto \prod_i \left(\frac{1}{\mathbb{e} + \mathrm{SSE}_i(\boldsymbol{\beta})}\right)^{\frac{\mathbb{d}+T_i}{2}}$$

where we have defined SSE_i in Equation 9.12 (page 191). Similarly we have:

$$\int d\theta \mathcal{P}(\boldsymbol{\beta} \mid \theta)\mathcal{P}(\theta) \propto \left(\frac{1}{\mathbb{g} + H[\boldsymbol{\beta}]}\right)^{\frac{\mathbb{f}+r}{2}}$$

where

$$H[\boldsymbol{\beta}] \equiv \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j.$$

Finally, we plug the expressions above in Equation 10.1, and take its log, which gives the marginal MAP estimator:

$$\beta_{\text{mMAP}} \equiv \arg\min_{\boldsymbol{\beta}} \sum_i (\mathbb{d} + T_i) \ln\left(1 + \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}\right) + (\mathbb{f} + r) \ln\left(1 + \frac{H[\boldsymbol{\beta}]}{\mathbb{g}}\right) \quad (10.3)$$

10.3 Conditional Maximum A Posteriori Estimator

We now compare Equation 10.3 with the one we would obtain if we had conditioned on the fixed values of the parameters σ_i and θ (corresponding to degenerate, point mass priors):

$$\beta_{cMAP} \equiv \arg\min_{\boldsymbol{\beta}} \sum_i \frac{1}{\sigma_i^2} \text{SSE}_i(\boldsymbol{\beta}) + \theta H(\boldsymbol{\beta}). \quad (10.4)$$

By comparing Equations 10.4 and 10.3, we first observe that the effect of uncertainty on σ_i and θ is to replace the quadratic cost functions in Equation 10.4 with the concave-shaped cost functions of Equation 10.3. Other functions than the logarithm could be obtained (for example, the absolute value) by choosing different densities for σ and θ (see Girosi, 1991, for a characterization of the class of functions that can be obtained in this way).

The effect of having concave functions in the likelihood is to provide some robustness against variation in squared error over the cross-sections: If one cross-section has large squared error and we use the estimator of Equation 10.4, we may end up with an estimator too focused on making the error in that particular cross-section small, at the expense of the squared error in the other cross-sections. Using Equation 10.3 instead will prevent such outliers from having undue effects.

When $\mathbb{e}$ is "large" with respect to SSE_k, we have $\ln\left(1 + \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}\right) \approx \frac{\text{SSE}_i(\boldsymbol{\beta})}{\mathbb{e}}$. This makes sense, since we can make $\mathbb{e}$ large, for example, by choosing a prior distribution for σ_i^{-2} with low variance, which is closer to the case considered in Equation 10.4 (with 0 variance). Similarly, if our prior for θ has low variance, so that we are fairly certain about the value of θ, the prior term in Equation 10.3 will tend to the corresponding term in Equation 10.4.

The role of the quantities in Equation 10.3 — $\mathbb{d}$, $\mathbb{f}$, r, and T_i — is also clear. The parameters $\mathbb{d}$ and $\mathbb{f}$ control the concentration of the densities for σ_i^{-2} and θ around their means. Therefore we expect them to appear as weights in Equation 10.3. The number T_i is the number of observations in cross-section i, and it appears as a weight

in the likelihood so that cross-sections with more observations are smoothed relatively less. The number r is the rank of the inverse covariance matrix in the prior for $\boldsymbol{\beta}$, and it reflects the amount of information carried by the prior (if r is small then the prior is constant on most of its domain). Notice that the counterpart of r in the likelihood is T_i, which measures how much information is in the likelihood.

Equation 10.3 provides an alternative estimator to the posterior mean. It does not require Gibbs sampling, but it requires minimizing a non-convex cost function. When σ_i^{-2} and θ have small variance, and therefore Equation 10.3 approaches Equation 10.4, it might be simpler to solve Equation 10.3 than to use the Gibbs sampler. However, since the objective function is not convex it is difficult to exclude the possibility of local minima which can can lead to sub-optimal solutions. On the other hand, Gibbs sampling algorithms are only guaranteed to converge asymptotically and so, as is usual with MCMC techniques, its never entirely clear when they have converged. Another disadvantage of this approach is that, unlike the Gibbs sampling, it does not provide model-based standard errors, since it does not provide estimates for σ and θ.

10.4 Concluding Remarks

This chapter offers three fast algorithms for estimating the mode of the posterior. Whether these point estimators or those in Chapter 9 are preferable in any one application is an empirical question. The estimators offered here are faster and also require less expertise of users than the MCMC algorithm in Chapter 9.

Part IV
Empirical Evidence

In this Part, we use a large and diverse collection of age-sex-country-cause-specific mortality data to illustrate our method and demonstrate how it improves forecasts.

Chapter 11

Examples

[The written portions of this chapter are currently under active revision and the others are incomplete. We hope to have it finished in a few weeks.]

We now offer a practical guide to making forecasts in real empirical data. We begin in Section 11.2 with the simple case of a forecasting model that uses a linear trend as the only covariate. We then discuss nonlinear trends in Section 11.3, and finally two examples using substantive covariates in Sections 11.4 and 11.5. Each of the data sets discussed here are fully documented and available with the software package we wrote to accompany this book, and a replication data file with all the computer codes we made is also available. The idea of this chapter is that by making the replication of the results presented here easy, it should be straightforward for users to extend the approach to different data sources.

11.1 Forecasting Choices

Our general approach throughout this book is to enable forecasters to include more information in their model, in order to improve forecasting accuracy. In addition to including covariates that may differ from one cross-section to the next, new information can enter the model in five ways, each of which represents a opportunity and choice for the user. This chapter delineates these choices and shows how they are made in practice. The choices include:

1. The baseline age profile, $\bar{\mu}$. This is the age profile towards which the forecasts will tend, due to the smoothing imposed.

2. The cross-sections across which we expect to see similar levels or trends in log-mortality. These may include priors for smoothness across age, time, and country, or interactions and combinations of these.

3. The amount of similarity and thus smoothing we will see (the weight of the prior, θ, which can be asertained by the standard deviation of the prior for a single forecast).

4. Where we should smooth more or less. By choosing weights, this enables users to loosen up the prior and let the data take over at chosen places within the age profile, time periods, or regions.

5. What we are ignorant about. This is the content of the null space, controlled by the degree of the derivative of the smoothness functional or functionals, $\mathtt{m}$.

11.2 Forecasts without Covariates: Linear Trends

In this section, we consider a simple application where forecasts in each cross-section are modeled as a linear trend over time. Thus, we include here only a single covariate representing the time period (the year), and exclude all other covariate information. The results here differ from that of Lee and Carter, a random walk with drift, or a least squares regression of log-mortality on the year primarily because of the priors we include that tie the forecasts in different cross-sections together in different ways.

This simple linear trend model is an especially useful starting point in practice for understanding a new set of data as well as for learning our general approach. It is also useful for real forecasting examples when time series are very short and autoregressive methods cannot be productively applied, or when covariates are unreliable or unavailable. In some data, such as U.S. male mortality, the observed data tend to be highly linear and so we might expect forecasts to be linear in time there as well.

The specification of expected mortality varying over age a and time t is therefore as follows:

$$\mu_{at} = \beta_a^{(0)} + \beta_a^{(1)} t \tag{11.1}$$

11.2.1 Smoothing over age groups only

Here we consider the case of acute respiratory infections in Belize, a time series with 23 observations. In Figure 11.1, we show in the top two panels 30 year forecasts of time series and age profiles obtained using the Lee-Carter method. A linear trend computed using least squares produces similar results. The forecast plot on the left shows the results to be highly variable, with each of the age group forecast lines shooting off in a different direction. The implausible results produce the highly unsmooth expected age profiles on the right. If the input data are noisy, substantive priors might yield large confidence intervals but not a highly jagged expected age profile. Of course, this figure is not surprising: The Lee-Carter method was never meant to be applied to such data, and least squares completely ignores any correlation among age groups.

Therefore, it is natural to introduce a non-zero-mean prior which smooths over the age groups. We consider the usual case with a derivative of order 2 and uniform weights over the age groups and over time, which we write below:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}H[\mu,\theta]\right) \tag{11.2}$$

$$H[\mu,\theta] \equiv \frac{\theta}{AT}\int_0^T dt \int_0^A da \left(\frac{d^2}{da^2}(\mu(a,t) - \bar{\mu}(a))\right)^2 \tag{11.3}$$

The choice of the second derivative is dictated by the choice of the null space, which we do not wish to be too restrictive. Using a second derivative means that the null space of this prior — the portion of the forecasts that depend entirely on the data and are not influence by the prior — is the set of age profiles that deviate linearly from the average age profile. Had we chosen a first derivative, the null space would include only age profiles which differ from the average age profile by a constant, a consraint which we find too strong.

Once the order of the derivative has been chosen, the only two remaining unknown quantities in this prior are the smoothness parameter θ and the mean age profile $\bar{\mu}(a)$). The mean age profile is a form of prior knowledge, which in principle could be elicited from an experts' panel. In practice we find that a satisfactory procedure is to use the empirical average of all the age profiles for respiratory infections in males in our data base, where the average is taken over all the countries with more than 10 observations (to exclude countries with very unreliable data).

Following the logic of Section 6.2 (Page 113), with $\theta \propto 1/\sigma^2$, we recognize σ as the standard deviation of the prior for any one log-mortality prediction (for one age, sex, country, cause, and year). We therefore chose the value $\sigma = 0.2$ for the prior, although we find that $\sigma = 0.1$ gives very similar results.

the bottom panel of Figure 11.1 gives 30 year forecasts and age profiles obtained using our Bayesian method. The coefficients were estimated using the MAP procedure described in Chapter 10, rather than the full Gibbs sampling from Chapter 9. While this result could be improved, it is already a substantial improvement on the forecasts shown in the top panel and faster to run than full Gibbs.

A side effect of the age profile smoothing is a relatively smooth pattern of time trends. This happens because if log-mortality were to decrease at significantly different rates in different age groups that might lead to age profiles which deviate non-smoothly from the average age profile. The mathematical reason for this can be seen by plugging specification 11.1 in prior 11.3, and deriving the prior imposed on the coefficients $\beta^{(0)}$ and $\beta^{(1)}$:

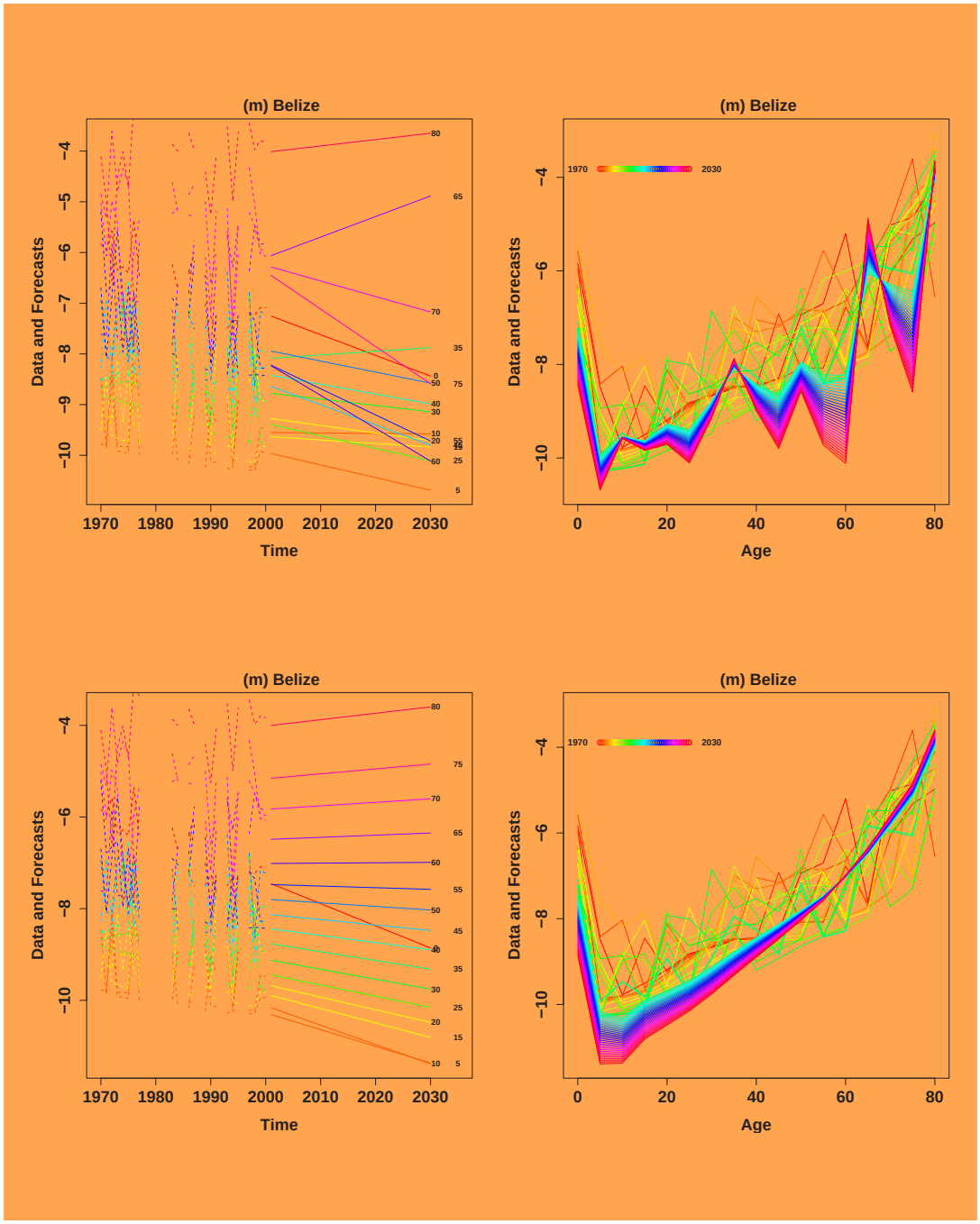

Figure 11.1: Log-mortality from respiratory infections in males in Belize. The time series includes 23 observations. The top two panels are the forecasts for time series (left) and age profiles (right) obtained using Lee-Carter. The bottom two are forecasts for time series (left) and age profiles (right) using the prior of Equation 11.3, which smooths over age groups. The standard deviation of the prior is set to 0.2, and the average age profile is the empirical average of all the age profiles for respiratory infections in males in our data base, taken over all the countries with more than 10 observations. Note how the priors used in the bottom two panels constrains the forecasts to be more plausible relative to each other.

$$
\begin{aligned}
H[\mu,\theta] &\equiv \frac{\theta}{AT}\int_0^T dt \int_0^A da \left(\frac{d^2}{da^2}(\mu(a,t)-\bar{\mu}(a))\right)^2 = \\
&= \frac{\theta}{AT}\int_0^T dt \int_0^A da \left(\frac{d^2}{da^2}(\beta_a^{(0)}+\beta_a^{(1)}t-\bar{\mu}(a))\right)^2 = \\
&= \frac{\theta}{A}\int_0^A da \left(\frac{d^2}{da^2}(\beta_a^{(0)}-\bar{\mu}(a))\right)^2 + \\
&+ \frac{\theta T^2}{3A}\int_0^A da \left(\frac{d^2}{da^2}\beta_a^{(1)}\right)^2 + \\
&+ \frac{\theta T}{A}\int_0^A da \left(\frac{d^2}{da^2}\beta_a^{(1)}\right)\left(\frac{d^2}{da^2}(\beta_a^{(0)}-\bar{\mu}(a))\right)
\end{aligned}
\tag{11.4}
$$

In the last expression in Equation 11.4, it is the second term that enforces smoothness of the time trend $\beta^{(1)}$ over the age groups. In many applications, this may not be enough to satisfy our needs, as another example will illustrate below, but the main point here is to realize that enforcing smoothness over age groups does enforce sometimes some smoothness of the trends over age groups.

Another way to look at the same phenomenon is to analyze in detail the null space of the prior 11.4. Simple algebra shows that in the limit as $\theta \to \infty$ the forecast assumes the following specification:

$$\mu_{at} = \bar{\mu}(a) + \alpha_1 + \alpha_2 t + \gamma_1 a + \gamma_2 at$$

The presence of the term with γ_2 in this expression shows that in the null space the time trend can only vary linearly with age. When a finite value of θ is used, the time trend will have more freedom to vary, but it will still be constrained up to a point.

In many applications, the user will know that, in addition to the age profiles being smooth, the time trend varies smoothly across age groups too, and so the amount of smoothing imposed by the prior over age groups may be insufficient. An example of such a case is shown in Figure 11.2, which has the same structure of Figure 11.1, but uses data from Bulgaria rather than Belize. In this case while the use of the prior 11.3 leads to smooth age profiles, it still leads to an unrealistic ranking of mortality across age groups. This problem originates from an unrealistic pattern in the time trend. We thus consider a prior that explicitly smooths the time trend over age groups in the following section.

11.2.2 Smoothing over age and time

One kind of prior knowledge is smoothness of the time trend over age groups. We have described a general smoothness functional that captures this type of prior information in section 7.4.1. A simple example of such a smoothness functional is the following:

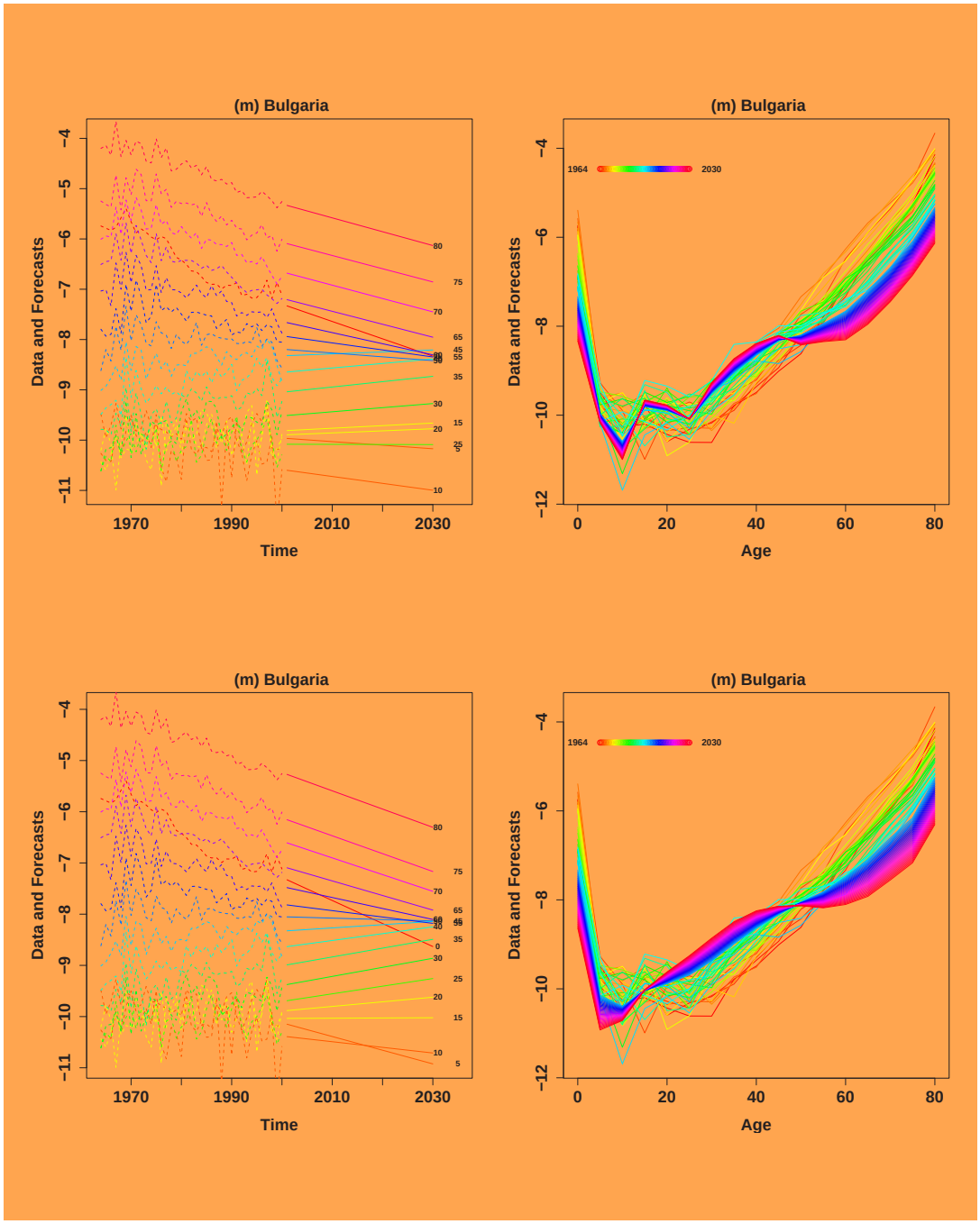

Figure 11.2: Log-mortality for respiratory infections in males, Bulgaria. There are 37 observations in this time series. The top two panels are the forecasts for time series (left) and age profiles (right) obtained using the Lee-Carter method. The bottom two panels are the forecasts for time series (left) and age profiles (right) obtained using the prior of equation 11.3, which smooths over age groups. The standard deviation of the prior has been set to 0.2, and the average age profile is the empirical average of all the age profiles for respiratory infections in males in our data base, taken over all the countries with more than 10 observations.

$$H[\mu, \theta] \equiv \frac{\theta}{TA} \int_0^T dt \int_0^A da \left(\frac{\partial^2 \mu(a,t)}{\partial a \partial t} \right)^2 . \tag{11.5}$$

The interpretation of this prior is very simple, and is best understood by plugging specification 11.1 into Equation 11.5, and obtaining the following prior for the coefficients $\boldsymbol{\beta}$:

$$H[\boldsymbol{\beta}, \theta] \equiv \frac{\theta}{A} \int_0^A da \left(\frac{\partial \beta_a^{(1)}}{\partial a} \right)^2 . \tag{11.6}$$

This prior is indifferent to the values of the intercept coefficients $\beta^{(0)}$, but it does enforce the constraint that similar age groups have similar time trends (i.e., $\beta^{(1)}$). Because this prior is indifferent to the intercept $\beta^{(0)}$, it does not enforce any kind of smoothness on the age profiles: The time trends could be exactly the same across ages, but the age profiles themselves could be far from being smooth. Therefore it will often be useful to use this prior in conjucntion with a prior like the one of Equation 11.3, which smooths the age profiles, in order to improve the behavior of the time trends.

We now demonstrate the use of this prior in conjunction with the age prior in Figure 11.3, which refers to the data for respiratory infections in Bulgarian males of Figure 11.2, where we showed the results obtained using the age prior by itself. Here we add the prior 11.6 to the age prior, using a value of the standard deviation $\sigma = 0.05$. This standard deviation follows the procedure described in Section 7.6 (Page 158). The main feature of the forecast in Figure 11.3 is that the basic shape of the age profile is kept similar in the far future in that the age profiles does not develop the curious hump seen in Figure 11.2.

11.3 Forecasts without Covariates: Nonlinear Trends

It is not unccomon to find ourselves in the situation of having to do a forecast for which no covariates are available and for which auto-regressive method cannot be used due to the shortness of the time series. While using a linear time trend is often an acceptable solution, in many instances the time series has clearly a non-linear component. In these instances we are forced to choose a functional form for the non-linear part of the time series. It is best if this choice is guided by at least some plausible modeling argument. For example, in the case of mortaliyt one could reason as follows.

Consider the case in which mortality historically has been decreasing over time. When we assume a linear trend in log-mortality we are implicitely assuming the following model:

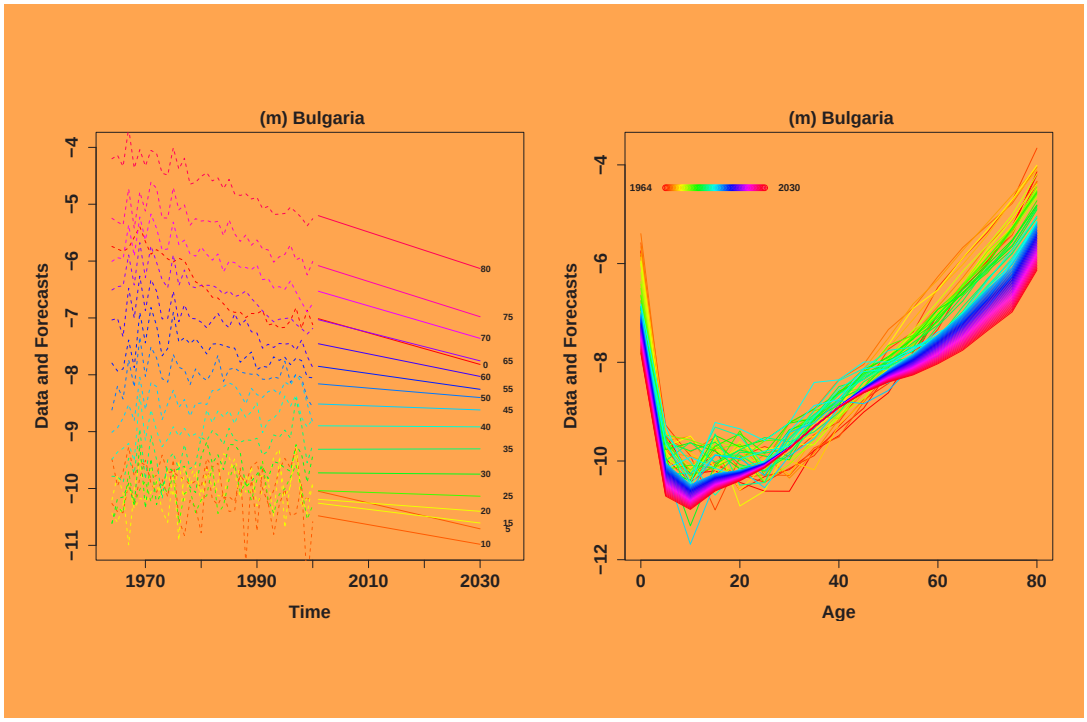

Figure 11.3: Log-mortality for respiratory infections in males, Bulgaria. There are 37 observations in this time series. Forecasts for time series (left) and age profiles (right) obtained using the prior of equation 11.3, which smooths over ages, in conjuction with the prior of equation 11.5, which smooths over ages and time. The standard deviation of the prior over ages was set to 0.2, while the standard deviation of the prior over age and time wast to 0.05.

$$\frac{d\mu(t)}{dt} = -\beta$$

where β is a positive coefficient which summarizes the effect of all the factors which drive mortality down over time, such as advances in medicine and medical care or improvements in public health infrastructure linked to economic growth. In many cases, however, there are also factors which drive mortality upward, such as increased access to tobacco or drugs, or increased volume of road traffic. If these upward trends are linear they are already netted out in the actual value of the parameter β. However, it is plausible that that in some cases the strength of the upward trend is a decreasing function of time. This may happen, for example, because the causes underlying the upward trends have been identified and policies are put in place in order to counteract them. For example, traffic safety rule may be implemented as a response to increased traffic accidents. If this is the case then one could model the dynamics of log-mortality as follows:

$$\frac{d\mu(t)}{dt} = -\beta + \frac{\alpha_1}{(\alpha_2 + t)^{\alpha_3}}$$

where $\alpha_1, \alpha_3 > 0$ and α_2 depends on the choice of the "origin" of time. The last term represents a positive rate of increase in log-mortality which decreases over time at rate controlled by α_3. Choosing, for example, $\alpha_3 = 1$ and integrating over time, one concludes that under this model log-mortality evolves over time as follows:

$$\mu(t) = \beta_0 + \beta t + \alpha_1 \log(\alpha_2 + t)$$

where we expect the sign of the coefficients β and α_1 to be opposites.

In order to set the parameter α_2 one needs to make some assumption on how fast the upward mortality trend changes over time. Measuring time in years, it is easy to see that the percentage change of this term going from on year to the next is $\frac{1}{\alpha_2+t+1}$. Setting this term to some small number, say, 1%, one obtains that $\alpha_2 = 100 - t - 1$. For a time series whose average time variable is, say 1975, one would then derive $\alpha_2 = -1876$, so that the nonlinear trend is $\log(t - 1876)$. In practice one does not need to be very precise, since these choice does not dramatically affect the results, as long as we get the right order of magnitude (a similar comments holds true for the value of α_3).

We apply these ideas to the forecast of death by lung cancer. In the past 50 years many countries have shown a mortality pattern for this disease which has an inverse U-shape. This is consistent with the story that as more people start smoking, more people die of lung cancer at a later date, triggering anti-smoking regulations aimed at decreasing lung cancer mortality rates. Therefore, in absence of a more appropriate covariate (for example, yearly tobacco consumption), we predict mortality by lung cancer using the specification:

$$\mu_{at} = \beta_a^{(0)} + \beta_a^{(1)} t + \beta_a^{(2)} \log(t - 1876)$$

In figures 11.4 and 11.5 we show forecasts of male mortality by lung cancer in Mauritius, Peru, Thailand and Lithuania obtained using OLS. We only consider age groups 30 and above since very few people die of lung cancer below age 30. Despite the fact that the regression specification has only 3 unknown parameters in each age group, OLS gives unacceptable results in all cases.
Here we consider a prior which includes three smoothness functionals:

$$\begin{aligned} H[\mu, \theta_{\text{age}}, \theta_{\text{time}}, \theta_{\text{age/time}}] \equiv & \frac{\theta_{\text{age}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{da^2} (\mu(a,t) - \bar{\mu}(a)) \right)^2 \\ & + \frac{\theta_{\text{time}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{dt^2} \mu(a,t) \right)^2 \\ & + \frac{\theta_{\text{age/time}}}{TA} \int_0^T dt \int_0^A da \left(\frac{\partial^3 \mu(a,t)}{\partial a \partial t^2} \right)^2 \end{aligned} \tag{11.7}$$

The first smoothness functional is the usual smoothness functional over age groups, that we have already discussed in section 11.2.1. The second smoothness functional smooths the time series over time. Since the functional involves the second derivative, it is used to make sure that the *curvature* of the time series stays within reasonable bounds. This functional is indifferent to linear trends, and therefore is only imposes a constraint on the logarithimic trend. The last term is a mixed age/time smoothness functional: it encodes knowledge that the curvature of the time series varies smoothly across age groups.

Now that the prior has been chosen we only need to choose reasonable values for the smoothness parameters θ_{age}, θ_{time} and $\theta_{\text{age/time}}$. Consistently with the notation of section 7.6, we re-parametrize our problem, and instead of the quantities θ_{age}, θ_{time} and $\theta_{\text{age/time}}$ we use the quantities σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$. We remind the reader that σ_{age} is simply the average standard deviation of the prior over age groups, if it were used in isolation, and it is linked to θ_{age} by equation 6.14, which we rewrite below (see equation 6.7 and 6.9 for the definition of other quantities in this formula):

$$\theta_{\text{age}} = \frac{\text{Tr}\left(\mathbf{Z} D_{\text{age}}^{+} \mathbf{Z}'\right)}{AT\sigma_{\text{age}}^2} \tag{11.8}$$

The same formula applies, with the obvious modifications, to θ_{time} and $\theta_{\text{age/time}}$.

In order to estimate the values of σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ we use the method described in section 7.6, which requires the definition of suitable summary measures. We use the following summary measures:

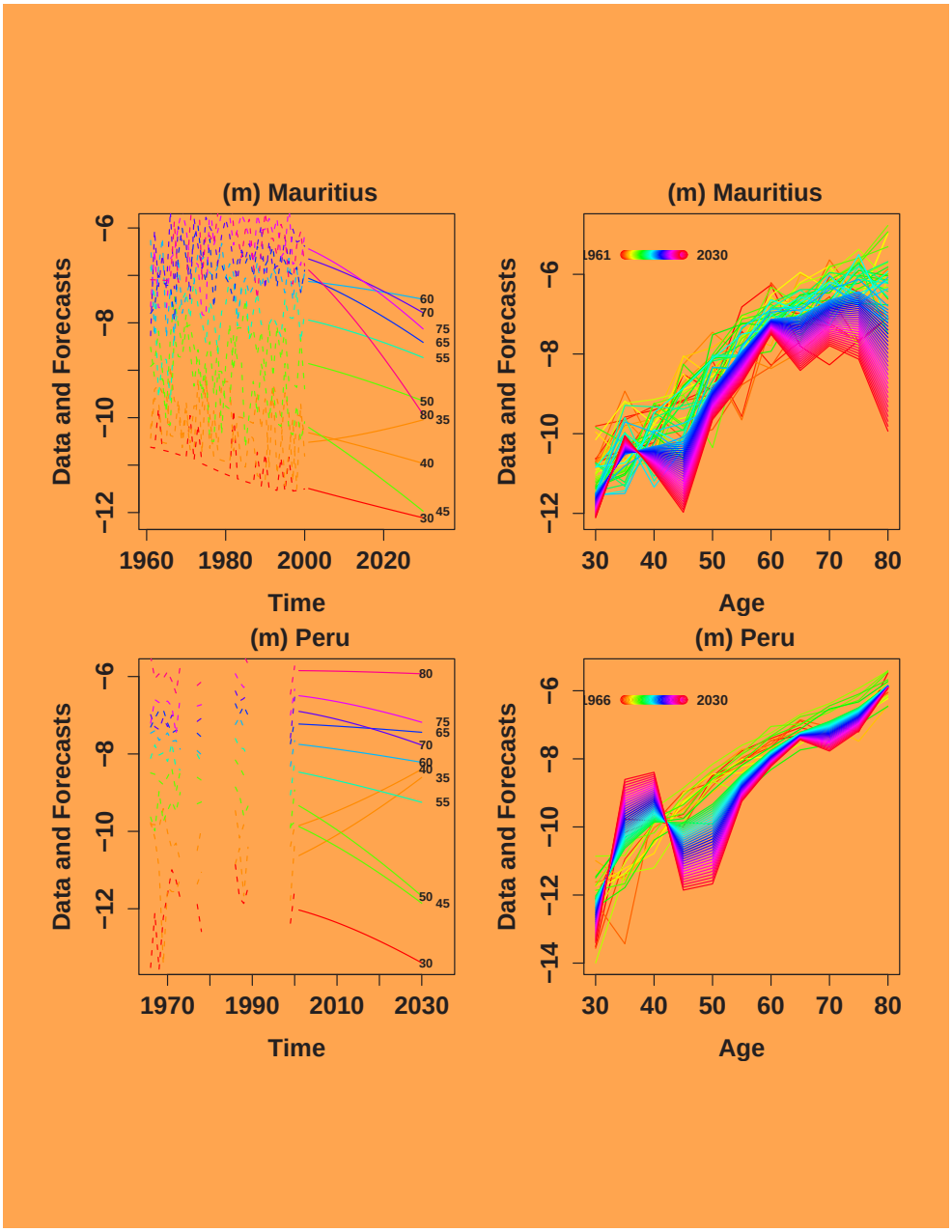

Figure 11.4: OLS forecasts of lung cancer log-mortality in males for Mauritius and Peru. In the left panels we plot the data and the forecast for age group 30 and above. In the right panels we plot the age profiles, both insample and out of sample.

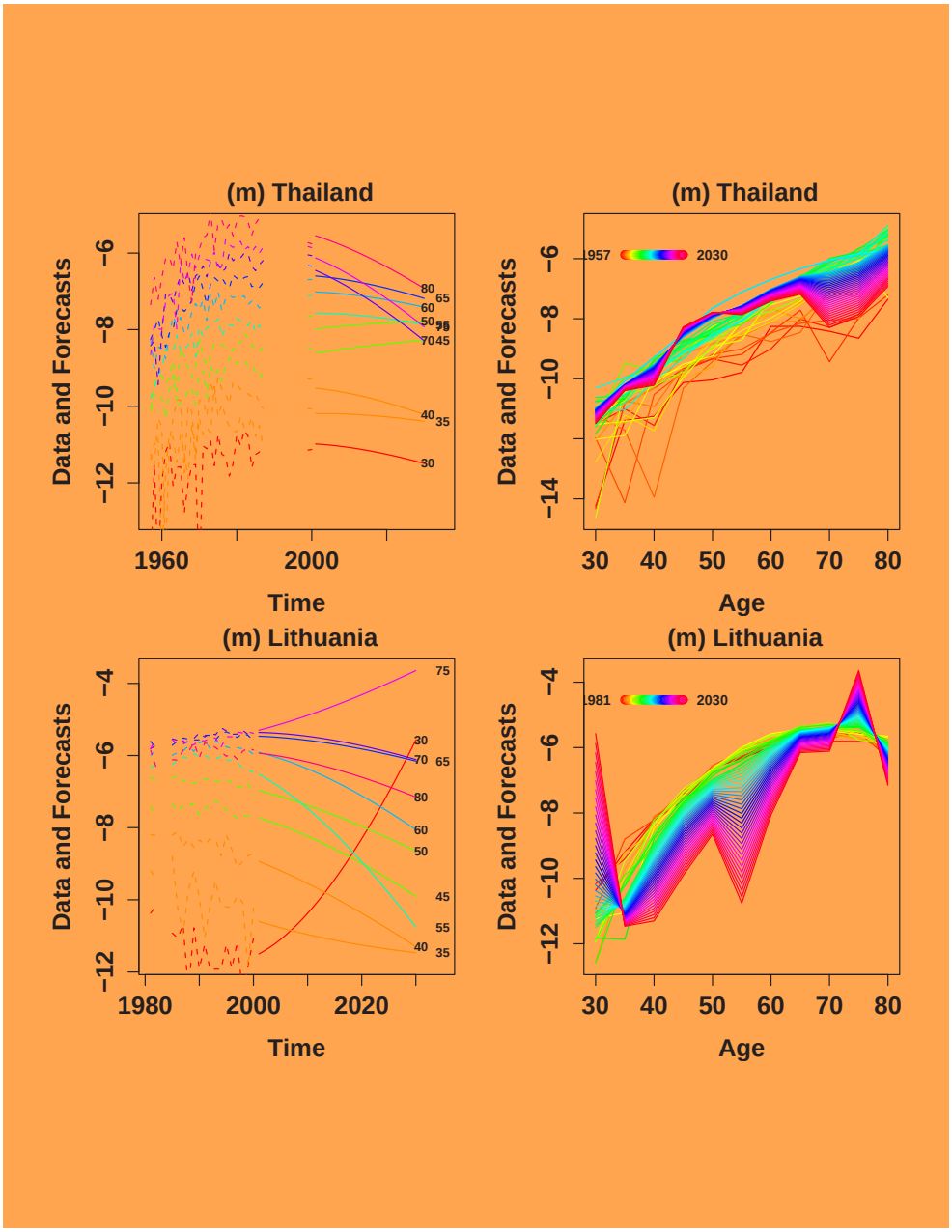

Figure 11.5: OLS forecasts of lung cancer log-mortality in males for Thailiand and Lithuania. In the left panels we plot the data and the forecast for age group 30 and above. In the right panels we plot the age profiles, both insample and out of sample.

$$\begin{aligned}
\mathrm{SD}(\mu) &\equiv \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T}(\mu_{at}-\bar{\mu}_a)^2 \\
F_{\mathrm{age}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|\mu_{at}-\mu_{a-1,t}| \\
F_{\mathrm{time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|\mu_{at}-\mu_{a,t-1}| \\
F_{\mathrm{age/time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=1}^{A}|(\mu_{at}-\mu_{a,t-1})-(\mu_{a-1,t}-\mu_{a-1,t-1})|
\end{aligned}$$

The summary measure SD is simply the average standard deviation of the prior, while F_{age} (F_{time}) measures how much log-mortality changes going from one age group (year) to the next. The quantity $F_{\mathrm{age/time}}$ is a measure of how much the time trend changes from one age group to the next.

Using the procedure described in section 7.6.2 we estimated that reasonable values for the expected values of the summary measures are approximately as follows:

$$\bar{\mathrm{SD}} \approx 0.33\ , \quad \bar{F}_{\mathrm{age}} \approx 0.56\ , \quad \bar{F}_{\mathrm{time}} \approx 0.029\ , \quad \bar{F}_{\mathrm{age/time}} \approx 0.006 \tag{11.9}$$

For each value of σ_{age}, σ_{time} and $\sigma_{\mathrm{age/time}}$ it is possible to draw samples from the prior defined by the smoothness functional 11.7 and compute the expected value of the summary measures above. Therefore the values of σ_{age}, σ_{time} and $\sigma_{\mathrm{age/time}}$ to be used in the forecast are those that generate expected values of the summary measures which are closest to the values in Equation 11.9.

In order to implement this method we considered values of σ_{age}, σ_{time} and $\sigma_{\mathrm{age/time}}$ in the following ranges:

$$\sigma_{\mathrm{age}} \in [0.1, 1]\ , \quad \sigma_{\mathrm{time}} \in [0.1, 2]\ , \quad \sigma_{\mathrm{age/time}} \in [0.01, 1]$$

We then let σ_{age}, σ_{time} and $\sigma_{\mathrm{age/time}}$ assume 10 equally spaced values, each in its own range, generating one thousand possible combinations. For each of these combinations we computed (empirically) the expected value of the 4 summary measures, storing the result is a table with 1,000 rows. Each row has the structure:

$$\left(\sigma_{\mathrm{age}},\ \sigma_{\mathrm{time}},\ \sigma_{\mathrm{age/time}},\ \bar{\mathrm{SD}}(\mu),\ \bar{F}_{\mathrm{age}}(\mu),\ \bar{F}_{\mathrm{time}}(\mu),\ \bar{F}_{\mathrm{age/time}}(\mu)\right)$$

A scatterplot of all the columns of such a table against each other is shown in Figure 11.6.

A cursory glance shows that the target values of Equation 11.9 are not strictly compatible with each other. In particular, it is difficult to find prior parameters which

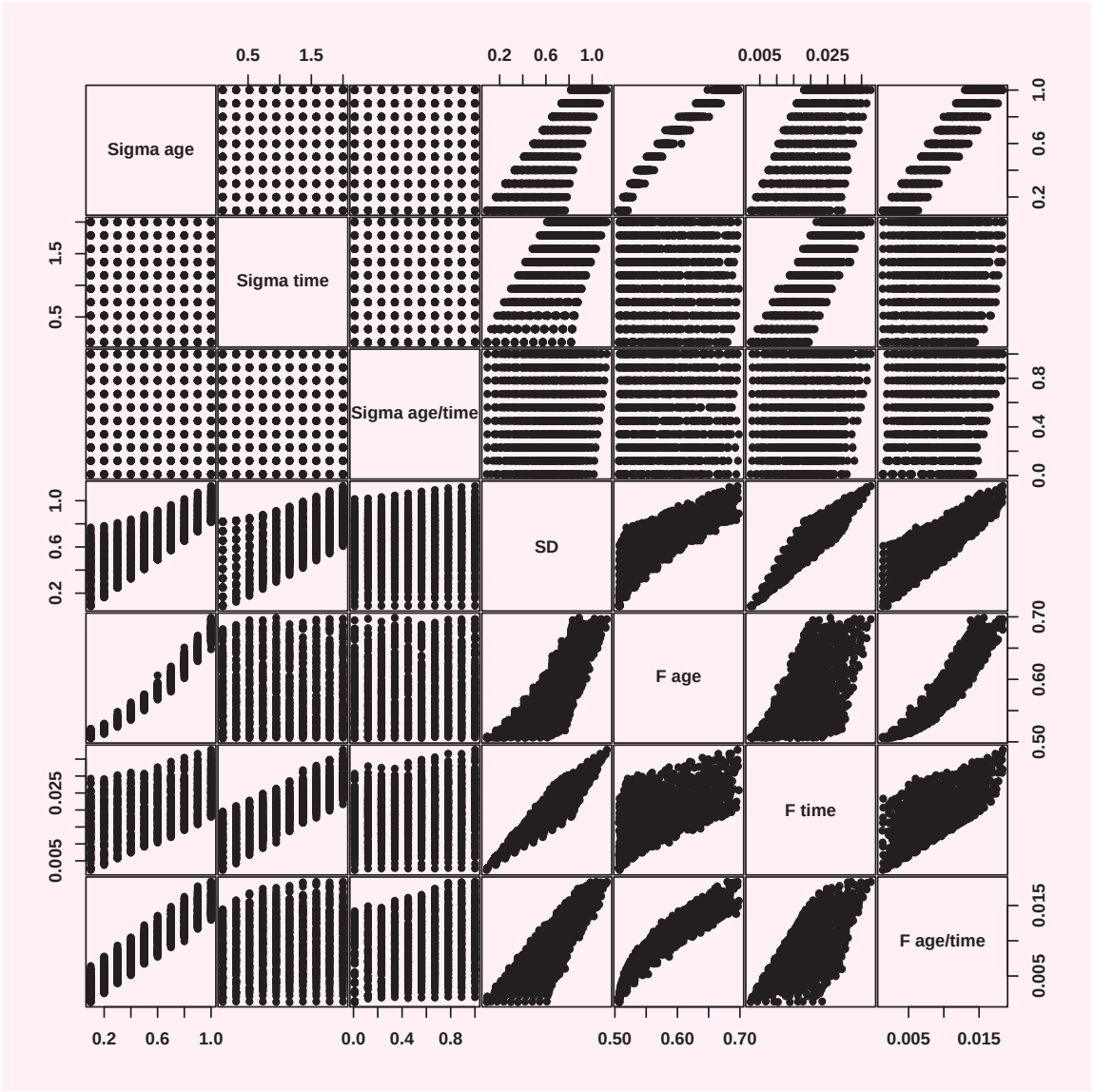

Figure 11.6: Pairwise plots of the relationship between the prior parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ and summary measures SD, F_{age}, F_{time} and $F_{\text{age/time}}$.

give a standard deviation SD of abour 0.3 and summary measures F_{age}, F_{time} and $F_{\text{age/time}}$ in the appropriate range. Therefore we search in our table of result for the set of parameters which deviates the least from the target values of Equation 11.9. Using the definition of distance provided in Section 7.6.1 we find the the optinal set of parameters is:

$$\sigma_{\text{age}} = 0.3 \ , \quad \sigma_{\text{time}} = 1.58 \ , \quad \sigma_{\text{age/time}} = 0.12 \tag{11.10}$$

which corresponds to the following values for the summary measures:

$$\bar{\text{SD}} \approx 0.54 \ , \quad \bar{F}_{\text{age}} \approx 0.53 \ , \quad \bar{F}_{\text{time}} \approx 0.022 \ , \quad \bar{F}_{\text{age/time}} \approx 0.005 \tag{11.11}$$

The optimal parameters of Equation 11.10 can now be plugged in our method, and produce the results shown in figures 11.7 and 11.8.

The forecasts shown in these figures are obviously not perfect, and do have some undesriable features. For example the forecast for Lithuania has an unlikely crossing of the time series at old ages. Similarly, the time series for the last age group seem to drop a bit too fast in the forecast for Thailand. However, one should bear in mind that these forecasts have been obtained as "first shot", without any expert's supervision. As usual, there is no substitute for good judgment, and some adjustment of the parameters in Equation 11.10 should be applied to produce forecasts which are more conform to our prior knowledge.

11.4 Forecasts with Few Covariates

11.5 Forecasts with Many Covariates

11.6 Concluding Remarks

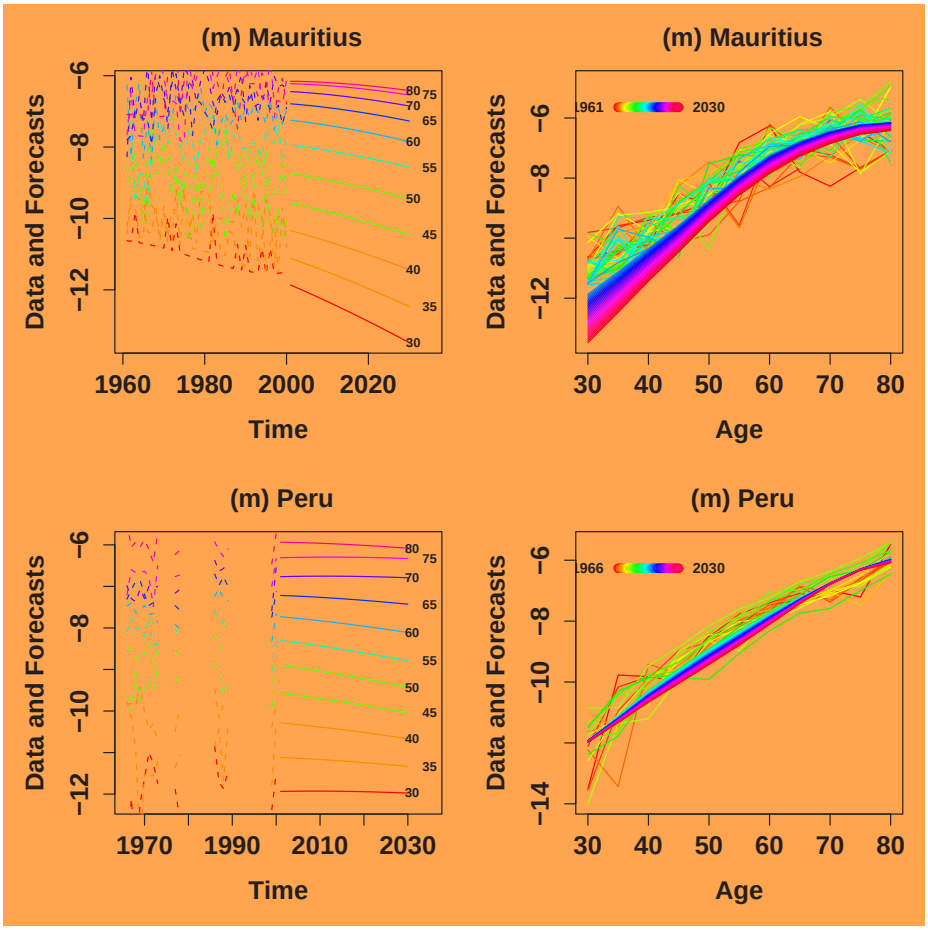

Figure 11.7: Forecasts of lung cancer log-mortality in males for Mauritius and Peru obtained using the Bayesian method. The prior used to obtain this forecast is given in Equation 11.7. The values of the prior parameters are given in Equation 11.10.

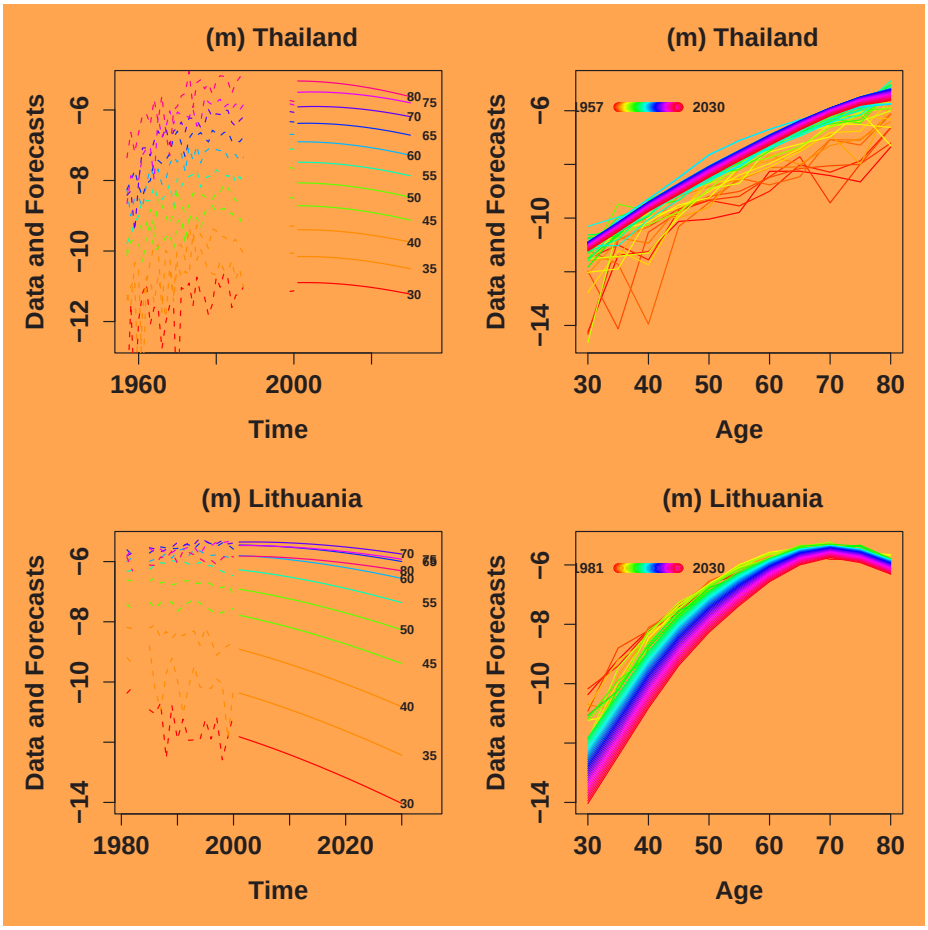

Figure 11.8: Forecasts of lung cancer log-mortality in males for Thailand and Lithuania obtained using the Bayesian method. The prior used to obtain this forecast is given in Equation 11.7. The values of the prior parameters are given in Equation 11.10.

Chapter 12

Concluding Remarks

We conclude with three brief points.

First, the framework we provide here can be used to improve forecasts beyond those in our empirical analyses. For example, the same ideas can be used to borrow strength between mortality rates for males and females, since we have prior knowledge about how long each live, and their relative frequencies of dying from particular causes. Using existing studies on co-morbidity, we could also include smoothness functionals that borrow strength across causes of death, or by smoothing across trends in similar causes.

Second, the procedures we introduce should be more widely applicable than for the important application of forecasting mortality. The statistical methods we introduce for modeling ignorance in Bayesian analysis, for smoothing based on the expected value of the dependent variable rather than the coefficients, and for connecting all hyperprior parameters to genuine knowledge available to subject matter experts should all extend beyond our application. And for any applications including mortality, they constitute a set of methods rather than just one, as well as a set of tools for constructing new methods and adapting the ones we have derived for new purposes.

Finally, as we emphasized in Chapter 1, no method can ever guarantee that real forecasts will be accurate in any application. Indeed, all good forecasters know perfectly well that unexpected events will sometimes occur and cause forecasts to fail miserably. In any one application, some models will be inappropriate; covariates can be picked badly so that they map idiosyncracies rather than systematic patterns likely to persist; choosing the wrong priors can cause us to propgate errors from neighboring cross-sections rather than to borrow statistical strength; and when the priors are not correct or strong enough to compensate, any of the usual problems with regression modeling can cause forecasts using these methods to miss their mark. In our view, the best anyone can do is to (a) gather as much information as possible, (b) ensure that in-sample modeling and out-of-sample forecasts include as much of this information

as possible, (c) verify in out-of-sample tests when a method works and when it fails, and (d) provide software to make sophisticated methods sufficiently intuitive so that they can be widely used and so researchers can build intuition and expertise in their use. These are the tasks we tried to accomplish in this book.

Part V

Appendices

Appendix A

Notation

A.1 Principles

Variables and Parameters We use Greek symbols for *unknown quantities*, such as regression coefficients ($\boldsymbol{\beta}$), expected values (μ), disturbances (ϵ), and variances (σ^2), and Roman symbols for *observed quantities*, such as y and m for the dependent variable, while the symbols $\mathbf{X}$ and $\mathbf{Z}$ refer to covariates.

Parameters that are *unknown, but are treated as known* rather than estimated, appear in the following font: $\mathtt{abcdef}$. Examples of these user-chosen parameters include the number of derivatives in a smoothing prior ($\mathtt{n}$) and some hyper-prior parameters (e.g., $\mathtt{f}, \mathtt{g}$).

Indices The indices $i, j = 1, \ldots, N$ refer to generic cross-sections. When the cross-sections are countries they may be labeled by the index $c = 1, \ldots C$, while when they are age groups, or specific ages, they may be labeled by the index $a = 1, \ldots, A$. Each cross-section also varies over time, which is indexed as $t = 1, \ldots, T$. Cross-sectional time series variables have the cross-sectional index (or indices) first and the time index last. For example, m_{it} denotes the value of the variable m in cross-section i at time t, and similarly m_{cat} is the value of the variable m in country c and age group a at time t.

Cross-section i contains k_i covariates. Therefore $\mathbf{Z}_{it}$ is a $k_i \times 1$ vector of covariates and $\boldsymbol{\beta}_i$ is a $k_i \times 1$ vector of coefficients. Every vector or matrix with one or more dimensions equal to k_i, such as $\mathbf{Z}_{it}$ or $\boldsymbol{\beta}_i$, will be in **bold**.

Dropping one index from a quantity with one or more indices implies taking the union over the dropped indices, possibly arranging the result in vector form. For example, if m_{it} is the observed value of the dependent variable in cross-section i at time t, then m_t is an $N \times 1$ column vector whose j-th element is m_{jt}. We refer to the vector m_t as the *cross-sectional profile* at time t. If the cross-sections i are age groups we call the vector m_t the *age profile* at time t. Applying the same in reverse,

we denote by m_i the $T \times 1$ column vector of the time series corresponding to cross-section i. Iterating this rule results in denoting by m the totality of elements m_{it}, and by $\boldsymbol{\beta}$ the totality of vectors $\boldsymbol{\beta}_i$. Similarly, $\mathbf{Z}_i$ denotes the standard $T \times k_i$ data matrix for cross-section i, with rows equal to the vector $\mathbf{Z}_{it}$.

If $\mathbf{X}$ is a vector, then diag[$\mathbf{X}$] is the diagonal matrix with $\mathbf{X}$ on its diagonal. If W is a matrix, then diag(W) is the column vector whose elements are the diagonal elements of W.

Sums We use the following short-hand for summation whenever it does not create confusion:

$$\sum_t \equiv \sum_{t=1}^{T}, \qquad \sum_i \equiv \sum_{i=1}^{N}, \qquad \sum_c \equiv \sum_{c=1}^{C}, \qquad \sum_a \equiv \sum_{a=1}^{A}$$

We also define the "summer" vector $\mathbf{1} \equiv (1, 1, \ldots, 1)$ so that for matrix X, $X\mathbf{1}$ denotes the row sums.

Norms For a matrix $\mathbf{x}$, we define the weighted Euclidean (or Mahalanobis) norm as $\|\mathbf{x}\|_\Phi^2 \equiv \mathbf{x}'\Phi\mathbf{x}$, with the standard Euclidean norm as a special case, so that $\|\mathbf{x}\|_I = \|\mathbf{x}\|$, with I as the identity matrix.

Functions We denote probability densities by capitalized symbols in calligraphic font. For example, the normal density with mean μ and standard deviation σ is $\mathcal{N}(\mu, \sigma^2)$. We denote generic probability densities by $\mathcal{P}$, and for ease of notation we distinguish one density from another only by their arguments. Therefore, for example, instead of writing $\mathcal{P}_\mathbf{x}(\mathbf{x})$ and $\mathcal{P}_\mathbf{z}(\mathbf{z})$ we simply write $\mathcal{P}(\mathbf{x})$ and $\mathcal{P}(\mathbf{z})$.

Sets Sets such as the real line $\mathbb{R}$ and its subsets ($\mathbb{S} \subset \mathbb{R}$), or the natural numbers $\mathbb{N}$ and the integers $\mathbb{Z}$ are denoted with these capital blackboard fonts. We denote the *null space* of a matrix, operator, or functional as $\mathfrak{N}$.

A.2 Glossary

- a: index for age groups.
- A: number of age groups.
- b_{it}: an exogenous weight for an observation at time t in cross-section i;
- $\boldsymbol{\beta}_i$: vector of regression coefficients for cross-section i;
- $\boldsymbol{\beta}_k^{\text{WLS}} \equiv (\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'y_k$, the vector of weighted least squares estimates.

- c: index for country.
- C: number of countries.
- d_{it}: the number of deaths in cross-sectional unit i occurring during time period t.
- δ_{ij}: Kronecker's delta function, equal to 1 if $i = j$ and 0 otherwise.
- $\mathrm{E}[\cdot]$: the expected value operator.
- ϵ: an error term.
- η: an error term.
- i: index for a generic cross-section (with examples being a for age, or c for country).
- I: the identity matrix (generic);
- I_d, $I_{d\times d}$: the $d \times d$ identity matrix;
- j: index for a generic cross-section.
- k_i: the number of covariates in cross-section i, and the dimension of all corresponding **bold-face** quantities, such as $\boldsymbol{\beta}_i$ and $\mathbf{Z}_{it}$.
- L: generic diagonal matrix.
- λ: mean of a Poisson event count (Section 3.1.1)
- $\ln(\cdot)$: the natural logarithm.
- M_{it}: mortality rate for cross-sectional unit i at time t: $M_{it} \equiv d_{it}/p_{it}$.
- m_{it}: a generic symbol for the observed value of the dependent variable in cross-section i at time t. When referring to an application, we use $m_{it} = \ln(M_{it})$, the natural log of the mortality rate.
- $\bar{m}_a$: The mean log-mortality age profile, averaging over time, $\sum_{t=1}^{T} m_{at}/T$.
- $\tilde{m}$: matrix of mean centered logged mortality rates, with elements $\bar{m}_{at} \equiv m_{at} - \frac{1}{T}\sum_t m_{at}$.
- μ_{it}: expected value of the dependent variable in cross-section i at time t;
- N: number of cross-sectional units.
- $\mathbb{N}$: the set of natural numbers.

- $\mathtt{m}$: generic order of the derivative of the smoothness functional.
- $\mathfrak{N}$: the null space of an operator or a functional.
- $\mathfrak{N}_\perp$: the orthogonal complement of the null space $\mathfrak{N}$.
- ν: an error term.
- $O_{q\times d}$: a $q \times d$ matrix of zeros;
- p_{it}: population (number of people) in cross-sectional unit i at the start of time period t.
- $\mathcal{P}$: denotes probability densities. The same $\mathcal{P}$ may refer to two different densities, with the meaning clarified from their arguments.
- Q: generic correlation matrix of the data.
- $\mathbb{R}$: the set of real numbers.
- s_{ij}: the weight describing how similar cross-sectional unit i is to cross-sectional unit j. This "similarity measure" s_{ij} is large when the two units are similar, except that, for convenience but without loss of generality, we set $s_{ii} = 0$.
- $s_i^+ \equiv \sum_j s_{ij}$. If s_{ij} is zero or one for all i and j, s_i^+ is known as the *degree* of i and is interpreted as the number of i's neighbors (or the number of edges connected to vertex i).
- Σ: an unknown covariance matrix.
- t: a generic time period (usually a year).
- T: total number of time periods (length of time series, when they all have the same length);
- T_i: number of observations for cross-section i (if $T_i = T_j \ , \ \forall i, j = 1, \ldots N$ then we set $T_i = T$);
- θ: The drift parameter in the Lee-Carter model. We reuse this symbol for the smoothing parameter in our approach.
- U_{it}: a missingness indicator equal to 0 if the dependent variable is missing in cross-section i at time t, and 1 if observed.
- V: generic orthogonal matrix.
- V[·]: the variance.

- W: A symmetric matrix constructed from the similarity matrix s. See Appendix B.2.6 (Page 253).
- $\mathbf{X}_{it} \equiv U_{it}\sqrt{b_{it}}\mathbf{Z}_{it}$, the explanatory variable vector ($\mathbf{X}_{it}$) weighted by the exogenous weights b_{it} when observed ($U_{it} = 1$) and 0 when missing.
- ξ denotes an error term.
- $x_\circ$: the projection of the vector x on some subspace.
- $x_\perp$: the projection of the vector x on the orthogonal complement of some subspace.
- $y_{it} \equiv U_{it}\sqrt{p_{it}}m_{it}$, the log-mortality rate (m_{it}) weighted by population (p_{it}), when observed ($U_{it} = 1$) and 0 when missing.
- $\mathbf{Z}_{it}$: a k_i-dimensional vector of covariates, for cross-sectional unit i at time t. The vector of covariates usually includes the constant.
- $\mathbf{Z}_i$: the $k_i \times T_i$ data matrix for cross-section i, whose rows are given by the vectors $\mathbf{Z}_{it}$;
- $\mathbb{Z}$: the set of integers.

Appendix B

Mathematical Refresher

This appendix presents mathematical concepts we use in developing our main arguments in the text of this book. This appendix can be read in the order in which it appears, or as a reference. Items are ordered so that simpler concepts appear earlier and, except where noted, each concept introduced does not depend on anything appearing after it.

B.1 Real Analysis

A *vector space* (Section B.1.1) is a set over which one can define a meaningful notion of a "sum" between two elements, and "multiplication of an element by a scalar". For our applications we impose additional structure on a vector space, and thus use *normed spaces* (Section B.1.3), which require also the notion of "length", and *scalar product and Euclidean spaces* (Section B.1.4), which add the notion of "projection". We only introduce the notion of a vector space here for its role as a basic building block for these other spaces, and do not use it by itself. We will see that the structure associated with a scalar product or Euclidean space is "more restrictive" than the one associated with a normed space, in the sense that with each Euclidean space we can associate in a natural way a normed space, but the opposite is in general not true.

Another useful construction is the *metric space* (Section B.1.2), which is a set over which we can define a notion of "distance" between two elements. One does not need the notion of sum between two elements in order to define the distance, and so a metric space is not necessarily a vector space. However, many vector spaces can be endowed with the structure of a metric space. In particular, with every scalar product, Euclidean, and normed spaces, we can always associate a metric space in a natural way.

B.1.1 Vector Space

Let X be a set where each element is defined as a vector, and where the following two operations are defined:

1. **Addition:** to every pair of vectors $x, y \in X$ corresponds an element $x + y \in X$ such that the communative and associative properties hold:

$$x + y = y + x \qquad \text{and} \qquad x + (y + z) = (x + y) + z \ , \quad \forall z \in X.$$

   In addition, X contains a unique vector, denoted by 0 (called the *zero vector*), such that $x + 0 = x \ , \forall x \in X$. Moreover, to each $x \in X$ corresponds a unique vector $-x$ such that $x + (-x) = 0$.

2. **Multiplication by a Scalar:** to every $a \in \mathbb{R}$ and $x \in X$ corresponds a vector ax such that:

$$1x = x \qquad \text{and} \qquad a(bx) = (ab)x \ , \quad \forall b \in \mathbb{R}$$

   In addition, the following distributive properties hold:

$$a(x + y) = ax + ay \quad \text{and} \quad (a + b)x = ax + bx \ , \quad \forall a, b \in \mathbb{R} \quad \forall x, y \in X$$

In this book if we refer to a set X as "the space X" we always mean "the vector space X".

Example 1 The set $\mathbb{R}^n$, $n = 1, 2, \ldots$ is a vector space with respect to the familiar operations of addition between two vectors and multiplication by a scalar.

Example 2 The set $C[a, b]$ of continuous functions over the interval $[a, b]$ is a vector space. Any continuous function on $[a, b]$ is a vector belonging to this vector space. The operations of addition and scalar multiplication corresponds to the usual operations of addition between two functions and multiplication of a function by a scalar.

Example 3 The three dimensional unit sphere $S^3 \equiv \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}$ is *not* a vector space with respect to the usual operations of addition between vectors and multiplication of a vector by a scalar. For example, the sum of two vectors on the sphere may be a vector which lies off of it.

B.1.2 Metric Space

A metric space is a pair (X, d), where X is a set and $d : X \times X \to [0, +\infty)$ is a function, called *distance* or *metric* with the following properties:

1. **Positiveness:** $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
2. **Symmetry:** $d(x, y) = d(y, x)$.
3. **Triangle Inequality:** $d(x, y) + d(y, z) \geq d(x, z)$.

A *semi-distance* is a distance except that $d(x, y) = 0$ does not imply that $x = y$.

Example 1 (R^n, d), where $d(x, y) = \sqrt{(x - y)'(x - y)}$ is a metric space. This distance is known as the *Euclidean distance*.

Example 2 (R, d), where $d(x, y) = \frac{|x-y|}{1+|x-y|}$ is a metric space.

Example 3 $(C[0, 1], d)$, where $d(f, g) = \max_{x\in[0,1]} |f(x) - g(x)|$, $f, g \in C[0, 1]$, is a metric space.

The set X of a metric space (X, d) does not have to be a vector space, as we see in the following:

Example 4 (S^n, d), where S^n is the n-dimensional sphere and $d(x, y)$ is the geodesic distance between two points (that is the distance measured along the shortest path) is a metric space. This metric space is not a vector space because there is no notion of a zero vector or addition on a sphere.

B.1.3 Normed Space

A vector space X is called a *normed space* if it is equipped with a *norm* $\|\cdot\|$. The norm is a function $\|\cdot\| : X \to [0, +\infty)$ with the following properties:

1. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$;
2. $\|ax\| = |a|\|x\|$ for all $a \in \mathbb{R}$;
3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

If condition (1) is replaced with the weaker condition that $\|x\| \geq 0$ then $\|x\|$ is called a *semi-norm*. The only difference between a semi-norm and a norm is that for a semi-norm it is possible that $\|x\| = 0$ without x being the zero vector.

A normed space is often denoted by the pair $(X, \|\cdot\|)$, since the same vector space X can be endowed with different norms. Every normed space $(X, \|\cdot\|)$ can be made into a metric space (X, d) by defining the distance $d(x, y) \equiv \|x - y\|$. We often refer to this distance as the distance *induced* by the norm $\|\cdot\|$.

Example 1 Denote a normed space as $(R^n, \|\cdot\|_A)$, where A is a strictly positive definite symmetric matrix (see B.2.3, Page 248), and we have defined,

$$\|x\|_A \equiv (x'Ax)^{\frac{1}{2}} .$$

which is known as the *Mahalanobis norm*. When $A = I$, this norm is called the *Euclidean norm* and is simply denoted by $\|x\|$.

Example 2 The pair $(R^n, \|\cdot\|_p)$, where $\|x\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ is a normed space if $p \geq 1$. This normed space is often denoted by $L_p(R^n)$, and the norm $\|\cdot\|_p$ is often referred to as the L_p norm.

Example 3 The set $(C[a,b], \|\cdot\|_{L_2})$, with the norm defined as

$$\|x\|_{L_2} \equiv \left(\int_a^b dx \ f(x)^2 \right)^{\frac{1}{2}} ,$$

is a normed space, which is usually denoted by $L_2[a,b]$.

Example 4 The set $(C[a,b], \|\cdot\|_{W_2^1})$, with the semi-norm defined as

$$\|x\|_{W_2^1} \equiv \left(\int_a^b dx \ \left(\frac{df(x)}{dx} \right)^2 \right)^{\frac{1}{2}} ,$$

is a semi-normed space, which is usually denoted by $W_2^1[a,b]$. In fact, $\|x\|_{W_2^1}$ is a semi-norm, rather than a norm, because there are functions which are not zero (like $f(x) = a$, for all $a \in \mathbb{R}$) but whose norm is zero.

B.1.4 Scalar Product Space

A vector space (or normed space) is called a *scalar product space* (or sometimes "inner product space") if, with each ordered pair of vectors x and y, we can associate a positive number (x, y), called the *scalar, or inner, product* of x and y, such that the following properties hold:

1. $(x, y) = (y, x)$, for all $x, y \in X$;
2. $(x + y, z) = (x, z) + (y, z)$, for all $x, y, z \in X$;
3. $(ax, y) = a(x, y)$ for all $x, y \in X$, $a \in \mathbb{R}$.
4. $(x, x) \geq 0$ and $(x, x) = 0$ only if $x = 0$.

If property (4) above is replaced by $(x, x) \geq 0$ the result is a semi-scalar, or semi-inner, product space. Since we can define different scalar products on the same vector space X, it is convenient to think of a scalar product space as a pair $(X, (\cdot, \cdot))$. To every scalar product space $(X, (\cdot, \cdot))$ we can associate in natural way a normed space $(X, \|\cdot\|)$ by defining the norm $\|x\| = \sqrt{(x, x)}$. We refer to this norm as the norm *induced* by the scalar product $(\cdot, \cdot)$. Therefore any scalar product space can be a normed space. The opposite of this proposition is not true in general, but it is true if the norm has the following property:

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2) \ , \quad \forall x, y \in X$$

When this is the case then it is possible to show that one can define a meaningful scalar product by setting $(x, y) \equiv \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2)$.

Example 1 Let A be any strictly positive definite $n \times n$ symmetric matrix. The set $\mathbb{R}^n$ with the following inner product

$$(x, y) \equiv x' A y \ , \quad \forall x, y \in \mathbb{R}^n$$

is a scalar product space. The norm induced by this scalar product is $\|x\|^2 = x' A x$ (see "Normed spaces", page 233). A *Euclidean space* is a scalar product space where $A = I$, and the norm induced by the scalar product in the Euclidean space is the Euclidean norm.

Example 2 Let $C[a, b]$ be the set of continuous functions on the interval $[a, b]$. We can make this set an inner product space by defining the following inner product between any two functions $f, g \in C[a, b]$:

$$(f, g) \equiv \int_a^b dx f(x) g(x)$$

Example 3 Let $C^1[a, b]$ be the set of continuous and differentiable functions on the interval $[a, b]$. We can make this set an inner product space by defining the following inner product between any two functions $f, g \in C[a, b]$:

$$(f, g) \equiv \int_a^b dx f(x) g(x) + \int_a^b dx \frac{df(x)}{dx} \frac{dg(x)}{dx}$$

Example 4 Let $C^1[a, b]$ be the set of continuous and differentiable functions on the interval $[a, b]$. We can make this set a semi-inner product space by defining the following semi-inner product between any two functions $f, g \in C[a, b]$:

$$(f, g) \equiv \int_a^b dx \frac{df(x)}{dx} \frac{dg(x)}{dx}$$

This semi-inner product naturally defines a semi-norm $\|f\| \equiv \sqrt{(f,f)}$, which coincides with the semi-norm defined in example 4 under the "Normed Spaces" heading (page 233).

Example 5 Let $M^{p,q}$ be the set of $p \times q$ matrices. This is a vector space which becomes an inner product space if endowed with the *Frobenius inner product* between any two of its elements A and B, which is defined as follows:

$$(A, B) \equiv \mathrm{tr}(AB')$$

The norm associated to this inner product is called the *Frobenius norm*. Given a generic matrix A its Frobenius norm can be easily computed as follows:

$$\|A\|^2 = \mathrm{tr}(AA') = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2$$

Therefore the square of the Frobenius norm of a matrix is the sum of the squares of its elements. The Frobenius norm can be used to define the distance between two matrices A and B by setting $d(A, B) \equiv \|A - B\|$. If the matrix B is an approximation of the matrix A, the Frobenius distance between A and B is a natural measure of the approximation error, since it coincides with the error given by a least squares criterion.

B.1.5 Functions, Mappings, and Operators

Let X and Y be two sets. A rule that associates each element $x \in X$ with a unique element $y \in Y$ is called a *mapping* from X into Y and is written as $y = f(x)$. The element y is called the *image* of x under the mapping f. The set X is the *domain* of the map f. The set of elements $y \in Y$ such that $y = f(x)$ for some $x \in X$ is called the *range* of f and is often denoted by $f(X)$. By definition $f(X) \subset Y$ (which, in words, means that that the set $f(X)$ is a subset of the set Y).

When $f(X) = Y$ we say that f maps X *onto* Y. If f maps X onto Y and, to every element $y \in Y$ we can associate a unique element $x \in X$ such that $f(x) = y$, then we say that f is *invertible*. In this case we denote the element $x \in X$ which corresponds to $y \in Y$ by $x = f^{-1}(y)$, and the mapping f^{-1} from Y to X is called the *inverse* of f.

Other words instead of "mapping" are sometimes used, depending on the properties of X and/or Y. For example when Y is the set of real numbers we also refer to f as a *real function* on X. Mappings are also called *operators*, although this word is usually reserved for cases in which neither X nor Y are the set of real numbers.

B.1.6 Functional

A *functional* is a real-valued function defined on a vector space X. When $X \subset \mathbb{R}^d$ this coincides with the definition of a real function on X.

Example 1 Let X be a normed space. The norm $\|x\|$ is a functional over X.

Example 2 Let $C[a,b]$ be the set of continuous functions on the interval $[a,b]$, and let $x_0 \in [a,b]$. Then, for $f \in C[a,b]$, we can define the functional $F[f] \equiv f(x_0)$.

Example 3 Let $C^1[a,b]$ be the set of functions whose first derivative is continuous on $[a,b]$. Then, for $f \in C^1[a,b]$, we can define the functional:

$$F[f] \equiv \int_a^b dx \ \left(\frac{df(x)}{dx}\right)^2 .$$

as one simple measure of the smoothness of the function f.

B.1.7 Span

Let $x_1, \ldots, x_l \in \mathbb{R}^d$ be l vectors in $\mathbb{R}^d$. The span of these vectors is a linear space defined by all the possible linear combinations of these vectors:

$$X \equiv \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=1}^{l} c_i x_i \ , \ c_i \in \mathbb{R} \ , i = 1, \ldots, l \right\}$$

We also say that the vectors $x_1, \ldots, x_l \in \mathbb{R}^d$ span the space X, or that X is spanned by the vectors $x_1, \ldots, x_l \in \mathbb{R}^d$.

B.1.8 Basis and Dimension

A basis for a vector space X is a set of vectors $x_1, \ldots, x_d$ which are linearly independent and that span X. If the vectors $x_1, \ldots, x_d$ form a basis for X then every vector $x \in X$ can be uniquely written as:

$$x = c_1 x_1 + c_2 x_2 + \cdots + c_d x_d, \qquad c_i \in \mathbb{R}, i = 1, \ldots, d$$

If a vector space X has a basis with d elements then the number d is known as the *dimension* of the vector space and denoted by $\dim(X)$. We then say that the vector space X is d-dimensional, or has dimension d, and write $\dim(X) = d$.

A vector space may have many different bases, but they must all have the same dimension. The number of elements of a basis can be infinite: in this case we refer to the vector space as an infinite dimensional vector space.

Example The vectors $x_1 = (1,0,0)$, $x_2 = (0,1,0)$ and $x_3 = (0,0,1)$ form a basis for $\mathbb{R}^3$. The vectors $y_1 = (2,1,5)$, $y_2 = (3,3,2)$ and $y_3 = (0,-1,6)$ also form a basis for $\mathbb{R}^3$. The vectors $y_1 = (2,1,5)$, $y_2 = (3,3,2)$ and $y_3 = (4,2,10)$ do not form a basis for $\mathbb{R}^3$, since they are not linearly independent ($y_3 = 2y_1$) and therefore do not span $\mathbb{R}^3$.

B.1.9 Orthonormality

Let $x_1, \ldots, x_l \in \mathbb{R}^d$ be l vectors in $\mathbb{R}^d$. We say that these vectors are orthonormal if they are mutually orthogonal and of length 1:

$$x_i' x_j = \delta_{ij} \ , \quad i, j = 1, \ldots, l$$

where, as always, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. If a set of orthonormal vectors forms a basis for a vector space X we refer to them as an orthonormal basis.

Example 1 The vectors $x_1 = (1, 0, 0)$ and $x_2 = (0, 1, 0)$ are orthonormal. However, they do not form an orthonormal basis for $\mathbb{R}^3$, since they do not span $\mathbb{R}^3$ and therefore they are not a basis.

Example 2 The vectors $x_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $x_2 = \frac{1}{\sqrt{2}}(1, -1)$ are orthonormal and form an orthonormal basis for $\mathbb{R}^2$.

B.1.10 Subspace

A set $Y \subset X$ (which should be read as "Y, which is a subset of the set X") is a *subspace* of the vector space X if Y itself is a vector space, with respect to the same operations.

Example 1 The set $\mathbb{R}^2$ is a subspace of $\mathbb{R}^4$, since $\mathbb{R}^2 \subset \mathbb{R}^4$ and both $\mathbb{R}^2$ and $\mathbb{R}^4$ are vector spaces.

Example 2 Let $a \in \mathbb{R}^3$ be a fixed column vector, and let V be the set $V \equiv \{x \in \mathbb{R}^3 \mid a'x = 0\}$. For any two elements $x, y \in V$ we have $x + y \in V$, and as such the set V is a subspace of $\mathbb{R}^3$. It is easy to see that V is a two dimensional plane going through the origin.

Example 3 Let $a \in \mathbb{R}^3$ be a fixed column vector and let V be the set $V \equiv \{x \in \mathbb{R}^3 \mid a'x = 1\}$. For any two elements $x, y \in V$ we have $x + y \notin V$. As such, the set V is *not* a subspace of $\mathbb{R}^3$. It is easy to see that V is a two dimensional plane which does *not* go through the origin.

Example 4 Let $C[a, b]$ be the set of continuous functions on the interval $[a, b]$. Polynomials of degree n are continuous functions, and the sum of two polynomials of degree n is also a polynomial of degree n. The set Π_n of polynomials of degree n, $n > 0$, is a subspace of $C[a, b]$.

Example 5 Let $M^{p,q}$ be the set of $p \times q$ matrices, and let $M_r^{p,q}$, with $r \leq \min(p,q)$, be the set of elements of $M^{p,q}$ with rank r. While $M^{p,q}$ is a vector space (of dimension pq), the subset $M_r^{p,q}$ of $M^{p,q}$ is *not* a subspace of $M^{p,q}$, since the sum of two matrices of rank r does not necessarily have rank r. ⊠

B.1.11 Orthogonal Complement

Subspaces of $\mathbb{R}^d$ (and of some infinite dimensional spaces) enjoy some particular properties of great usefulness in linear algebra. Let Y be an n dimensional subspace of $\mathbb{R}^d$, with $n < d$ and let us endow $\mathbb{R}^d$ with the Euclidean scalar product. Then we can define the *orthogonal complement* of Y in $\mathbb{R}^d$, and denote it by $Y_\perp$, as the set of all vectors in $\mathbb{R}^d$ which are orthogonal to every element $y \in Y$:

$$Y_\perp = \{x \in \mathbb{R}^d \mid x'y = 0 \ \ \forall y \in Y\}$$

The set $Y_\perp$ is a subspace of $\mathbb{R}^d$ and has dimension $r = d - n$. An important result is the following: if X is an inner product space and $Y \subset X$ is a subspace of X, then every element x of X has a unique representation as the sum of an element $x_\circ$ of Y and an element $x_\perp$ of $Y_\perp$. In other words, the following representation is unique:

$$\forall x \in X: \quad x = x_\circ + x_\perp \ , \quad x_\circ \in Y \ , \ \ x_\perp \in Y_\perp$$

The statement above is often summarized by writing: $X = Y \oplus Y_\perp$, where the symbol $\oplus$ is defined in Section B.1.12. The vectors $x_\circ$ and $x_\perp$ are called the *projections* of x onto Y and $Y_\perp$ respectively. Given a vector $x \in X$ and a subspace Y, we can always find the projection of x onto Y using the projection operator defined in Section B.1.13 (Page 240).

B.1.12 Direct sum

Let X be a vector space and let $Y, Z \subset X$ be subspaces of X. We say that X is the *direct sum* of Y and Z and write $X = Y \oplus Z$ if every $x \in X$ can be written in a unique way as:

$$x = y + z \ , \ \ y \in Y \ , \ \ z \in Z$$

If $X = Y \oplus Z$ then we say that Y and Z are *complementary* subspaces. It is important to note that if Y and Z are complementary subspaces then $Y \cap Z = 0$. The notion of a direct sum applies to generic vector spaces, and no inner product structure is required. When X is an inner product space the subspaces Y and Z are orthogonal complements (see Section B.1.11).

Example 1 Let $X = \mathbb{R}^3$, $Y = \{x \in X \mid x = (a, b, 0)\ , a, b \in \mathbb{R}\}$ and $Z = \{x \in X \mid x = (c, c, c)\ , c \in \mathbb{R}\}$. The subspace Y is the two dimensional (x_1, x_2) plane, and the subspace Z is the diagonal of the positive orthant (that is the line at a 45° angle with all the coordinate axis). Then Y and Z are complementary subspaces, and $X = Y \oplus Z$. This is equivalent to saying that that every vector in $\mathbb{R}^3$ can be uniquely written as the sum of a vector in the two dimensional plane and a vector "slanted" of 45° with respect to all the axis.

Example 2 Let $X = \mathbb{R}^3$, $Y = \{x \in X \mid x = (a, b, 0)\ , a, b \in \mathbb{R}\}$ and $Z = \{x \in X \mid x = (0, c, d)\ , c, d \in \mathbb{R}\}$. The subspace Y is the two dimensional (x_1, x_2) plane, and the subspace Z is the the two dimensional $(x_2 - x_3)$ plane. Although it is true that every $x \in X$ can be written as the sum of an element of Y and an element of Z, this representation is clearly not unique, and therefore X is not the direct sum of Y and Z, and Y and Z are *not* complementary subspaces (in fact, $Y \cap Z = \{x \in X \mid x = (0, b, 0)\ , b \in \mathbb{R}\} \neq \{0\}$).

B.1.13 Projection Operators

Definition and Properties Let X be a vector space, and let $P : X \to X$ be a linear operator which maps X into itself. The operator P is called a *projection operator*, or a *projector* if $P(Px) = Px\ , \forall x \in X$ (in short: $P^2 = P$). In all the cases in which we use projectors in this book the vector space X is $\mathbb{R}^d$, and therefore we can think of a projector P simply as an $d \times d$ matrix such that $P^2 = P$. In addition, where projectors are involved, we always assume that $\mathbb{R}^d$ is endowed with the usual Euclidean inner product. Projectors are very useful whenever we are given a vector $x \in X$ and are interested in picking the part of X which lies in a subspace Y (that is, which is "explained" by Y). In order to see the connection between projectors and subspaces, remember that if Y is a subspace of an inner product vector space X, then $X = Y \oplus Y_\perp$, that is any vector $x \in X$ can be uniquely written as $x = x_\circ + x_\perp$, with $x_\circ \in Y$ and $x_\perp \in Y_\perp$. This means that there is a well defined map $P_\circ$ that associates with the vector $x \in X$, the vector $x_\circ \in Y$, so that $x_\circ = P_\circ x$. This map is clearly linear and since $x_\circ \in Y$ it satisfies the identity $P_\circ(P_\circ x) = P_\circ x_\circ = x_\circ$. Therefore $P_\circ^2 = P_\circ$, and $P_\circ$ is a projector, which is often referred to as the projector of X onto Y. We also say that $x_\circ$ is the projection of x onto Y.

Example 1 Let $X = \mathbb{R}^3$ and Y be the (x_1, x_2) plane, that is $Y = \{x \in X \mid x = (a, b, 0)\ , a, b \in \mathbb{R}\}$. Since Y is a subspace of X, $X = Y \oplus Y_\perp$. The orthogonal complement $Y_\perp$ of Y is the vertical axis: $Y_\perp = \{x \in X \mid x = (0, 0, c)\ , c \in \mathbb{R}\}$. In words: any vector $x = (a, b, c)$ can be written as $x = (a, b, 0) + (0, 0, c)$, and the vectors $(a, b, 0)$ and $(0, 0, c)$ are orthogonal. Therefore the projection $x_\circ$ of $x = (a, b, c)$ onto Y is $x_\circ = (a, b, 0)$, and it is easily verified that the projector $P_\circ$ of X onto Y has the form:

$$P_\circ = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \boxtimes$$

If Y is a subspace of X, with its associated projector $P_\circ$, its orthogonal complement is also a subspace, and it has a corresponding projector that we define as $P_\perp$, with the property that if $x = x_\circ + x_\perp$ then $x_\perp = P_\perp x$. Therefore the two projectors $P_\circ$ and $P_\perp$ are related:

$$P_\circ + P_\perp = I.$$

Writing this relationship as $P_\perp = I - P_\circ$ makes clear that $P_\perp$ should be interpreted as the "residual making matrix" with respect to Y. In fact $x_\circ$ is the component of x "explained" by Y, and the residual is $x_\perp = x - x_\circ = P_\perp x$. Since the residual $x_\perp$ is orthogonal to $x_\circ$, it also follows that the two projectors $P_\circ$ and $P_\perp$ have the property that $P_\circ P_\perp = P_\perp P_\circ = 0$. We summarize the properties of these projectors as

$$P_\circ P_\circ = P_\circ \ , \quad P_\perp P_\perp = P_\perp \ , \quad P_\circ + P_\perp = I \ , \quad P_\circ P_\perp = P_\perp P_\circ = 0.$$

Constructing Projectors Assume that $X = \mathbb{R}^d$ and Y is an n-dimensional subspace of X. Let $\{u_1, \ldots, u_n \in X\}$ be a basis for Y, that is a set of n vectors which span Y (they do not have to be orthonormal). This means that any vector in Y can be written as a linear combination of the vectors $u_1, \ldots, u_n$. Our goal is to find a unique decomposition of an arbitrary vector x as $x_\circ + x_\perp$, with $x_\circ \in Y$ and $x_\perp \in Y_\perp$. Since $x_\circ \in Y$ then it can be written as $x_\circ = \sum_{i=1}^n a_i u_i$ for some coefficients a_i. Therefore our goal is to find, for every $x \in X$, coefficients a_i such that the following two conditions are satisfied:

$$x = \sum_{i=1}^n a_i u_i + x_\perp \ , \quad x'_\perp \sum_{i=1}^n a_i u_i = 0 \tag{B.1}$$

where the last condition ensures that $x_\perp$ and $x_\circ$ are orthogonal. Defining an n-dimensional vector $a = (a_1, \ldots, a_n)$ and the $n \times d$ matrix U with the vectors $u_1, \ldots, u_n$ on its columns the expression above can be rewritten as:

$$x = Ua + x_\perp \ , \quad x'_\perp Ua = 0 \tag{B.2}$$

Substituting $x_\perp = x - Ua$ in the orthogonality condition above, we conclude that the vector of coefficients a must satisfy:

$$(x - Ua)'Ua = 0 \ , \quad \Rightarrow a'(U'x - U'Ua) = 0$$

Since we know that the decomposition of Equation B.1 is unique, the solution of the equation above is obtained by setting $U'x - U'Ua$ to 0, so that we obtain $a =$

$(U'U)^{-1}U'x$. Therefore $x_\circ = Ua = U(U'U)^{-1}U'x$. This implies that the projector $P_\circ$ of X on the subspace Y is the following matrix:

$$P_\circ = U(U'U)^{-1}U'. \tag{B.3}$$

Therefore all we need in order to construct the projector $P_\circ$ onto an arbitrary subspace Y of X is a basis for the subspace Y. If the basis for Y is orthonormal then $U'U = I$, and the formula above simplifies to:

$$P_\circ = UU'. \tag{B.4}$$

The derivation of $P_\circ$ has an obvious interpretation in linear regression theory. In fact, Equation B.2 can be seen as a linear specification for the vector of observations x, in which a is the vector of unknown regression coefficients, U is the the matrix of covariates, and $x_\perp$ is a residual disturbance. The condition $x'_\perp Ua = 0$ expresses the well known fact that the residuals and the fitted values (Ua) are mutually orthogonal, and the residual making matrix $I - U(U'U)^{-1}U'$ is immediately identified with $P_\perp$. Equation B.3 is often called the "hat matrix."

The connection with linear regression helps to explain an important property of projection operators, and one of the reasons for which they are so useful in practice: the projection of x on the subspace Y is the vector of Y which has minimum distance from x. This follows from the observation that the vector of coefficients a which we have derived above is also the vector which minimize the least squares error $\|x-Ua\|^2$, which is exactly the Euclidean distance between x and a generic element Ua of Y. In other words, if we have a vector space X and a subspace Y, and we want to approximate a vector $x \in X$ with an element of Y, the solution of this problem is simply $P_\circ x$, where $P_\circ$ is the projector of X on Y and can be computed using Equation B.3.

Example 2 Let $X = \mathbb{R}^3$, $w \in X$ be a given vector of norm 1, and $Y = \{x \in X \mid w'x = 0\}$. Y is a two-dimensional subspace, and more precisely a two-dimensional "slanted" plane going through the origin. Since Y is constructed as the set of points which are orthogonal to the vector w, the orthogonal complement $Y_\perp$ of Y is simply the set of vectors which are multiples of w, that is a line through the origin. Let x be a generic point in X: we wish to find the closest point to x on the plane Y. From what we have seen above this point is simply $P_\circ x$, where $P_\circ$ is the projector of X onto Y. In order to find $P_\circ$ we need a basis for Y, which is not readily available. However, we have a basis for $Y_\perp$, which is given by w. Therefore we can find the projector $P_\perp$ of X on $Y_\perp$ and obtain $P_\circ$ as $I - P_\perp$. Applying formula B.4 we have $P_\perp = ww'$, and therefore:

$$P_\circ = I - ww'.$$

B.2 Linear Algebra

B.2.1 Range, Null Space, Rank, and Nullity

In order to understand the properties of a matrix it is important to understand the effect it has when it acts on a vector. Crucial to this understanding are two subspaces associated with a matrix: its range and its null space, which we now describe. Let A be a $q \times d$ matrix, and let x be a d-dimensional column vector. The first question we ask is what happens (that is, what kind of vectors do we obtain) when we operate with A on x? To answer this question we need to study the *range* of A, that is the set of all the vectors $y \in \mathbb{R}^q$ which can be written as $y = Ax$ for some $x \in \mathbb{R}^d$. Formally we have:

$$\text{range}(A) \equiv \{y \in \mathbb{R}^q \mid y = Ax \text{ , for some } x \in \mathbb{R}^d\} \tag{B.5}$$

Since the expression Ax can be read as "a linear combination of the columns of A with coefficients given by the components of x", we can also define the range of A as the vector space spanned by the columns of A. For this reason the range of A is often referred to as the *column space* of A. Denoting by $a_1, \dots a_d \in R^q$ the columns of A, this definition is formalized as follows:

$$\text{range}(A) \equiv \{y \in \mathbb{R}^q \mid y = \sum_{i=1}^{d} x_i a_i \text{ , for some } x_i \in \mathbb{R} \text{ , } i = 1, \dots, d\} \tag{B.6}$$

Let us assume that $q \geq d$: this definition makes clear that in this case the range of A is a vector space whose dimensionality r is at most d. The reason for which we say "at most" d, rather than equal to d, is that the columns of A may not be linearly independent, and therefore they may not span $\mathbb{R}^d$. Let us assume instead that $q < d$: then, since the d vectors a_i are q-dimensional, they can span at most a q-dimensional space. Therefore, by defining the *rank* of A as the dimensionality r of range(A), and denoting it by rank(A), we conclude that:

$$\text{rank}(A) \leq \min(q, d)$$

When rank$(A) = \min(q, d)$ we say that the matrix A has *full rank*, otherwise we say that is *rank deficient*.

To summarize: the matrix A takes $\mathbb{R}^d$ and maps into a subspace of $\mathbb{R}^q$ whose dimensionality is rank(A) and is at most d. A fundamental result of linear algebra, which we do not prove here but that we will use later, is the following:

$$\text{rank}(A) = \text{rank}(A') \tag{B.7}$$

We now present 4 examples which exhaust all the possibilities for the values of the rank of a matrix.

Example 1: $q \geq d$, full rank Consider the following matrix A:

$$A \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{B.8}$$

The rank of A is 2, since the two columns of A are linearly independent. Since $2 = \min(3, 2)$ the matrix has full rank. The range of A is a 2-dimensional subspace of $\mathbb{R}^3$, that is a slanted plane going through the origin. Every point on the slanted plane is the image of at least one point $x \in \mathbb{R}^2$ (see Section B.1.5 , Page 236 for definition of image). Vectors in $\mathbb{R}^3$ which are not in the range of A, that is do not lie on the slanted plane, are not the image of any point in $\mathbb{R}^2$.

Example 2: $q \geq d$, rank deficient Consider the following matrix A:

$$A \equiv \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \tag{B.9}$$

The rank of A is 1, since the second column is a multiple of the first. Since $1 < \min(3, 2)$ the matrix is rank deficient. The range of A is a 1-dimensional subspace of $\mathbb{R}^3$, that is the set of vectors x which are multiples of the vector $(1, 1, 1)$. This subspace is therefore the diagonal of the positive orthant. Every point on the diagonal is the image of at least a point $x \in \mathbb{R}^2$. Vectors in $\mathbb{R}^3$ which are not in the range of A, that is do not lie on the diagonal of the positive orthant, are not the image of any point in $\mathbb{R}^2$.

Example 3: $q < d$, full rank Consider the following matrix A:

$$A \equiv \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{B.10}$$

The rank of A is 2, since the third column of A is the sum of the first 2. Since the columns are 2-dimensional vectors the rank is as high as it can be and the matrix has full rank ($2 = \min(2, 3)$). The range of A is the entire space $\mathbb{R}^2$: every two-dimensional vector y is the image of at least one point x under A.

Example 4: $q < d$, rank deficient Consider the following matrix A:

$$A \equiv \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \tag{B.11}$$

The rank of A is 1, since the 2nd and 3rd column of A are both multiple of the first. Since $1 < \min(2, 3)$ the matrix is rank deficient. The range of A is not the entire space $\mathbb{R}^2$, but a one dimensional subspace, which is easily identified with the diagonal

of the positive quadrant: only points on the diagonal are images of points in $\mathbb{R}^3$ under A. ⊠

The range of a matrix is important because it allows us identify those vectors for which the matrix equation $Ax = y$ has at least one solution. Clearly, if y is not in the range of A then there is no solution and if y is in the range of A there is at least one solution. The reason for which there may be more than one solution is that there may be vectors $x_\circ$ such that $Ax_\circ = 0$. In this case, if we have $Ax = y$ then we also have $A(x + x_\circ) = y$. The set of vectors $x \in \mathbb{R}^d$ such that $Ax = 0$ is called the *null space* of the matrix A, and it is denoted by $\mathfrak{N}(A)$ (or just $\mathfrak{N}$ if no confusion arises). Formally we have:

$$\mathfrak{N}(A) \equiv \{x \in \mathbb{R}^d \mid Ax = 0\} \tag{B.12}$$

By rewriting the condition $Ax = 0$ as $\sum_{i=1}^d x_i a_i = 0$ we see that if the columns of A are linearly independent then $\mathfrak{N}(A) = \{0\}$, and we say that null space is trivial. In fact, if the columns of A are linearly independent the only numbers x_i such that $\sum_{i=1}^d x_i a_i = 0$ are zeros (if the x_i were not 0 one could express one of the a_i as a linear combination of the others). Therefore if A has full rank its null space is trivial, and if a solution to $Ax = y$ exists it is unique. When A is rank deficient we can expect a non-trivial null space: in this case the equation $Ax = y$ has an infinite set of solutions, which differ by an element of the null space of A (if x_1 and x_2 are solutions then $A(x_1 - x_2) = 0$ and $x_1 - x_2 \in \mathfrak{N}(A)$).

We will also need to know "how big" is the set of solutions making up the null space. To do this we note that $Ax = 0$ implies that x is orthogonal to every row of A, or every column of A', or any linear combination of columns of A'. This is equivalent to saying that x is orthogonal to the span of the columns of A', which in turn is the same as saying that x is orthogonal to the range(A') (since range(A') is the span of the columns of A'). This sequence of reasoning leads us to the key result that *the null space of a matrix is the orthogonal complement of the range of its transpose*:

$$\mathfrak{N}(A) = \text{range}(A')_\perp \tag{B.13}$$

This result is important because it allows us to compute the dimensionality of the null space of A, which is called the *nullity* of A and denoted by $\text{nullity}(A) \equiv \dim(\mathfrak{N})$. In fact, since the range of A' is a subspace of $\mathbb{R}^d$ of dimension $\text{rank}(A')$, we know that its orthogonal complement must have dimension $\dim(\text{range}(A')_\perp) = d - \text{rank}(A)$. Therefore we conclude with the fundamental decomposition:

$$\text{nullity}(A) = d - \text{rank}(A) \tag{B.14}$$

where throughout d is the number of columns of A. As anticipated above, then, the nullity of a matrix is zero (and the null space is trivial) only if the matrix has full rank.

The results of this section can be summarized as follows: the range of A allows us to characterize the vectors y for which the linear equation $Ax = y$ has a solution. The null space of A allows us to characterize whether this solution is unique, and in the case it is not, the whole set of solutions.

Example 1 (continued): $q \geq d$, full rank We have considered the matrix A:

$$A \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{B.15}$$

We have already seen that the rank of A is $2 = \min(3, 2)$, so that the matrix has full rank. Therefore its null space is trivial and every point in the range of A (the slanted plane) is the image of only one point in $\mathbb{R}^2$: the map between $\mathbb{R}^2$ and the slanted plane is one-to-one and therefore invertible. In fact, if $y \in \text{range}(A)$ we can solve $Ax = y$ with the usual formula:

$$x = (A'A)^{-1}A'y \tag{B.16}$$

Notice that this shows that if A has full rank then $A'A$ must be invertible.

Example 2 (continued): $q \geq d$, rank deficient We have considered the following matrix A:

$$A \equiv \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \tag{B.17}$$

We have already seen that A is rank deficient, since its rank is 1. As a consequence its null space is not trivial: from Equation B.14 we have that $\text{nullity}(A) = 2 - \text{rank}(A) = 1$, so $\mathfrak{N}(A)$ is one dimensional. We now define $\mathfrak{N}(A)$ explicitly. From Equation B.13 we know that it is the orthogonal complement of the range of A':

$$A' \equiv \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix}.$$

This expression confirms the fact that the rank of A and A' are the same, since only 1 of the three column vectors of A' is linearly independent of the others. The range of A' is therefore given by the multiples of the vector $(1, 2)$. Its orthogonal complement, that is $\mathfrak{N}(A)$, is easily seen to be the set of vectors which are multiples of the vector $(-2, 1)$ (since $(-2, 1)(1, 2)' = 0$). Therefore, every point in $\mathbb{R}^2$ is mapped into a point on the diagonal of the positive quadrant (range(A)), but all the points of the form $x + \alpha(-2, 1)$, for any $\alpha \in \mathbb{R}$, which lie on a straight line through x, are mapped into the same point on the diagonal. Therefore the solution to the linear equation $Ax = y$, when it exists, is not unique, and Equation B.16 does not hold anymore. This implies that when A is rank deficient the matrix $A'A$ must be not invertible. As we will see in Section B.2.5 in this case we can still give a meaning to the problem of solving $Ax = y$, but Equation B.16 must be replaced by something else.

B.2.2 Eigenvalues and Eigenvectors for Symmetric Matrices

Symmetric matrices and their eigenvectors and eigenvalues play a special role in this book, so we list here some of their properties.

Let A be a $d \times d$ symmetric matrix. If we can find a non-null vector $v \in \mathbb{R}^d$ and a number ℓ such that:

$$Av = \ell v$$

then we say that v is an eigenvector of A with eigenvalue ℓ. Notice that if v is an eigenvector with eigenvalue ℓ, then kv, with $k \in \mathbb{R}$, is also an eigenvector with eigenvalue ℓ. We eliminate this trivial redundancy by using the convention that eigenvectors always have length 1, so that $\|v\| = 1$ unless otherwise specified.

An important property of symmetric $d \times d$ matrices is that they always have exactly d mutually orthogonal eigenvectors $v_1, \ldots v_d$. We denote by $\ell_1, \ldots \ell_d$ the corresponding eigenvalues.

Let R be a $d \times d$ matrix with the eigenvectors of A, $v_1, \ldots v_d$, as its columns. By virtue of the orthogonality of the eigenvectors and the convention that they have length one it follows that R is an orthogonal matrix, so that $R' = R^{-1}$. Let L be the diagonal matrix with the eigenvalues $\ell_1, \ldots, \ell_d$ on the diagonal. One can prove that the matrix A can always be written as:

$$A = RLR' \tag{B.18}$$

Equation B.18, the *eigenvalue/eigenvector decomposition*, tells us everything we may need to know about the matrix A. If the eigenvalues are all strictly positive then the matrix is invertible, and the inverse is simply:

$$A^{-1} = RL^{-1}R'$$

where L^{-1} is a diagonal matrix with the reciprocals of the eigenvalues $(1/\ell_1, \ldots, 1/\ell_d)$ on the diagonal.

The rank r of A is the number of non-zero eigenvalues, and the nullity n of A is the number of zero eigenvalues. The eigenvectors corresponding to the non-zero eigenvalues, that is the first r columns of R, span the range of A, while the eigenvectors corresponding to the zero eigenvalues (the last n columns of R) span the null space of A.

When A does not have full rank, the decomposition B.18 can be written in a simplified form, often useful for computational purposes. Let us write the matrix L in block form:

$$L = \begin{pmatrix} \ell & 0 \\ 0 & 0 \end{pmatrix}, \qquad \ell = \text{diag}[(\ell_1, \ldots \ell_r)]$$

Let us also write R as $R = (R_\perp R_\circ)$, where $R_\perp$ is a $d \times r$ matrix whose columns are the first r eigenvectors (a basis for $\mathfrak{N}_\perp$), and $R_\circ$ is a $d \times n$ matrix whose columns are the last n eigenvectors (a basis for $\mathfrak{N}$). Then we have the following identity:

$$A = RLR' = (R_\perp R_\circ) \begin{pmatrix} \ell & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R'_\perp \\ R'_\circ \end{pmatrix} = R_\perp \ell R'_\perp . \tag{B.19}$$

B.2.3 Definiteness

Let A be an $n \times n$ symmetric matrix. If we have

$$x'Ax > 0 \ , \quad \forall x \in \mathbb{R}^n \ , \ x \neq 0$$

then we say that A is *positive definite*. If this condition is substituted with the weaker condition

$$x'Ax \geq 0 \ , \quad \forall x \in \mathbb{R}^n$$

then we say that A is *positive semi-definite*. The only difference between positive definite and positive semi-definite matrices is that for a positive semi-definite matrix A the fact that $x'Ax = 0$ does not imply $x = 0$. Similar definitions for negative definite and negative semi-definite matrices are obtained by switching the sign in the inequalities above.

If a matrix A is positive definite then the quantity $x'Ax$ is a norm, while if A is positive semi-definite then $x'Ax$ is a semi-norm.

It can be shown that A is positive definite if and only if its eigenvalues are all strictly positive, where we use the word "strictly" to emphasize the fact that they cannot be equal to 0. Similarly, A is positive semi-definite if and only if its eigenvalues are either positive or equal to zero.

Example 1 Let A be a $n \times d$ matrix of full rank. Then the $d \times d$ matrix $C = A'A$ is positive definite. This is seen by noticing that $x'Cx = x'A'Ax = (Ax)'Ax = \|Ax\|^2 > 0$. The strict inequality is a consequence of the fact that $Ax = 0$ implies $x = 0$, since A has full rank. If A has rank smaller than d then C is positive semi-definite.

B.2.4 Singular Values Decomposition (SVD)

For symmetric matrices the eigenvalue/eigenvector decomposition tells us everything we need to know about a matrix. The analog of this decomposition for generic, rectangular matrices is the Singular Value Decomposition (SVD). SVD is one the most useful and common tools of linear algebra, and it has been known for more than a century (Beltrami, 1873; Jordan, 1874; see Stewart, 1992, for its history). Here we give the basics facts, needed in this book, and refer the reader to a linear algebra book (e.g., Strang, 1988) for a full explanation.

Definition

Let A be a $q \times d$ matrix, with $q \geq d$ (if $q < d$ we can look at the transpose of A). It is possible to show that one can always write A as follows:

$$A = UWV' \tag{B.20}$$

where U is an $q \times d$ matrix whose columns are mutually orthonormal ($U'U = I_d$), W is a $d \times d$ diagonal matrix with positive or zero values, and V' is a $d \times d$ orthogonal matrix ($V'V = V'V = I_d$). The diagonal elements of W, denoted by $w_1, \ldots w_d$, are called the *singular values* of the matrix A. The decomposition B.20 is known as the *Singular Values Decomposition*, and it is unique, up to a permutation of the columns of U and V and of the corresponding singular values. Here we consider the general case in which n singular values are 0, with $n \geq 0$, and denote by r the number of non-zero singular values, so that $r + n = d$ and

$$w_1 \geq w_2 \geq \ldots w_r \geq w_{r+1} = w_{r+2} \ldots w_{r+n} = 0$$

The singular values play here the role played by the eigenvalues in the eigenvalue/eigenvector decomposition of symmetric matrices: The rank of A is equal to r, the number of non-zero singular values, and therefore the nullity of A is n, which is the number of singular values equal to zero. We now list some useful properties of the SVD decomposition. To this end, we define $u_1, \ldots u_d$ as the column vectors of U and $v_1, \ldots v_d$ as the column vectors of V.

- rank$(A) = r$;
- nullity$(A) = n$;
- The vectors $u_1, \ldots u_r$ (first r columns of U) form a basis for the range of A.
- The vectors $v_{r+1}, \ldots v_d$ (last n columns of V) form a basis for the null space of A.
- The following relationships hold:

$$Av_i = w_i u_i \ , \quad A'u_i = w_i v_i \ , \quad i = 1, \ldots d$$

- If the matrix A is square, then it is invertible if and only if the singular values are all strictly positive. In this case one can easily verify that the inverse of A is given by:

$$A^{-1} = VW^{-1}U'$$

For Approximation

We now show how to use SVD to approximate a matrix A as a linear combination of "simpler" matrices, and how to bound the corresponding approximation error. Using this notation Equation B.20 can be rewritten as:

$$A = \sum_{i=1}^{r} w_i u_i v_i' \equiv \sum_{i=1}^{r} w_i a_i \tag{B.21}$$

where we have defined the $q \times d$ matrices $a_i = u_i v_i'$, which are all of rank 1 (the fact that a_i has rank 1 follows from its SVD). Equation B.21 is a powerful result: it says that *any matrix A of rank r can be written as a linear combination of r matrices of rank 1.* Because of the orthonormality properties of U and V, the matrices a_i are mutually orthonormal under the Frobenius inner product (see examples in Appendix B.1.4, Page 234). In particular, they all have the same "size", where the size is measured by their Frobenius norm, which is equal to 1.

Therefore, if some singular values are much bigger than the others, then the corresponding terms in the expansion B.21 will dominate the others. This suggests that a good approximation to the matrix A can be obtained by retaining, in Equation B.21, only the largest singular values. To quantify this observation, define $\tilde{A}_k$ as the approximation of A obtained by retaining only the k ($k < r$) largest singular values in Equation B.21:

$$\tilde{A}_k \equiv \sum_{i=1}^{k} w_i a_i$$

We quantify the approximation error as $\|A - \tilde{A}_k\|^2$, where $\|\cdot\|$ is the Frobenius norm (see Appendix B.1.4, page 234). However, this definition of error is not always useful because it depends on the scale of matrix A. Therefore we introduce a relative measure of error, defined as

$$\Delta E_k \equiv \frac{\|A - \tilde{A}_k\|^2}{\|A\|^2}.$$

Substituting the expansion B.21 in the equation above we obtain:

$$\Delta E_k = \frac{\|\sum_{i=k+1}^{d} w_i a_i\|^2}{\|\sum_{i=1}^{d} w_i a_i\|^2} = \frac{\sum_{i=k+1}^{d}\sum_{j=k+1}^{d} w_i w_j (a_i, a_j)}{\sum_{i=1}^{d}\sum_{i=j}^{d} w_i w_j (a_i, a_j)} = \frac{\sum_{i=k+1}^{d} w_i^2}{\sum_{i=1}^{d} w_i^2}, \tag{B.22}$$

where we have used the definition of the Frobenius norm $\|\cdot\|$ in terms of Frobenius inner product $(\cdot,\cdot)$ ($\|A\|^2 = (A, A)$) and the orthonormality of the matrices a_i under the Frobenius inner product ($(a_i, a_j) = \delta_{ij}$). Equation B.22 is a very useful result: it allows us to estimate precisely the error we make approximating a matrix A of rank r by a linear combinations of k matrices of rank 1 only in terms of the singular values

of A: The faster the decay rate of the singular values, the better the approximation using a small number of terms. The relative error ΔE_k is usually referred to as "the percentage of the variance linearly accounted for by the first k singular values". The term "variance" here refers to the square of the Frobenius norm of A. The reason for this terminology is that in some applications the rows of A are realizations of a random d-dimensional variable with zero mean, and therefore the Frobenius norm of A (squared) is proportional to the variance of this random variable.

B.2.5 Generalized Inverse

Let A be a $d \times d$ matrix. In this section we consider the problem of finding a solution to the linear system $Ax = y$. We do not consider the more general problem in which A is rectangular to keep notation simple.

When A has full rank the solution to this problem is obviously unique and given by $x = A^{-1}y$. When A is rank deficient, with $\text{rank}(A) = r$, $r < d$, two things happen:

- the range of A is an r-dimensional subspace of $\mathbb{R}^d$, and therefore the linear system $Ax = y$ has a solution only if $y \in \text{range}(A)$;
- the null space of A is a subspace of $\mathbb{R}^d$ with dimensionality $n = d - r$ (see Equation B.14, Page 245). Therefore, when $y \in \text{range}(A)$ and a solution exist, it is not unique. In fact, if x is a solution then $x + x_\circ$, where $x_\circ \in \mathfrak{N}(A)$, is also a solution.

Here we assume that $y \in \text{range}(A)$ (otherwise an exact solution does not exist), and focus on the problem of having an infinite number of solutions. Even if all these solutions are equivalent, in the sense that they differ from each other for an element of $\mathfrak{N}(A)$, which is "invisible" to A, it is important to have a consistent criterion to pick a particular one, that we can consider as "representative". In order to choose a criterion we reason as follows.

Since $\mathfrak{N}(A)$ is a subspace of $\mathbb{R}^d$ we write $\mathbb{R}^d = \mathfrak{N}(A) \oplus \mathfrak{N}(A)_\perp$ (see Sections B.1.10 and B.1.11 on subspaces and orthogonal complements). Let x be such that $Ax = y$, and let us decompose it as $x = x_\circ + x_\perp$, with $x_\circ \in \mathfrak{N}(A)$ and $x_\perp \in \mathfrak{N}(A)_\perp$. Since x is a solution of $Ax = y$, by adding or subtracting any element of $\mathfrak{N}(A)$ we obtain another solution, and since $x_\circ \in \mathfrak{N}(A)$ therefore $x - x_\circ = x_\perp$ is also a solution. Therefore there is always a well defined solution which lies in $\mathfrak{N}(A)_\perp$, that is a solution whose projection on the null space of A is zero. We take this as the "representative" solution. We will see later how this is equivalent to choosing the solution of with minimum norm.

To summarize, we wish to find, among all the possible vectors x such that $Ax = y$, the one such $x_\circ = P_\circ x = 0$. This problem is solved easily using the SVD of A. First we notice that the condition $P_\circ x = 0$ can be written as $P_\perp x = x$, since $P_\circ + P_\perp = I$. Therefore we substitute $P_\perp x$ for x in $Ax = y$ and obtain $AP_\perp x = y$. Now let us use

the SVD of A and write $A = UWV'$, where U and V are orthogonal $d \times d$ matrices. The equation $AP_\perp x = y$ becomes:

$$UWV'P_\perp x = y \ , \Rightarrow \ WV'P_\perp x = U'y \ , \tag{B.23}$$

By the properties of the SVD the matrix W is diagonal, and since $\text{rank}(a) = r$, $\text{nullity}(A) = n$, and $d = r + n$, it has the following structure:

$$W \equiv \begin{pmatrix} w & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix}$$

where w is an $r \times r$ diagonal matrix with the first r non-zero singular values on the diagonal. Since our goal is to "isolate" $P_\perp x$ in Equation B.23, ideally we would multiply both sides of B.23 by the inverse of W. The inverse of W does not exist, but we can define something which resembles it and see whether that it is enough. We define the matrix W^+ below, which we list together with one useful identity it satisfies:

$$W^+ \equiv \begin{pmatrix} w^{-1} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix} \ , \quad W^+W = \begin{pmatrix} I_{r\times r} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix}$$

Now we premultiply both sides of Equation B.23 by W^+ and obtain:

$$\begin{pmatrix} I_{r\times r} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix} V'P_\perp x = W^+U'y$$

Now we remember that from SVD the matrix V is orthogonal and has the form $V = (V_\perp \ V_\circ)$, where $V_\circ$ is a basis for $\mathfrak{N}(A)$ and $V_\perp$ is a basis for $\mathfrak{N}(A)_\perp$. Premultiplying both sides of the equation above by V we obtain:

$$V \begin{pmatrix} I_{r\times r} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix} V'P_\perp x = (V_\perp \ \ V_0) \begin{pmatrix} I_{r\times r} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \end{pmatrix} \begin{pmatrix} V'_\perp \\ V'_\circ \end{pmatrix} P_\perp x$$

As a final step we remember that since the columns of $V_\perp$ form an orthonormal basis for $\mathfrak{N}(A)_\perp$, the projector $P_\perp$ on $\mathfrak{N}(A)$ is simply $P_\perp = V_\perp V'_\perp$. Therefore:

$$V_\perp V'_\perp P_\perp x = P_\perp P_\perp x = P_\perp x = VW^+U'y$$

Since we started with the assumption $P_\perp x = x$, the solution to our problem is finally:

$$x = VW^+U'y \equiv A^+y \tag{B.24}$$

where the matrix $A^+ = VW^+U'$ is the so-called *generalized inverse* of A. Notice that applying this definition the generalized inverse inverse of W is W^+, which justifies our notation. To summarize: among the infinite number of solutions of $Ax = y$, with $y \in \text{range}(A)$, the solution computed using the generalized inverse is the one whose projection on the null space of A is zero.

In many book the generalized inverse is defined in the same way, but it is derived according to a different criterion: among all the solutions of $Ax = y$ the solution computed using the generalized inverse is the one with minimum norm. We nowshow that these two criteria are equivalent. The set of all solutions can be obtained by adding to a known solution (for example $x^* = A^+ y$) arbitrary points in $\mathfrak{N}(A)$, which can always be written as $P_\circ z$, for arbitrary $z \in \mathbb{R}^d$. Therefore the set of solutions is the set of vectors which can be written as $x = x^* + P_\circ z$, with z varying in $\mathbb{R}^d$. Let us find the vector of this form with minimum norm:

$$\min_z \ \|x^* + P_\circ z\|^2 = \min_z \ [2(P_\circ z)' x^* + (P_\circ z)'(P_\circ z)].$$

We notice that $(P_\circ z)' x^* = z' P_\circ' x^* = z' P_\circ x^* = 0$, where we have used the fact that $P_\circ' = P_\circ$ and $P_\circ x^* = 0$ by definition of x^*. Therefore

$$\min_z \ \|x^* + P_\circ z\|^2 = \min_z \ (P_\circ z)'(P_\circ z)$$

The minimum is attained for any z such that $P_\circ z = 0$, and its value is $\|x^*\|^2$. Therefore the solution whose projection in the null space of A is 0 and the solution of minimum norm coincide.

B.2.6 Quadratic Form Identity

Let $b_1, \ldots, b_N \in R^d$ be a collection of N vectors in R^d. For any $N \times N$ symmetric matrix s and for any $d \times d$ symmetric matrix Φ the following identity holds:

$$\frac{1}{2} \sum_{i,j=1}^{N} s_{ij} (b_i - b_j)' \Phi (b_i - b_j) = \sum_{i,j=1}^{N} W_{ij} b_i' \Phi b_j \tag{B.25}$$

where

$$W \equiv s^+ - s \ , \quad s^+ \equiv \mathrm{diag}[s_i^+] \ , \quad s_i^+ \equiv \sum_{j=1}^{N} s_{ij} \ . \tag{B.26}$$

Since the rows of the matrix W sum to 0 ($W\mathbf{1} = 0$), W is not full rank. If the elements of s are all positive, the matrix W is positive semi-definite, but the reverse does not generally hold. The values of the expression in Equation B.25 do not depend on the diagonal elements of s. Therefore, for all expressions of the form in Equation B.25, we assume $s_{ii} = 0 \ , \forall i$. Under this assumption, given any matrix W such that $W\mathbf{1} = 0$, we can always find a matrix s such that $W \equiv s^+ - s$: it is sufficient to take $s = \mathrm{diag}(W) - W$.

A particular case of Equation B.25 which appears often in the book is when $d = 1$ and $\Phi = 1$. We restate it as follows. Let $u \equiv (u_1, \ldots, u_N)'$ be a column vector, then

$$\frac{1}{2} \sum_{i,j=1}^{N} s_{ij} (u_i - u_j)^2 = \sum_{i,j=1}^{N} W_{ij} u_i u_j = u' W u \tag{B.27}$$

B.3 Probability Densities

B.3.1 The Normal Distribution

Let D be a strictly positive definite $d \times d$ matrix and $\theta > 0$. We say that a d-dimensional random variable x has normal distribution with mean $\bar{x}$ and covariance D^{-1}, and write $x \sim \mathcal{N}(\bar{x}, D^{-1})$, if its probability density is:

$$\mathcal{P}(x) = \left(\frac{\theta}{2\pi}\right)^{\frac{d}{2}} \sqrt{\det D}\ \exp\left(-\frac{1}{2}\theta(x-\bar{x})'D(x-\bar{x})\right)$$

Since a density must integrate to 1 we have the multidimensional Gaussin integral:

$$\int_{\mathbb{R}^d} dx \exp\left(-\frac{1}{2}\theta x'Dx\right) = \left(\frac{2\pi}{\theta}\right)^{\frac{d}{2}} \frac{1}{\sqrt{\det D}} \tag{B.28}$$

B.3.2 The Gamma Distribution

We say that a random variable x with values on the positive real axis has a Gamma density, with $\mathbb{a}, \mathbb{b} > 0$, if its probability density is

$$\mathcal{P}(x) = \frac{1}{\Gamma(\mathbb{a})} \mathbb{b}^{\mathbb{a}} x^{\mathbb{a}-1} e^{-\mathbb{b}x} \quad . \tag{B.29}$$

We also write the equation above as:

$$x \sim \mathcal{G}(\mathbb{a}, \mathbb{b}) \tag{B.30}$$

The mean and variance of the Gamma$(\mathbb{a}, \mathbb{b})$ density are as follows:

$$\mathbb{E}[x] = \frac{\mathbb{a}}{\mathbb{b}}\ , \quad \mathbb{V}[x] = \frac{\mathbb{a}}{\mathbb{b}^2} \tag{B.31}$$

B.3.3 The Lognormal Distribution

We say that a random variable x with values on the positive real axis has a lognormal density with parameters ν and ϱ, and write $x \sim \log\mathcal{N}(\nu, \varrho^2)$, if its probability density is:

$$\mathcal{P}(x) = \frac{1}{\sqrt{2\pi}\varrho x} \exp\left[-\frac{1}{2}\left(\frac{\log x - \nu}{\varrho}\right)^2\right] \tag{B.32}$$

The lognormal density has the property that:

$$x \sim \log\mathcal{N}(\nu, \varrho^2) \Longleftrightarrow \log x \sim \mathcal{N}(\nu, \varrho^2)$$

The mean and variance of the lognormal density are as follows:

$$\mathrm{E}[x] = e^{\nu + \frac{\varrho^2}{2}} \ , \quad \mathrm{V}[x] = e^{2(\nu+\varrho^2)} - e^{2\nu+\varrho^2} \tag{B.33}$$

It is often useful to be able to express ν and ϱ as functions of the mean and the variance. This is done below:

$$\varrho^2 = \log\left(1 + \frac{\mathrm{V}[x]}{\mathrm{E}[x]^2}\right) \ , \quad \nu = \log\left(\frac{\mathrm{E}[x]^2}{\sqrt{\mathrm{V}[x] + \mathrm{E}[x]^2}}\right) \tag{B.34}$$

Appendix C

Improper Normal Priors

Most of the prior densities considered in this book are improper. The fact that they are improper has no negative consequences, since our likelihood and therefore our posterior distribution is always proper. What is relevant is the reason they are improper. Roughly speaking, the reason is that we only have partial prior knowledge: That is we know certain things but are ignorant about others. This implies that our prior density is flat over unbounded subsets of its support, which causes it to fail to be integrable.

C.1 Definitions

In this section we describe the mathematical tools needed to deal with the kinds of improper prior densities we use. These densities have the following form:

$$\mathcal{P}(x;\theta) \propto \exp\left(-\frac{1}{2}\theta x' D x\right) \ , \quad x \in \mathbb{R}^d \tag{C.1}$$

where $\theta > 0$ and D is a symmetric, $d \times d$ positive semi-definite matrix with $\text{rank}(D) = r < d$ and $\text{nullity}(D) = n = d - r$. The obvious way to see why this density is improper is to notice that its covariance, which would be calculated by taking the inverse of the matrix D, does not exist since D does not have full rank. We now develop a richer understanding of improperness through a geometric interpretation. The key to understanding and manipulating improper densities of the form C.1 is noting that, since D is symmetric, it can be diagonalized. That is, every such D can be uniquely decomposed in the following (see Appendix B.2.2 , Page 247):

$$D = RLR', \tag{C.2}$$

where R is an orthogonal matrix (so that $R^{-1} = R'$ and $\det(R) = 1$) and L is diagonal. This implies that we can always put ourselves in a reference system where the density

in Equation C.1 is the product of one dimensional densities. In fact, if we make the change of variables $x = Rz$, the density above can be written as a function of z:

$$\mathcal{P}(z;\theta) \propto \exp\left(-\frac{1}{2}\theta z' L z\right) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\theta \ell_i z_i^2\right). \tag{C.3}$$

where ℓ_i are the diagonal elements of L. This observation suggests that everything we need to know about improper densities of the type C.1 can be learned analyzing the simpler densities like that in Equation C.3. For this reason we next analyze in detail a simple but highly informative example in 3 dimensions.

C.2 An Intuitive Special Case

Here we take $\theta = 1$ and we consider a diagonal matrix D:

$$D \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

This matrix is positive semi-definite, since its eigenvalues, which coincide with the elements on the diagonal, are greater than or equal to zero. The rank of this matrix is 2 and its nullity is $1 = 3 - 2$. The range of D is the (x_1, x_2) plane: $\text{range}(D) \equiv \{(x_1, x_2, 0) \in \mathbb{R}^3\}$, while its null space is the x_3-axis: $\mathfrak{N}(D) \equiv \{(0, 0, x_3) \in \mathbb{R}^3\}$. Since the matrix is symmetric, its range coincides with the orthogonal complement of its null space: $\text{range}(D) = \mathfrak{N}(D)_\perp$. In the following we refer to the null space of D and its orthogonal complement as $\mathfrak{N}$ and $\mathfrak{N}_\perp$.

The density corresponding to this matrix is

$$\mathcal{P}(x_1, x_2, x_3) \propto \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \tag{C.4}$$

and is not integrable, since we have:

$$\int_{\mathbb{R}^3} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) = \int_{\mathbb{R}} dx_1 \exp\left(-\frac{1}{2}x_1^2\right) \int_{\mathbb{R}} dx_2 \exp\left(-\frac{1}{2}x_2^2\right) \int_{\mathbb{R}} dx_3 = +\infty.$$

This density has two important features:

1. It does not carry any information about the probability of the realization of values of x_3: it is ignorant with respect to x_3. More formally, we have:

$$\mathcal{P}(x_1, x_2, x_3) = \mathcal{P}(x_1, x_2, x_3') \ , \quad \forall x_3, x_3' \in \mathbb{R}$$

   The fact that the density is uniform in the direction of x_3 is what causes the density not to be integrable. The key observation here is that the direction along which the density is uniform coincides with the null space $\mathfrak{N}$.

2. Since the density C.4 does not depend on x_3 we can always set x_3 to 0 (or any other value) in its argument, and look at it as a density over $\mathbb{R}^2$ rather than $\mathbb{R}^3$. In this case the density is informative and proper:

$$\int_{R^2} dx_1 dx_2 \mathcal{P}(x_1, x_2, 0) < +\infty \tag{C.5}$$

The set of points in $\mathbb{R}^3$ of the form $(x_1, x_2, 0)$ is $\mathfrak{N}_\perp$, the orthogonal complement of the null space.

We conclude from this simple analysis that the support of the density C.4, that is $\mathbb{R}^3$, can be decomposed in two orthogonal parts: One is the null space $\mathfrak{N}$, over which the density is flat and that represents our ignorance, and the other is $\mathfrak{N}_\perp$, the orthogonal complement of $\mathfrak{N}$, over which the density is proper and carries information.

More formally, every vector $x \in \mathbb{R}^3$ can be decomposed as follows:

$$x = (x_1, x_2, x_3) = (x_1, x_2, 0) + (0, 0, x_3) = P_\perp x + P_\circ x$$

where $P_\perp$ is the projector onto $\mathfrak{N}_\perp$, and $P_\circ = I - P_\perp$ is the projector onto the null space $\mathfrak{N}$. In our case we have:

$$P_\perp \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Using this notation we can summarize what we have learned so far as:

$$\mathcal{P}(P_\circ x) = \text{ is uniform on } \mathfrak{N} \qquad \mathcal{P}(P_\perp x) = \mathcal{P}(x) \text{ is proper on } \mathfrak{N}_\perp$$

We conclude from these observations that while expected values of functions of x cannot be computed under the density C.4, since x as a random variable is ill-defined, expected values of functions of $P_\perp x$ are well defined, as long as we remember that in order to compute them we must restrict the support of the density to $\mathfrak{N}_\perp$. For example, while the covariance matrix of x does not exist, the covariance matrix of $P_\perp x$ is well defined, since it is the covariance matrix of the density $\mathcal{P}(x_1, x_2) \propto \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right)$, which is the identity.

C.3 The General Case

We now move to the general case, where the null space is a generic subspace of $\mathbb{R}^d$. As we have seen above, meaningful expectations can be computed if we restrict the support of $\mathcal{P}$ to $\mathfrak{N}_\perp$: The only difficulty here is that computing integrals over the subspace $\mathfrak{N}_\perp$ is a bit more complicated. As an exercise, and in order to convince

ourselves that the density C.1 is indeed well defined on $\mathfrak{N}_\perp$, we compute its integral, and show that it is finite. Therefore we want to compute the quantity

$$K = \int_{\mathfrak{N}_\perp} dx \exp\left(-\frac{1}{2}\theta x' D x\right)$$

In the following we denote by r the rank of D, and by $n = d - r$ the dimensionality of its null space. We adopt the convention that the eigenvalues of D, that is the diagonal elements of L in Equation C.2, are sorted in descending order, so that $\ell_1 \geq \ell_2 \geq \ldots \ell_r > \ell_{r+1} = \ell_{r+2} = \cdots = \ell_{r+n} = 0$. We start by substituting Equation C.2 in the equation above:

$$K = \int_{\mathfrak{N}_\perp} dx \exp\left(-\frac{1}{2}\theta (R'x)' L R' x\right)$$

This expression suggests that we perform the change of variable $R'x = y$, which makes the integrand a product of independent terms. The Jacobian of this transformation is one since R is a rotation matrix. The only thing left to do is to figure out how the domain of integration, $\mathfrak{N}_\perp$, changes under this transformation. Remember that $\mathfrak{N}_\perp$ is the subspace of vectors which are orthogonal to the null space, that is such that $P_\circ x = 0$. Since $P_\circ = I - P_\perp$ we have:

$$\mathfrak{N}_\perp = \{x \in \mathbb{R}^d \mid x = P_\perp x\}.$$

This can be rewritten in terms of the variable y as:

$$\mathfrak{N}_\perp = \{y \in \mathbb{R}^d \mid Ry = P_\perp Ry\} = \{y \in \mathbb{R}^d \mid y = R' P_\perp R y\}.$$

In order to understand the form $R'P_\perp Ry$, denote by r_i the columns of R. As shown in Section B.2.2, the first r columns form a basis for range$(D) = \mathfrak{N}_\perp$, and the last n columns form a basis for $\mathfrak{N}$. Denoting by y_i the components of y, we have:

$$Ry = \sum_{i=1}^{d} y_i r_i \ , \quad \Rightarrow P_\perp Ry = \sum_{i=1}^{d} y_i P_\perp r_i = \sum_{i=1}^{r} y_i r_i$$

where we have used the fact that $P_\perp r_i = r_i$ if $r_i \in \mathfrak{N}_\perp$ and $P_\perp r_i = 0$ if $r_i \in \mathfrak{N}$. Therefore :

$$R' P_\perp R y = \sum_{i=1}^{r} y_i R' r_i = \sum_{i=1}^{r} y_i e_i = (y_1, y_2, \ldots, y_r, 0, \ldots, 0)$$

where e_i is the vector $(0, \ldots, 0, 1, 0, \ldots 0)$, with the 1 in the i-th place, and we have used the fact that the columns of R are mutually orthogonal. We conclude that the only vectors y such that $y = R'P_\perp Ry$ are those whose last n components are zero, and therefore:

$$\mathfrak{N}_\perp = \{y \in \mathbb{R}^d \mid y_{r+1} = y_{r+2} = \cdots = y_{r+n} = 0\}$$

This last expression implies that when we integrate over the variable y we should keep the values of the last n components of y fixed at 0, while the first r components of y are unconstrained. Since integrating over a variable held constant is equivalent to not integrating over that variable, we summarize this finding as

$$\int_{\mathfrak{N}_\perp} dx = \int_{\mathbb{R}^d} dy \prod_{i=r+1}^{d} \delta(y_i) \ , \quad x = Ry. \tag{C.6}$$

where $\delta(y_i)$ stands for the probability density (as a function of y_i) with unit mass at the origin. Now we complete our change of variables, and rewrite:

$$K = \int_{\mathbb{R}^d} dy \prod_{i=r+1}^{d} \delta(y_i) \exp\left(-\frac{1}{2}\theta y' L y\right).$$

The exponent in this expression does not depend on $y_{r+1}, \ldots, y_{r+n}$, and so simplifies to:

$$K = \int_{\mathbb{R}^r} dy_1 \ldots dy_r \exp\left(-\frac{1}{2}\theta \sum_{i=1}^{r} \ell_i y_i^2\right) = \prod_{i=1}^{r} \left[\int_{\mathbb{R}} dy_i \exp\left(-\frac{1}{2}\theta \ell_i y_i^2\right)\right].$$

From the Gaussian integral identity in Equation B.28, we have:

$$\int_{\mathbb{R}} dy_i \exp\left(-\frac{1}{2}\theta \ell_i y_i^2\right) = \left(\frac{2\pi}{\theta \ell_i}\right)^{\frac{1}{2}}.$$

Putting everything together we obtain:

$$K = \left(\frac{2\pi}{\theta}\right)^{\frac{r}{2}} \frac{1}{\sqrt{\prod_{i=1}^{r} \ell_i}}.$$

Defining the quantity:

$$\det D_\perp \equiv \prod_{i=1}^{r} \ell_i$$

we can finally rewrite the expression above as the following identity:

$$\int_{\mathfrak{N}_\perp} dx \exp\left(-\frac{1}{2}\theta x' D x\right) = \left(\frac{2\pi}{\theta}\right)^{\frac{r}{2}} \frac{1}{\sqrt{\det D_\perp}}. \tag{C.7}$$

Notice the complete analogy between this expression and the identity in Equation B.28. The number above can be thought of as the normalization constant for the density C.1 restricted to $\mathfrak{N}_\perp$, that we present here for completeness:

$$\mathcal{P}(x;\theta) = \left(\frac{\theta}{2\pi}\right)^{\frac{r}{2}} \sqrt{\det D_\perp} \exp\left(-\frac{1}{2}\theta x' D x\right) \ , \quad x \in \mathbb{R}^d \ , \quad x \in \mathfrak{N}_\perp. \tag{C.8}$$

Using a similar technique we can now compute the expected values of several quantities of interest. In the following we denote by $\mathrm{E}_\perp[\cdot]$ the expected value with respect to the density C.8. It is possible to show that the covariance matrix is:

$$\mathrm{E}_\perp[xx'] = \frac{1}{\theta} D^+ \tag{C.9}$$

where D^+ is the pseudo-inverse of D, as defined in Section B.2.5 (Page 251). Using this result it is then easy to see that:

$$\mathrm{E}_\perp[x' D x] = \frac{r}{\theta}. \tag{C.10}$$

This last result can also be derived by first computing the distribution of the quantity $H = x'Dx$, and then computing its expected value. The quantity H, which in general is interpreted as a smoothness functional, has the following distribution:

$$\mathcal{P}(H) = K H^{\frac{r}{2}-1} e^{-\frac{1}{2}\theta H} \tag{C.11}$$

In the notation of Chapter 9 we would write $H \sim \mathcal{G}(r, \theta)$.

C.4 Drawing Random Samples

Here we consider the problem of sampling from the improper normal density of Equation C.1. We have seen in the previous sections that this density is not proper, but it becomes proper if we restrict its argument to the r-dimensional subspace $\mathfrak{N}_\perp$, that is the portion of $\mathbb{R}^d$ on which the prior is informative. This restriction means that we should think of the argument x of the density as a linear combination of r independent elements $\mathfrak{N}_\perp$. We know from results about the eigenvectors/eigenvalue decomposition (Section B.2.2 , Page 247) that a basis for $\mathfrak{N}_\perp$ is given by the first r columns of R (where $D = RLR'$), which we collect as columns of a $d \times r$ matrix $R_\perp$. Therefore the restriction of the domain of the improper density $P(x;\theta)$ is simply expressed as $x = R_\perp z,\ z \in \mathbb{R}^r$. The role of the matrix $R_\perp$ here is simply to take $\mathbb{R}^r$ and rotate it to obtain $\mathfrak{N}_\perp$. We notice that this transformation is invertible ($z = R'_\perp x$) and that, since it is a rotation, its Jacobian is 1. Therefore the density of the variable z is:

$$\mathcal{P}(z;\theta) \propto \exp\left(-\frac{1}{2}\theta (R_\perp z)' D R_\perp z\right) = \exp\left(-\frac{1}{2}\theta z' R'_\perp D R_\perp z\right)$$

This probability density is now well defined, and therefore all we have to do in order to sample from the improper normal of Equation C.1 is the following:

$$x = R_\perp z \ , \quad z \sim \mathcal{N}\left(0, \frac{1}{\theta}\left(R'_\perp D R_\perp\right)^{-1}\right) \tag{C.12}$$

The covariance of the variable z in the expression above is unnecessarily complicated, and can be greatly simplified. To this end it suffices to remember that, from Equation B.19, the matrix D can be written as $D = R_\perp \ell R'_\perp$, where ℓ is the $r \times r$ diagonal matrix whose diagonal elements are the non-zero eigenvalues of D. Substituting this formula in Equation C.12, and remembering that $R'_\perp R_\perp = I$, we arrive at the following simple formula:

$$x = R_\perp z \ , \quad z \sim \mathcal{N}\left(0, \frac{1}{\theta}\ell^{-1}\right) . \tag{C.13}$$

Appendix D

Discretization of the Derivative Operator

In this section we consider the problem of evaluating the derivative of a function known at a finite number of points on an equally spaced grid. This is the case for examples when the variable is age or time. We report here a simple way to obtain an approximate derivative of arbitrary order, and refer the reader to texts on numerical analysis for an analysis of the error. The equations given here will reproduce, with appropriate choices of the parameters, well known formulas (like the Richardson extrapolation; Press et al., 1987).

Let $f(x)$ be a function whose values are known on an integer grid (so x is always an integer in the following). Our problem is to evaluate the derivative of some order at the point x. The starting point is to assume that f can be represented by a polynomial of degree k:

$$f(x+n) = f(x) + \sum_{j=1}^{k} \frac{n^j}{j!} f^{(j)}(x) \tag{D.1}$$

where the k derivatives $f^{(j)}(x)$ in this expression are unknown. The idea here is that if we evaluate Equation D.1 at k different points we obtain a linear system that we can solve for $f^{(j)}(x)$.

Let $\mathbf{n}$ be a set of k integers that excludes 0, for example $\mathbf{n} = \{-2, -1, 1, 2\}$. We then rewrite Equation D.1 as

$$f(x+n_i) - f(x) = \sum_{j=1}^{k} A_{ij} f^{(j)}(x) \quad i = 1, \ldots k,$$

where the matrix A has elements $A_{ij} \equiv \frac{n_i^j}{j!}$. By inverting this linear system, we obtain the intuitively pleasing formula:

$$f^{(j)}(x) = \sum_{i=1}^{k} A_{ji}^{-1}[f(x+n_i) - f(x)] \tag{D.2}$$

This is just the statement of the fact that the derivative of order j at a point is a weighted combination of the values of the function at nearby points: the points $f(x+n_i)$ receive a weight A_{ji}^{-1}, while $f(x)$ receives a weight $-\sum_{i=1}^{k} A_{ji}^{-1}$, so that the sum of the weights is zero.

With reasonable choices of the set of integers $\mathbf{n}$ Equation D.2 is sufficient to recover the standard formulas for numerical derivatives. For example, choosing $k = 4$ and $\mathbf{n} = \{-2, -1, 1, 2\}$ we obtain, for $j = 2$, the classic 5-point central second derivative formula, which is fourth-order accurate:

$$f^{(2)}(x) \approx \frac{1}{12}\left(f(x-2) - 8f(x-1) + 8f(x+1) - f(x+2)\right).$$

While choosing a set $\mathbf{n}$ which is symmetric around the origin is always recommended, a non-symmetric $\mathbf{n}$ is needed when the function is defined over a finite segment and we need to evaluate the derivative near the end points. For example if we know the values of the function at the integers $0, 1, \ldots 10$ and we want the derivative at $x = 0$ using a five point formula (that is $k = 4$), the boundary at 0 forces us to choose $\mathbf{n} = \{1, 2, 3, 4\}$. Following the criterion that the set $\mathbf{n}$ should be as symmetric as possible, for the evaluation of the second derivative at $x = 1$ we will make the choice $\mathbf{n} = \{-1, 1, 2, 3\}$. Therefore if we denote by $\mathbf{f}$ the vector whose elements are the values of the function f on an integer grid, its derivative of order j is the vector $D^j\mathbf{f}$ where D^j is a square matrix whose rows contain the weights that can be computed according to Equation D.2. The matrix D^j is a k-band matrix: except for the first and last few rows, where we cannot use a symmetric choice of n, the elements along the band are equal to each other. In addition, we know from Equation D.2 that each row sums up to 0, that is $\sum_{ij} D^j_{ik} = 0$. This property is just a reflection of the fact that any differential operator of degree at least one annihilates the constant function, that is $D^j\mathbf{f} = 0$ if $\mathbf{f}$ is a a constant vector.

For example, using $k = 4$ and $j = 1$ we have:

$$D^1 = \begin{vmatrix}
-2.0833 & 4 & -3 & 1.333 & -0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-0.25 & -0.8333 & 1.5 & -0.5 & 0.083 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0833 & -0.6667 & 0 & 0.6667 & -0.0833 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0833 & -0.6667 & 0 & 0.6667 & -0.0833 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0833 & -0.6667 & 0 & 0.6667 & -0.0833 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.083 & 0.5 & -1.5 & 0.8333 & 0.25 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & -1.333 & 3 & -4 & 2.0833
\end{vmatrix}$$

The matrix D^1 is nearly antisymmetric (that is is antisymmetric except for the "border" effects). This is not an accident: it follows from simple properties of differential operators that D^j will be almost antisymmetric for j odd and almost symmetric for j even.

Appendix E

Smoothness over Graphs

This appendix defines the mathematical notion of smoothness over geographic areas using concepts from graph theory. The ideas here have relatively few practical uses in the book, but they do convey the essential unity of the priors we offer defined across any type of underlying variable. For priors defined over discretized continuous variables, we use analogous ideas to reduce the task of specifying the (non)spatial contiguity matrix to the choice of a single number (the order of the derivative in the smoothing functional). Unfortunately, a parallel reduction of effort is not possible for geographic space, although similar ideas apply.

We begin by denoting by G the set of cross-sectional indexes. When cross-sections vary by age, or similar variables, the set G is a discrete or continuous set endowed with a natural metric, and it is easy to formalize the notion of smoothness using derivatives or their discretized versions. When the cross-sectional indexes are labels like country, familiar calculus does not help to define a notion of smoothness, but graph theory does.

In fact, when we have a number of cross-sections that we want to pool (partially), it is natural to represent each of them as a point on the 2-dimensional plane, and join by a line segment points corresponding to cross-sections that we consider "neighbors" of each other. This construction is called a *graph*, where we call the points *vertices* and the line segments *edges* of the graph. We denote the vertices and edges by V and E, respectively. Both vertices and edges are numbered using increasing positive integers: Any numbering scheme is allowed, as long as it is used consistently. If i and j are two vertices connected by the edge e, we assign to e a *weight* $w(e) \equiv s_{ij}$, which represents our notion of how close the two cross-sections are. The quantity $\rho(i,j) \equiv \frac{1}{\sqrt{s_{ij}}}$ is called the *length* of the edge e and it is thought of as the distance between cross-sections i and j. In the simplest case s, which is called the *adjacency matrix* of the graph, is a matrix of zeros and ones, where the ones denote country pairs which are considered as neighbors. If no edge exists between i and j, we set $s_{ij} = 0$. Vertex i could be connected to itself (in which case we would have a *loop*),

so that we could have $s_{ii} \neq 0$: We will see below that for our purposes the value of s_{ii} is irrelevant, so we arbitrarily set $s_{ii} = 0$. The number of edges connected to the vertex i is called the *degree* of i, and we denote it by $s_i^+ \equiv \sum_j s_{ij}$ (in other words s_i^+ is the sum of the elements of the i-th row of s, which is the number of neighbors).

If we have a function f defined over the graph, that is $f : V \to V$, it is possible to introduce the notion of a gradient. This is done by introducing the *oriented incidence matrix* Q, which is a $\#V \times \#E$ matrix whose rows and columns are indexed by the indices for V and E. To define Q, we first need to *orient* the graph, that is we assign a direction to each edge, so that edge e will point from some vertex i (the initial vertex) to some other vertex j (the terminal vertex). The orientation is arbitrary for our purposes, as long as it is fixed once for all. The matrix Q is built by setting entry Q_{ie} to $\sqrt{w(e)}$ if i is the initial vertex of e, to $-\sqrt{w(e)}$ if i is the terminal vertex of e, and 0 otherwise.

For example, if the first of row of Q is $(0, 0, 1, 0, -2)$, then vertex 1 is the terminal vertex of edge 5, which has a weight $w(5) = 4$, and the initial vertex of edge 3, which has a weight $w(3) = 1$. Notice that since each edge must have one initial and one terminal point, the columns of Q have even numbers of non-zero elements, and must sum up to 0.

Now that we have the matrix Q we define a meaningful differential operator. At any given vertex we think of the edges connected to that point as abstract "directions" from which one can leave that vertex. Therefore, given a function defined over V, it makes sense to characterize its local variation in terms of how much the function changes along each direction, that is to assign to the edge e running from vertex i to vertex j the quantity

$$\frac{f(j) - f(i)}{\rho(i, j)} ,$$

which obviously resembles a derivative. The matrix Q allows us to group all the quantities of this type in one single vector, which we think of as the gradient of the function f. In fact, denote by f the vector $(f)_i \equiv f(i) \ \ \forall i \in V$, and suppose e is the edge that runs from i to j, then by construction

$$(Q'f)_e = \frac{f(j) - f(i)}{\rho(i, j)} ,$$

Therefore we think of $Q'f$ as the gradient of f, and as Q' as the gradient operator. A measure of the smoothness of a function f defined over V is therefore obviously the quantity $\|Q'f\|^2$. The definition of smoothness we give in Chapter 4 is operationally equivalent to this. A simple result of graph theory shows, however, that we need not compute the matrix Q' of the gradient in order to compute $\|Q'f\|^2$. In fact, the gradient operator is strictly connected to another important operator defined over the graph, that is the *Laplacian*, defined more simply in terms of the adjacency matrix s:

$$W \equiv s^{+} - s$$

where following relationship is well known:

$$W = QQ'.$$

A general smoothness functional for a function f defined over the graph therefore has the following form:

$$H[f] = \|Q'f\|^2 = f \cdot Wf = \sum_{ij} W_{ij} f_i f_j,$$

which can alternatively be written as:

$$H[f] = \frac{1}{2} \sum_{ij} s_{ij} (f_i - f_j)^2.$$

Appendix F

Software Implementation

[Our software should be ready in a few months. When available, it will be posted at `http://GKing.Harvard.edu/stats.shtml`, and described in this appendix. The first version released will be a command driven program that works within R. A few months afterwards, we plan to have a menu-based version too.]

Bibliography

Abu-Mostafa, Y. 1992. A Method for Learning from Hints. In *Advances in Neural information processings systems 5*, ed. S. J. Hanson, Jack D. Cowan and C. Lee Giles. San Mateo, CA: Morgan Kaufmann Publishers.

Alho, J. M. 1992. "Comment on "Modeling and Forecasting U.S. Mortality" by R. Lee and L. Carter." *Journal of the American Statistical Association* 87(419, September):673–674.

Armstrong, J. Scott. 2001. Extrapolation of Time Series and Cross-Sectional Data. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, ed. J. Scott Armstrong. Kluwer pp. 217–243.

Beck, Nathaniel and Jonathan Katz. 1995. ""What to Do (and Not to Do) with Time-Series-Cross-Section Data"." *American Political Science Review* 89:634–647.

Beck, Nathaniel and Jonathan Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Model." *Political Analysis* VI:1–36.

Beck, Nathaniel, Jonathan Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Political Science Review* 42:1260–1288.

Bell, W.R. 1997. "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates." *Journal of Official Statistics* 13(3):279–303.

Bell, W.R. and B.C. Monsell. 1991. "Using Principal Components in time Series modeling and Forecasting of Age-Specific Mortality Rates." *Proceedings of the American Statistical Association, Social Statistics Section* pp. 154–159.

Beltrami, E. 1873. "Sulle funzioni bilineari." *Giornale di Matematiche ad Uso degli Studenti Delle Universitá* 11:98–106. An English translation by D. Boley is available as University of Minnesota, Department of Computer Science, Technical Report 90-37, 1990.

Berger, James. 1994. "An Overview of Robust Bayesian Analysis (With Discussion)." *Test* 3:5–124.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems (With Discussion)." *Journal of the Royal Statistical Society, B* 36:192–236.

Besag, Julian. 1975. "Statistical Analysis of Non-Lattice Data." *The Statistician* 24(3):179–195.

Besag, Julian. 1986. "On the Statistical Analysis of Dirty Pictures." *Journal of the Royal Statistical Society B* 48(3):259–302.

Besag, Julian and Charles Kooperberg. 1995. "On Conditional and Intrinsic Autoregressions." *Biometrika* 82(4):733–746.

Besag, Julian and David M. Higdon. 1999. "Bayesian Analysis of Agricultural Field Experiments (With Discussion)." *Journal of the Royal Statistical Society, B* 61(4):691–746.

Biggs, N.L. 1993. *Algebraic Graph Theory*. 2nd ed. Cambridge, UK: Cambridge University Press.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, Y.M. M., S.E. Fienberg and P.W. Holland. 1975. *Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Blattberg, R.C. and E.I. George. 1991. "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations." *Journal of the American Statistical Association* 86(414, Jun):304–315.

Booth, Heather, John Maindonald and Len Smith. 2002. "Applying Lee-Carter Under Conditions of Variable Mortality Decline." *Population Studies* 56(3):325–336.

Bozik, J.E. and W.R. Bell. 1987. "Forecasting Age Specific Fertility Using Principal Components." *Proceedings of the Americal Statistical Association, Social Statistics Section* pp. 396–401.

Cameron, A. Colin and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Carlin, Bardley P. and Thomas A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. CRC Press.

Clements, M.P. and D.F. Hendry. 1998. *Forecasting Economic Time Series*. Cambridge, U.K.: Cambridge University Press.

Coale, Ansley J. and Paul Demeny. 1966. *Regional Model Life Tables and Stable Populations*. Princeton, N.J.: Princeton University Press.

Collier, David and Jr. James E. Mahon. 1993. "Conceptual 'Stretching' Revisited: Adapting Categories in Compartive Analysis." *American Political Science Review* 87(4, December):845–855.

Dawid, A. P. 1983. Invariant Prior Distributions. In *Encyclopedia of Statistical Sciences*, ed. S. Kotz, S. Johnson and C.B. Read. Vol. 4 Wiley-Interscience pp. 228–236.

de Boor, C. 1978. *A Practical Guide to Splines*. New York: Springer-Verlag.

Edwards, A.W.F. 1972. *Likelihood*. New York: Cambridge University Press.

Efron, B. 1979. "Bootstrap methods: another look at the jackknife." *Annals of Statistics* 7:1–26.

Efron, B. 1982. *The Jacknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B. and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapmand and Hall.

Eubank, R.L. 1988. *Spline Smoothing and Nonparametric Regression*. Vol. 90 of *Statistics, textbooks and monographs* Basel: Marcel Dekker.

Gelfand, A.E. and Smith, A.F.M. 1990. "Sampling-based approach to calculating marginal densities." *Journal of the American Statistical Association* 85:398–409.

Gelman, Andrew, Gary King and Chuanhai Liu. 1999. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93(433, September):846–857. http://gking.harvard.edu/files/abs/not-abs.shtml.

Gelman, Andrew, J.B. Carlin, H.S. Stern and D.B. Rubin. 1995. *Bayesian Data Analysis*. Chapman & Hall.

Geman, Stuart and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *I.E.E.E. Transactions: Pattern Analysis and Machine Intelligence* 6:721–741.

Gilks, W.R., S. Richardson and D.J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Gill, Jeff. 2002. *Bayesian Methods for the Social and Behavioral Sciences*. London: Chapman and Hall.

Girosi, F. 1991. Models of noise and robust estimates. A.I. Memo 1287 Artificial Intelligence Laboratory, Massachusetts Institute of Technology. ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1287.pdf.

Girosi, Federico and Gary King. 2005. "A Reassessment of the Lee-Carter Mortality Forecasting Method." http://gking.harvard.edu/files/abs/lc-abs.shtml.

Golub, G., M. Heath and G. Wahba. 1979. "Generalized cross validation as a method for choosing a good ridge parameter." *Technometrics* 21:215–224.

Gompertz, B. 1825. "On the Nature of the Function Expressive of the Law of Mortality." *Philosophical Transactions* 27:513–585.

Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663–666.

Goss, Stephen C., Alice Wade, Felicitie Bell and Bernard Dussault. 1998. "Historical and Projected Mortality for Mexico, Canada, and the United States." *North American Actuarial Journal* 4(2):108–126.

Government's Actuary Department. 2001. *National Population Projections: Review of Methodology for Projecting Mortality*. London: National Statistics Direct, UK. http://www.statistics.gov.uk/.

Graunt, John. 1662. *Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality*. London: John Martyn and James Allestry.

Griffiths, W. E., R. Carter Hill and C. J. O'Donnell. 2001. Including Prior Information in Probit Model Estimation. Working Papers 816 Department of Economics University of Melbourne: .

Gutterman, Sam and Irwin T. Vanderhoof. 1998. "Forecasting Changes in Mortality: A Search for a Law of Causes and Effects." *North American Actuarial Journal* 4(2):135–138.

Haberland, J. and K.E. Bergmann. 1995. "The Lee-Carter Model of the Prognosis of Mortality in Germany." *Gesundheitswesen* 57(10, October):674–679. article in German.

Hastie, Trevor, R. Tibshirani and J. Friedman. 2001. *The Elements of Statistical Learning*. New York: Springer Verlag.

Heligman, L. and J. H. Pollard. 1980. "The Age Pattern of Mortality." *Journal of the Institute of Acturaries* 107:49–80.

Henderson, R. 1924. "A new method of graduation." *Transaction of the Actuarial Society of America* 119:457–526.

Ibrahim, Joseph G. and Ming-Hui Chen. 1997. "Predictive Variable Selection for the Multivariate Linear Model." *Biometrics* 53(June):465–478.

Jee, Sun Ha, Il Soon Kim, Il Suh, Dongchun Shin and Lawrence J Appel. 1998. "Projected Mortality from Lung Cancer in South Korea, 1980-2004." *International Journal of Epidemiology* 27:365–369.

Jeffreys, H. 1961. *Theory of Probability.* 3rd (1st edn., 1939 ed. Oxford: Clarendon Press.

Jordan, C. 1874. "Mémoire sur les formes bilinéaires." *Comptes Rendus de l' Académie des Sciences, Paris* 78:614–617.

Kadane, Joseph B. 1980. Predictive and Structural Methods for Eliciting Prior Distributions. In *Bayesian Analysis in Econometrics and Statistics*, ed. Arnold Zellner. North-Holland.

Kadane, Joseph B., J.M. Dickey, R.L. Winkler, W.S. Smith and S.C. Peters. 1980. "Interactive Elicitation of Opinion for a Normal Linear Model." *Journal of the American Statistical Association* 75:845–854.

Kadane, Joseph B. and Lara J. Wolfson. 1998. "Experiences in Elicitation." *The Statistician* 47(1):3–19.

Kass, Robert E. and Larry Wasserman. 1996. "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association* 91(435, September):1343–1370.

Keyfitz, N. 1968. *Introduction to the Mathematics of Population.* Reading, MA: Addison Wesley.

Keyfitz, N. 1982. "Choice of Function for mortality Analysis: Effective Forecasting Depends on a Minimum Parameter Representation." *Theoretical Population Biology* 21:239–252.

Kimeldorf, G.S. and G. Wahba. 1970. "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines." *Ann. Math. Statist.* 41(2):495–502.

King, Gary. 1988. "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for The Exponential Poisson Regression Model." *American Journal of Political Science* 32(3, August):838–863. http://gking.harvard.edu/files/abs/epr-abs.shtml.

King, Gary. 1989*a*. "Representation Through Legislative Redistricting: A Stochastic Model." *American Journal of Political Science* 33(4, November):787–824. http://gking.harvard.edu/files/abs/repstoch-abs.shtml.

King, Gary. 1989*b*. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Michigan University Press.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

King, Gary and Curtis S. Signorino. 1996. "The Generalization in the Generalized Event Count Model." *Political Analysis* 6:225–252. http://gking.harvard.edu/files/abs/generaliz-abs.shtml.

King, Gary and Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409–1427. http://gking.harvard.edu/files/abs/1s-abs.shtml.

King, Gary and Langche Zeng. 2005. "The Dangers of Extreme Counterfactuals." http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Künsch, Hans R. 1987. "Intrinsic Autoregressions and Related Models on the Two-Dimensional Lattice." *Biometrika* 74(3):517–524.

LaPalombara, Joseph. 1968. "Macrotheories and Microapplications in Comparative Politics: A Widening Chasm." *Comparative Politics* (October):52–78.

Laplace, P.S. 1951, original: 1820. *Philosophical Essays on Probaiblities*. New York: Dover.

Laud, Purushottam W. and Joseph G. Ibrahim. 1995. "Predictive Model Selection." *Journal of the Royal Statistical Society, B* 57(1):247–262.

Laud, Purushottam W. and Joseph G. Ibrahim. 1996. "Predictive Specification of Prior Model Probabilities in Variable Selection." *Biometrika* 83(2):267–274.

Ledermann, S. and J. Breas. 1959. "Les Dimensions de la Mortalité." *Population* 14:637–682. [in French].

Lee, Ronald D. 1993. "Modeling and Forecasting the Time Series of US Fertility: Age Patterns, Range, and Ultimate Level." *International Journal of Forecasting* 9:187–202.

Lee, Ronald D. 2000*a*. "The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications." *North American Actuarial Journal* 4(1):80–93.

Lee, Ronald D. 2000*b*. "Long-Term Projections and the US Social Security System." *Population and Development Review* 26(1, March):137–143.

Lee, Ronald D., Lawrence Carter and S. Tuljapurkar. 1995. "Disaggregation in Population Forecasting: Do We Need It? And How to Do it Simply." *Mathematical Population Studies* 5(3, July):217–234.

Lee, Ronald D. and Lawrence R. Carter. 1992*a*. "Modeling and Forecasting U.S. Mortality." *Journal of the American Statistical Association* 87(419, September).

Lee, Ronald D. and Lawrence R. Carter. 1992*b*. "Modeling and Forecasting U.S. Mortality." *Journal of the American Statistical Association* 87(419, September).

Lee, Ronald D. and R. Rofman. 1994. "Modeling and Projecting Mortality in Chile." *Notas Poblacion* 22(59, Jun):183–213.

Lee, Ronald D. and S. Tuljapurkar. 1994. "Stochastic Population Projections for the U.S.: Beyond High, Medium and Low." *Journal of the American Statistical Association* 89(428, December):1175–1189.

Lee, Ronald D. and S. Tuljapurkar. 1998*a*. Stochastic Forecasts for Social Security. In *Frontiers in the Economics of Aging*, ed. David Wise. Chicago: University of Chicago Press pp. 393–420.

Lee, Ronald D. and S. Tuljapurkar. 1998*b*. "Uncertain Demographic Futures and Social Security Finances." *American Economic Review: Papers and Proceedings* (May):237–241.

Lee, Ronald D. and Timothy Miller. 2001. "Evaluating the Performance of the Lee-Carter Approach to Modeling and Forecasting Mortality." *Demography* 38(4, November):537–549.

Li, S.Z. 1995. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag.

Lilienfeld, David E. and Daniel P. Perl. 1994. "Projected Neurodegenerative Disease Mortality among Minorities in the United States." *Neuroepidemiology* 13:179–186.

Macridis, Roy C. 1955. *The Study of Comparative Government*. New York: Doubleday and Co.

Marshall, R.J. 1991. "Mapping Disease and Mortality Rates using Empirical Bayes Estimators." *Applied Statistics* 40(2):283–294.

McNown, Rober and Andrei Rogers. 1989. "Forecasting Mortality: A Parameterized Time Series Approach." *Demography* 26(4):645–660.

McNown, Robert. 1992. "Comment." *Journal of the American Statistical Association* 87(419):671–672.

McNown, Robert and Andrei Rogers. 1992. "Forecasting Cause-Specific Mortality Using Time Series Methods." *International Journal of Forecasting* 8:413–432.

Morozov, V.A. 1984. *Methods for solving incorrectly posed problems*. Berlin: Springer-Verlag.

NIPSSR. 2002. *Population Projections for Japan (January, 2002)*. National Institute of Population and Social Security Research.

Niyogi, P., F. Girosi and T. Poggio. 1998. "Incorporating Prior Information in Machine Learning by Creating Virtual Examples." *Proceedings of the IEEE.* 86(11):2196–2209.

Oman, Samuel D. 1985. "Specifying a Prior Distribution in Structured Regression Problems." *Journal of the American Statistical Association* 80(389, March):190–195.

Plackett, R.L. 1981. *The Analysis of Categorical Data*. London: Griffin.

Press, William H., Saul Teukolsky, William T. Vetterling and Brian P. Flannery. 1987. *Numerical Recipes: the Art of Scientific Computing*. Cambridge: Cambridge University Press.

Preston, Samuel H. 1991. Demographic Change in the United States, 1970–2050. In *Forecasting the Health of Elderly Populations*, ed. K.G. Manton, B.H. Singer and R.M. Suzman. New York: Springer-Verlag pp. 51–77.

Preston, Samuel H., Patrick Heuveline and Michel Guillot. 2001. *Demography: Measuring and Modeling Population Processes*. Oxford, England: Blackwell.

Rogers, Andrei. 1986*a*. "Parameterized Multistate Population Dynamics and Projections." *Journal of the American Statistical Association* 81(393, March):48–61.

Rogers, Andrei. 1986*b*. "Parameterized Multistate Population Dynamics and Projections." *Journal of the American Statistical Association* 81:48–61.

Rogers, Andrei and James Raymer. 1999. "Fitting Observed Demographic Rates with the Multiexponential Model Schedule: An Assessment of Two Estimation Programs." *Applied Regional Science Conference* 11(1):1–10.

Salomon, Joshua A., Milton C. Weinstein, James K. Hammitt and Sue J. Goldie. 2002. "Empirically Calibrated Model of Hepatitis C Virus Infection in the United States." *American Journal of Epidemiology* 156:761–773.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4, December):1033–1053.

Schoenberg, I.J. 1946. "Contributions to the problem of approximation of equidistant data by analytic functions, Part A: On the problem of smoothing of graduation, a first class of analytic approximation formulae." *Quart. Appl. Math.* 4:45–99.

Schumaker, L.L. 1981. *Spline functions: basic theory.* New York: John Wiley and Sons.

Sivamurthy, M. 1987. Principal Components Representation of ASFR: Model of Fertility Estimation and Projection. In *CDC Research Monograph.* Cairo Demographic Center: pp. 655–693.

Speckman, P. and D. Sun. 2001. "Bayesian Nonparametric Regression and Autoregression Priors." www.stat.missouri.edu/~speckman/report/bnpreg.ps.

Stewart, G.W. 1992. On the Early History of the Singular Value Decomposition. Institute for Advanced Computer Studies TR-92-31 University of Maryland, College Park.

Stimson, James A.̇1985. "Regression Models in Space and Time: A Statistical Essay." *American Journal of Political Science* 29:914–947.

Strang, G. 1988. *Linear Algebra and Its Applications.* Saunders.

Sun, D., R. Tsutakawa, H. Kim and Z. He. 2000. "Spatio-temporal Interaction with Disease Mapping." *Statistics in Medicine* 19:2015–2035.

Tabeau, Ewa. 2001. A Review of Demographic Forecasting Models for Mortality. In *Forecasting Mortality in Developed Countries*, ed. Anneke van de Berg Jeths Ewa Tabeau and Christopher Heathcoate. The Netherlands: Kluwer Academic Publishers chapter 1, pp. 1–32.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer-Verlag.

Taylor, G. 1992. "A Bayesian interpretation of Whittaker-Henderson graduation." *Insurance: Mathematics and Economics* 11:7–16.

Tikhonov, A. N. 1963. "Solution of incorrectly formulated problems and the regularization method." *Soviet Math. Dokl.* 4:1035–1038.

Tikhonov, A. N. and V. Y. Arsenin. 1977. *Solutions of Ill-posed Problems.* Washington, D.C.: W. H. Winston.

Tuljapurkar, S., N. Li and C. Boe. 2000. "A Universal Pattern of Mortality Decline in the G7 Countries." *Nature* 405(June):789–792.

Tuljapurkar, Shripad and Carl Boe. 1998. "Mortality Change and Forecasting: How Much and How Little Do We Know?" *North American Actuarial Journal* 2(4).

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. New York: Wiley.

Verdecchia, Arduino, Giovanni De Angelis and Riccardo Capocaccia. 2002. "Estimation and Projections of Cancer Prevalence from Cancer Registry Data." *Statistics in Medicine* 21:3511–3526.

Verrall, R.J. 1993. "A state space formulation of Whittaker graduation, with extensions." *Insurance: Mathematics and Economics* 13:7–14.

Wahba, G. 1975. "Smoothing noisy data by spline functions." *Numer. Math* 24:383–393.

Wahba, G. 1978. "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression." *Journal of the Royal Statistical Society B* 40(3):364–372.

Wahba, G. 1980. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. In *Approximation theory III*, ed. W. Cheney. New York: Academic Press pp. 905–912.

Wahba, G. 1990. *Splines Models for Observational Data*. Philadelphia: Series in Applied Mathematics, Vol. 59, SIAM.

Waller, L.A., B.P. Carlin, H. Xia and A.E. Gelfand. 1997. "Hierarchical Spatio-Temporal Mapping of Disease Rates." *Journal of the American Statistical Association* 92(438):607–617.

Weinstein, Milton C., Pamela G. Coxson, Lawrence W. Williams, Theodore M. Pass, William B Stason and Lee Goldman. 1987. "Forecasting Coronary Heart Disease Incidence, Mortality, and Cost: The Coronary Heart Disease Policy Model." *American Journal of Public Health* 77(11):1417–1426.

Weiss, Robert E., Yan Wang and Joseph G. Ibrahim. 1997. "Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors." *Biometrics* 53(June):592–602.

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: a Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42(4, October):1233–1259.

Whittaker E.T. 1923. "On a New Method of Graduation." *Proceedings of the Edinburgh Mathematical Society* 41:63–75.

Wilmoth, John. 1993. Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change. Technical report Department of Demography, University of California, Berkeley.

Wilmoth, John R. 1996. Mortality Projections for Japan: A Comparison of Four Methods. In *Health and Mortality Among Elderly Populations*, ed. G. Caselli and Alan Lopez. Oxford: Oxford University Press pp. 266–287.

Zellner, A. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association* 57(298, June):348–368.

Zorn, Christopher. 2001. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." *American Journal of Political Science* 45(April):470–490.